Genetic Diversity of Bundibugyo Ebolavirus from Uganda and the Democratic Republic of Congo

Isaac Emmanuel Omara[1,2], Sylvia Kiwuwa-Muyingo[3,7], Stephen Balinandi[1], Luke Nyakarahuka[1,5], Jocelyn Kiconco[1], John Timothy Kayiwa[1], Gerald Mboowa[2,4], Daudi Jjingo[4,6], Julius J. Lutwama[1]

Affiliations

1. Department of Arbovirology, Emerging and Re-emerging Infectious Diseases, Uganda Virus Research Institute, Entebbe, Wakiso, Uganda

2. Department of Immunology and Molecular Biology, School of Biomedical Sciences, College of Health Sciences, Makerere University, Kampala, Uganda

3. Department of Data Measurement and Evaluation, African Population and Health Research Center, Nairobi, Kenya

4. The African Center of Excellence in Bioinformatics and Data-Intensive Sciences, The Infectious Disease Institute, Makerere University, Kampala, Uganda

5. School of Bio-security, Bio-technical and Laboratory Sciences, College of Veterinary Medicine, Animal Resources and Bio-security, Makerere University, Kampala, Uganda

6. Department of Computer Science, College of Computing and Information Sciences, Makerere University, Kampala, Uganda

7. MRC/UVRI and LSHTM Uganda Research Unit, Entebbe, Uganda

Corresponding author

Email: omara.isaac.88@gmail.com (OIE)

Author Contributions

Isaac E. Omara: Conceptualization, retrieved and curated the data, formal analysis, Interpretations, drafted original manuscript, coordinated manuscript writing and editing

Sylvia Kiwuwa-Muyingo: Made comments, guided and provided mentorship throughout the manuscript writing up to submission.

Stephen Balinandi: Involved in conceptualization, made comments and reviewed manuscript

Luke Nyakarahuka: Involved in conceptualization, made comments and reviewed manuscript

John T. Kayiwa: Made comments and reviewed manuscript

Jocelyn Kiconco: Made comments and reviewed manuscript

Gerald Mboowa: Guidance during conceptualization, overall over sight during project execution

Daudi Jjingo: Guidance during conceptualization, overall over sight during project execution

Julius J. Lutwama: Guidance during conceptualization, overall over sight during project execution

Abstract

Background

The Ebolavirus is one of the deadliest viral pathogens which was first discovered in the year 1976 during two consecutive outbreaks in the Democratic Republic of Congo and Sudan. Six known strains have been documented. The *Bundibugyo Ebolavirus* in particular first emerged in the year 2007 in Uganda. This outbreak was constituted with 116 human cases and 39 laboratory confirmed deaths. After 5 years, it re-emerged and caused an epidemic for the first time in the Democratic Republic of Congo in the year 2012 as reported by the WHO. Here, 36 human cases with 13 laboratory confirmed deaths were registered. Despite several research studies conducted in the past, there is still scarcity of knowledge available on the genetic diversity of *Bundibugyo Ebolavirus*. We undertook a research project to provide insights into the unique variants of *Bundibugyo Ebolavirus* that circulated in the two epidemics that occurred in Uganda and the Democratic Republic of Congo

Materials and Methods

The Bioinformatics approaches used were; Quality Control, Reference Mapping, Variant Calling, Annotation, Multiple Sequence Alignment and Phylogenetic analysis to identify genomic variants as well determine the genetic relatedness between the two epidemics. Overall, we used 41 viral sequences that were retrieved from the publicly available sequence database, which is the National Center for Biotechnology and Information Gen-bank database.

Results

Our analysis identified 14,362 unique genomic variants from the two epidemics. The Uganda isolates had 5,740 unique variants, 75 of which had high impacts on the genomes. These were 51 frameshift, 15 stop gained, 5 stop lost, 2 missense, 1 synonymous and 1 stop lost and splice region. Their effects mainly occurred within the L-gene region at reference positions 17705, 11952, 11930 and 11027. For the DRC genomes, 8,622 variant sites were identified. The variants had a modifier effect on the genome occurring at reference positions, 213, 266 and 439. Examples are C213T, A266G and C439T. Phylogenetic reconstruction identified two separate and unique clusters from the two epidemics.

Conclusion

Our analysis provided further insights into the genetic diversity of *Bundibugyo Ebolavirus* from the two epidemics. The *Bundibugyo Ebolavirus* strain was genetically diverse with multiple variants. Phylogenetic reconstruction identified two unique variants. This signified an independent spillover event from a natural reservoir, rather a continuation from the ancestral outbreak that initiated the resurgence in DRC in the year 2012. Therefore, the two epidemics were not genetically related.

Keywords: Bundibugyo, Ebolavirus, RT-PCR, DRC, RNA, Viral Hemorrhagic Fever

Introduction

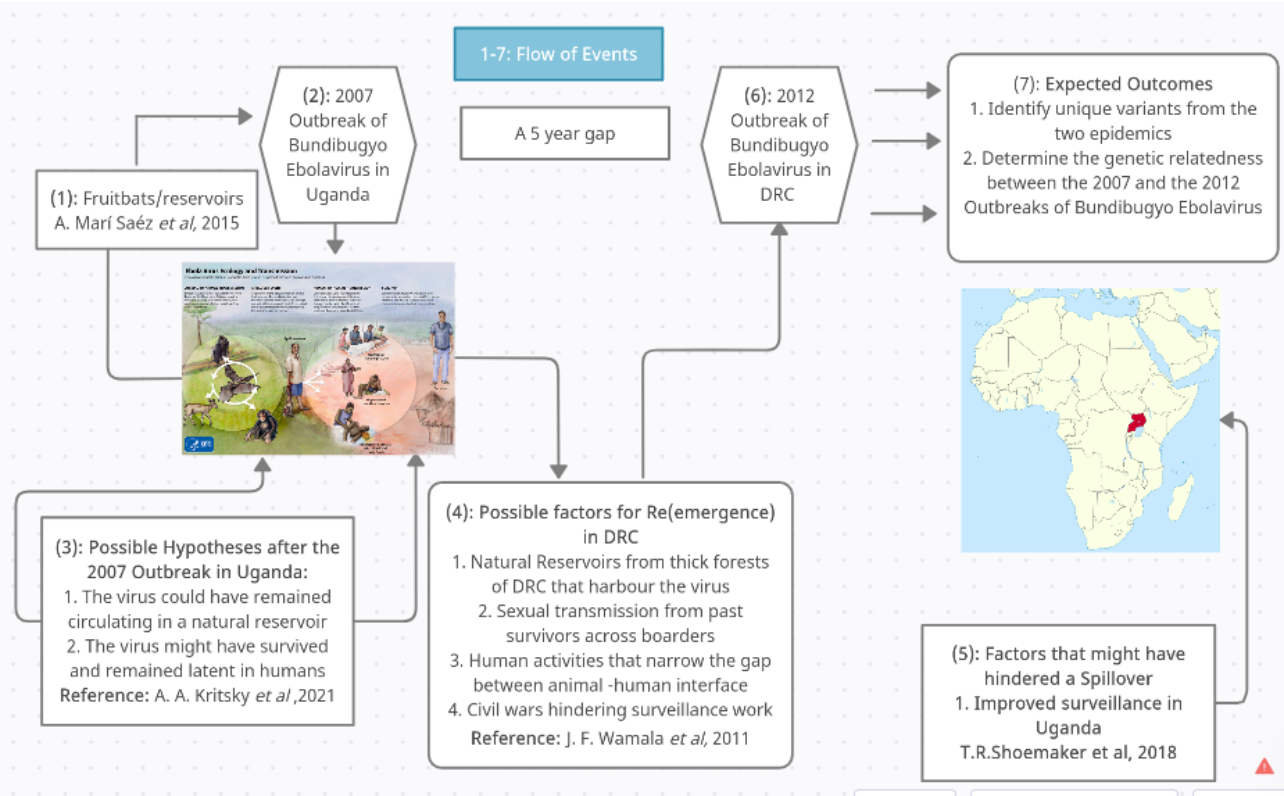
The Ebolavirus is one of the deadliest viral pathogens which was first discovered in the year 1976 during two consecutive outbreaks in the Democratic Republic of Congo (DRC) and Sudan(1). Since then, over 30 different outbreaks have been reported in Sub-Saharan Africa with an estimated 14,000 deaths and case fatality rates of up to 90% (2)(1). These viruses belong to the family Filoviridae and Genus Ebolavirus (2). There are six known strains in the genus Ebolavirus, all of which have a negative sense-single stranded RNA genome of approximately 18 -19 kilo base pairs (3). They include; *Zaire Ebolavirus*, *Sudan Ebolavirus*, *Bundibugyo Ebolavirus*, *Reston Ebolavirus*, *Tai Forest Ebolavirus* (4) and *Bombali Ebolavirus (5)*. The first three strains have been documented to cause severe illness and death in both humans and non-human primates with case fatality rates ranging from 40%-90% (6) (7). The *Reston* and *Tai Forest Ebolavirus* have not yet been discovered to cause human mortalities(1). Since its first discovery in the year 1976, there has been recurrent outbreaks of Ebolaviruses in Sub Saharan African countries(8)(9)(1)(10). With new cases reported almost every after five years in East and Central Africa for example, Uganda has reported seven different Ebolavirus outbreaks since the year 2000 and the DRC has recorded its 12th outbreak this year in February 2021(1)(11)(12). In particular, the B*undibugyo Ebolavirus* has a genome size of 18,940 base pairs and its RNA genome encodes seven structural proteins namely; Nucleoprotein (NP), two virion proteins (VP35 and VP40), a surface Glycoprotein (GP) and additional two virion proteins (VP30 and VP24). The genome also consists of an RNA- dependent, RNA polymerase (L) and a non-structural soluble protein (sGP proteins). The L gene codes for the RNA Polymerase, which is the most conserved region where as the VP40 virion protein is the most polymorphic gene in the Ebolavirus. The Bundibugyo Ebolavirus made its first appearance on the 1st August 2007, when there were reported cases of a viral hemorrhagic fever in Bundibugyo and Kikyo townships, a district in the western part of Uganda (11). This outbreak resulted into 116 human cases and 39 laboratory confirmed deaths (13). The index case was suspected to be a 26-year-old woman from Kabango village in Bundibugyo district. She presented with general weakness, fever and diarrhea after which she was hospitalized (13). Together with other suspect cases, blood samples were collected, sent to the Uganda Virus Research Institute (UVRI) and the US Centers for Disease Control and Prevention. Several laboratory investigations were performed and they confirmed on the 29th November 2007, a very unique and therefore novel strain of Ebolavirus, that was named Bundibugyo (13) (14). The Epidemiological data collected from this investigation found hunting spears near her home but hunting as a practice was denied.

In order to strengthen the response preparedness of the Viral Hemorrhagic Fever (VHF) in Uganda, the UVRI re-initiated the VHF National Surveillance programme in the year 2010 [34]. This was through an agreement between the Uganda Ministry of Health, Uganda Virus Research Institute and the US Centers for Disease Control and Prevention [32]. To date, UVRI serves as the national and regional reference laboratory for detection and response to VHF outbreaks which are of public health relevance in the region [35]. Currently, there is improved diagnostics to provide real time reporting of VHF cases detected. Laboratory diagnostic assays that have been implemented include; IgM and IgG ELISA for antigen-detection, RT-PCR as well as sequencing [36].

Since then, several research studies have been conducted for example; the work done by J.S. Towner *et al,* 2008 highlighted the high level of genetic diversity at amino acid level in the encoded virus proteins computing to over 27% and 35% for *Bundibugyo* and *Zaire Ebolavirus* respectively (11). Secondly, other research studies done elsewhere have also reported that variations in the Ebolavirus genome might have effects on the efficacy of virus detection at a sequence based level and design of candidate therapeutics (15). Thirdly, several years back, a research study that was conducted following two simultaneous occurrences of Ebolavirus in the DRC and South Sudan in 1976 (16), found a correlation between Ebolavirus disease and animal disease outbreaks (17). This is because Ebolavirus is transmitted by direct contact with the blood or any other secretions from animals or persons (18) (19). In addition, more recent studies that involve the Polymerase Chain Reaction (PCR) and antibody tests have identified cave-dwelling fruit bats as the possible natural reservoirs to most Ebolavirus strains (20) (21). Spillover events therefore occur when the animal and human interface is bridged through human activities such as; hunting wildlife for bush meat (22). This then sparks of epidemics which is most often followed by sustained human to human transmissions (23) (24). Despite all these research studies, five years after the ancestral outbreak was declared over in Uganda, the *Bundibugyo Ebolavirus* re-emerged and caused an epidemic for the first time in the DRC (25). This was reported by the WHO on the 17th August 2012 in Isiro Province (25) (26). The putative index case for this epidemic remains unidentified [14]. However, the earlier laboratory investigations using the RT-PCR assays confirmed, a clinic nurse in Isiro Province whose symptoms began on the 28th June 2012 (25). She reported with multiple potential exposures like human contact with other sick people, exposure to bats and as well she attended a funeral service (25). This outbreak resulted into 36 human cases with 13 laboratory confirmed deaths (26). Despite all these research studies, there is still limited scientific information available to explain the genetic diversity of this strain.

Further to this, in light of new evidence from the February 2021 outbreaks of Ebolavirus in the Republic of Guinea and the DRC, a new and unique paradigm or pattern for how these outbreaks spark off has been identified (27) (17). This new research findings suggests that the putative index cases leading to the resurgences of the February 2021 outbreaks are linked to contacts with survivors from past Ebolavirus outbreaks (28). Surprisingly, the previous outbreak of Ebolavirus in Guinea occurred 5-7 years ago at the time of the West African outbreak (29) (30). Whereas the resurgence in the DRC occurred a year after the 2020 outbreak was declared over (17) This cases have already raised important new research questions such as; "How do we need to change our response to escape from the cycle of outbreak-response-re-introduction-outbreak", "can new therapeutics be used to clear viruses from survivors" and the immediate question is, what these new findings mean for Ebolavirus survivors who are already faced with a lot of challenges (31). This therefore has created a need for reconsideration into local and scientific accounts of past Ebolavirus outbreaks (27) for example; the two epidemics of *Bundibugyo Ebolavirus* in Uganda in the year 2007 (11) and the DRC in the year 2012 (25). We therefore undertook a research study to get a better understanding of these two epidemics. Our main aim was to determine the genetic diversity of *Bundibugyo Ebolavirus* from Uganda and the Democratic Republic of Congo. The specific goals were to; i) To identify the unique variants in isolates of *Bundibugyo Ebolavirus* from the epidemics that occurred in Uganda and the Democratic Republic of Congo, ii) To determine the genetic relatedness between the *Bundibugyo Ebolavirus* outbreaks in Uganda (2007) and the Democratic Republic of Congo (2012). Ultimately, we aimed to determine whether the resurgence of *Bundibugyo Ebolavirus* in DRC was an independent spillover event from nature or a continuation from the ancestral outbreak, possibly through contacts with past survivors.

Fig 1: The working hypotheses for the resurgence in DRC in the year 2012



Materials and Methods

Study Design

This was a retrospective descriptive study. The source organism in our analysis was the *Bundibugyo Ebolavirus* strain, which has a genome size of 18,940 base pairs (32). We used publicly available sequence data that was retrieved from the National Center for Biotechnology and Information Gen-bank database (33) The NCBI, has the Sequence Read Archive (SRA) repository and the Nucleotide sequence database (34). The SRA is the largest publicly available repository having raw sequence data from high throughput sequencers (35). The 31 raw sequence data which represented isolates collected from the epidemic in Uganda was retrieved from this repository. Whereas the Nucleotide sequence database has assembled genomes deposited from different experiments (36). The 4 nucleotide sequences that represented isolates from the DRC in our analysis was retrieved from this database.

Sample Size Determination

The isolates were; 31 fastq sequences, 6 fasta sequences from the Uganda outbreak in the year 2007. The isolates from the DRC outbreak were represented by the 4 fasta sequences that were retrieved from the nucleotide sequence database (36). The table 1 below shows the characteristics of the isolates which the DRC sequences were generated (25)

Table 1: Patient and Sample characteristics from the DRC outbreak in the year 2012

Case ID	Gene-bank Number	Demographics	Occupation	Sampling Location	Clinical Status
112	KC545393	44/F	homemaker	Isiro Province	deceased
120	KC545394	77/M	unknown	Vungba	deceased
122	KC545395	Unknown	unknown	Unknown	survived
37	KC545396	18/F	student	Isiro Province	deceased

F, female; M, Male

Bioinformatics Analysis

The fastq-dump tool (37) was used to download all the sequence data from the NCBI database. This included the sequences with their metadata which were all stored in a High-Performance Computing (HPC) server at the African Center of Excellence in Bioinformatics and Data Intensive Sciences (38). Quality Assessment was not performed on the DRC sequences. This is because, they were assembled genomes. However for the 31 raw sequence data (fastq format) collected from the Uganda outbreak in the year 2007, a quality control check was performed comprehensively in order to ensure they were of good quality before downstream analysis (39). This assessment was performed using tools; Fast-QC (v0.11.9) (40) and Multi-QC (v1.9) (41) respectively. The low

quality regions were then trimmed, including adapter sequences setting the phred threshold at 20 (42).

Variant Analysis

For the Uganda genomes, the quality filtered raw sequence data were referenced mapped against a reference genome using the Burrow's Wheeler Aligner tool (0.7.17-r1188) (43) The reference genome used was an isolate from the 2007 outbreak in Uganda (Gene-bank accession number FJ217161). This isolate was used because it has a complete genome size of 18,940 base pairs and it is also an original isolate from the ancestral outbreak. Variants were then called using freebayes tool (v1.3.1-dirty) (44)(45). This tool has advantages over other tools because, it is haplotype based, also a Bayesian genetic variant detector and outputs a variant call format (VCF) file, which consists of small polymorphisms specifically SNPs (single-nucleotide polymorphisms), Indels (Insertions and Deletions), MNPs (multi-nucleotide polymorphisms) (45). We then used SnpEff tool to perform variant annotation (46). This tool predicts the functional effect of the variants on proteins or amino acid changes (47).

To annotate variants, a database from the reference genome has to be built. This was performed using "SnpEff build" tool. To create the SnpEff database, we downloaded sequence data from NCBI for the reference genome of *Bundibugyo Ebolavirus* with accession number, FJ217161. We also downloaded the corresponding General Feature Format (GFF) file, which contains the annotations and the FASTA file, with its entire genome (48) The SnpEff tool was then used to annotate variants. Once the analysis was executed, the annotation data was outputted as an annotated Variant Call Format (VCF) and an HTML report file containing all the summary statistics for the different variants (46). In addition, Python v3.6.3 (49) was then used to construct a bar plot to show the frequency of unique variants with high impacts on the genome.

On the other hand, all the DRC sequences including an isolate of the *Bundibugyo Ebolavirus* from the 2007 outbreak as the reference sequence (Gene-bank accession number FJ217161) were concatenated in to a single multi-fasta file and saved as a FASTA format. This reference sequence was used in order to determine how the variants from the 2012 outbreak were phylogenetically distinct from the 2007 outbreak in Uganda. Multiple Sequence Alignment was performed on the fasta sequences using MAFFT v7.310 tool (50). After this, variants were then called using the alignment FASTA file as input and the SNP extraction tool, SNP-sites v2.3.3 (51). This tool restructures the aligned data as a Variant Call Format (VCF) file. This VCF file provides a clear

mapping of SNPs from the aligned sequences. This then allowed easy identification of the SNP location and the genotype for each sample at a given locus (52). In the outputted VCF file, the rows correspond with each unique variant and the columns provides the genotype at that given site (53). A summary of the SNPs relative to the reference sequence was then visualized using the snipit tool (https://github.com/aineniamh/snipit) and SnpEff tool was used to annotate the variants. Using different bash scripts, a report showing the effect of the variants on the different sequences was extracted (54)

Phylogenetic Analysis to determine the genetic relatedness between the two Epidemics

All the quality filtered raw sequence data from Uganda, were assembled using both SPades v3.13.1 (55) and abyss 1.9.0 assemblers (56) (57) in order to obtain a consensus sequence. These two genome assemblers are best suited for assembly of short paired end reads (58). A draft scaffold was then obtained with the use of SSPACE tool (59). This is a standalone tool and was used for scaffolding the paired end reads. It enabled read orientation into connected sequences by allowing mean values and standard deviations of the insert sizes for each read library (59). GapFiller tool was then used to find and fill gaps generated in the contiguous sequence (60). All the obtained fasta sequences from both Uganda and the DRC were concatenated in to a single FASTA file including the reference sequence. Multiple Sequence Alignment was then performed using MAFFT v7.310 (50) and manually checked in Ali-view v.1.27(61). The 5' and the 3' untranslated regions were then trimmed to remove any remaining gaps. The maximum-likelihood phylogenetic tree was then constructed using IQ-TREE (62) and Phyml (63) and the best suited substitution model was determined and run for 1000 replicates. The resulting newick file was uploaded to the interactive tree of life, iTOL v4.0(64), which is an online tool for phylogenetic tree visualization. The tree was rooted at mid-point to split variants from Uganda and the Democratic Republic of Congo.

RESULTS

Variant Analysis

In the Uganda genomes, 37 sequences of *Bundibugyo Ebolavirus* were analyzed. We identified 5,740 distinct genome variants and they are recorded in table 2 below. Generally, the variants were distributed according to the different regions (downstream, upstream, exon intergenic, 3 and 5 prime Untranslated Regions (UTR). However, majority of the variants were found downstream and upstream regions of the genome (non-coding regions). The viral sequences showed multiple diversity with most variants occurring at reference positions 17705, 11952, 11930 and 11027 appearing in most of the isolates collected from this outbreak.

Table 2: **The frequency of each unique variant type in isolates of the *Bundibugyo Ebolavirus* collected from the 2007 outbreak in Uganda.**

Number of Effects by Type	
Annotation	**Counts**
downstream_gene_variant	2,375
upstream_gene_variant	1,945
missense_variant	543
synonymous_variant	284
3_prime_UTR_variant	268
intergenic_region	103
5_prime_UTR_variant	79
frameshift_variant	68
stop_gained	57
splice_region_variant	7
stop_lost	5
5_prime_UTR_premature_start_codon_gain_variant	3
conservative_in-frame_deletion	2
stop_retained_variant	1
Total number of unique genomic variants	5,740

In addition, we then identified 75 unique variants of high impacts on the genome. They were; 51 frameshift, 15 stop gained, 5 stop lost, 2 missense, 1 synonymous and 1 stop lost and splice region. Among these variants, the most common impacts were majorly frame-shifts and stop gained. They include; T**A**T17705TT, GAAAAAATTTTG11952GAAAAAA**A**TTTTG, G11930T, CAAAAAACCCG11027CAAAAAA**A**CCCG. Their effects occurred mostly on the L-gene region of the *Bundibugyo Ebolavirus*. Refer to S2 in Appendix for the supporting information showing a table indicating the frequency of unique variants which had high impacts on the genome.

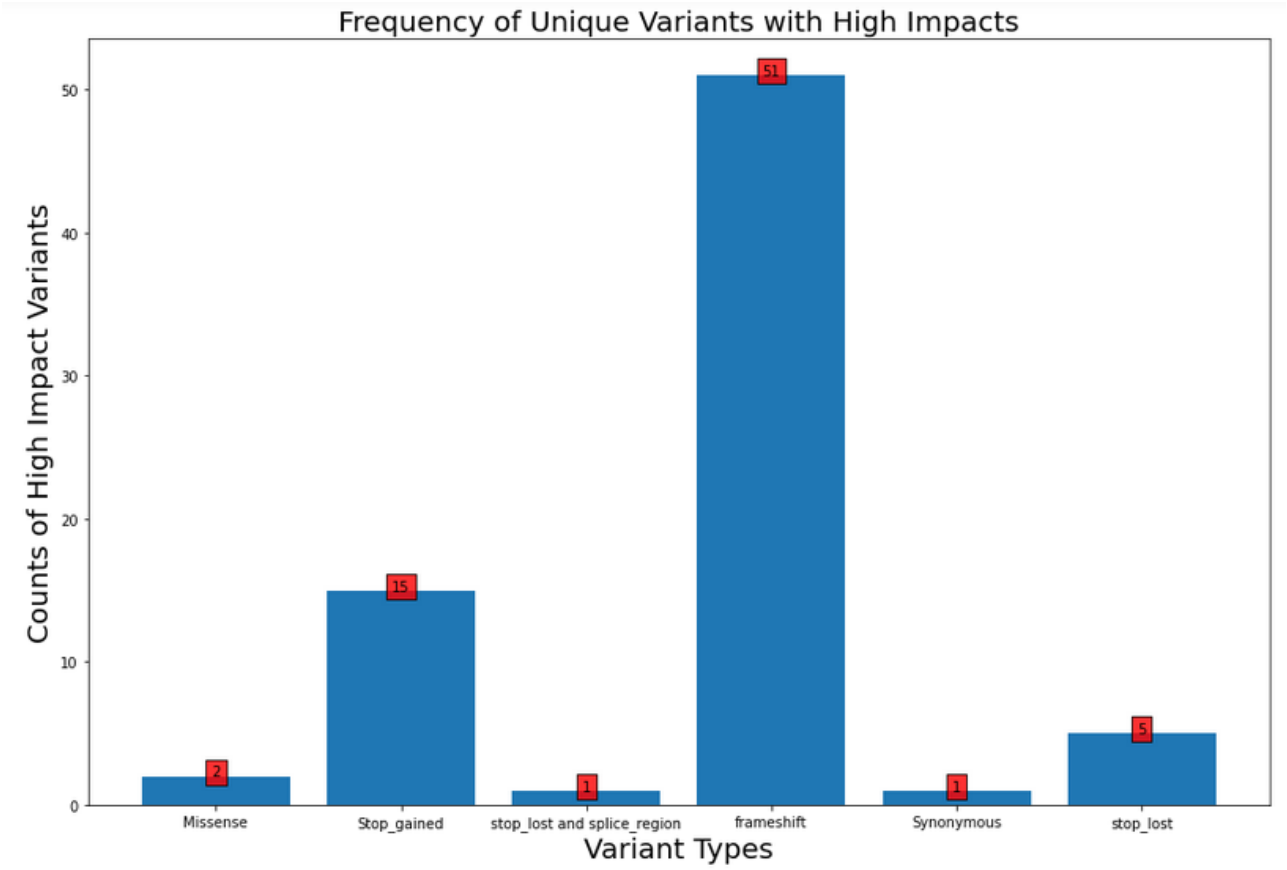


Fig 2: A Bar-plot showing the frequency of unique variants with high impact on the genomes of Bundibugyo Ebolavirus isolated from the 2007 outbreak in Uganda

On the other hand, we identified 8,622 nucleotide variant sites from the isolates of *Bundibugyo Ebolavirus* collected from the 2012 outbreak. The variants identified here all had a modifier effect on the genome. This effect was predetermined by the variant type in each of the isolates. Some of them include; C213T, A266G and C439T. Fig 3 below shows the different nucleotide variant sites in the DRC sequences relative to the reference sequence with Gene-bank accession number of FJ217161. This was a complete genome and isolated from the 2007 outbreak. The purpose of using this as a reference sequence was to find out how the sequences from the 2012 outbreak in the DRC were genetically distinct and unique from the ancestral outbreak of 2007 which occurred in Uganda.

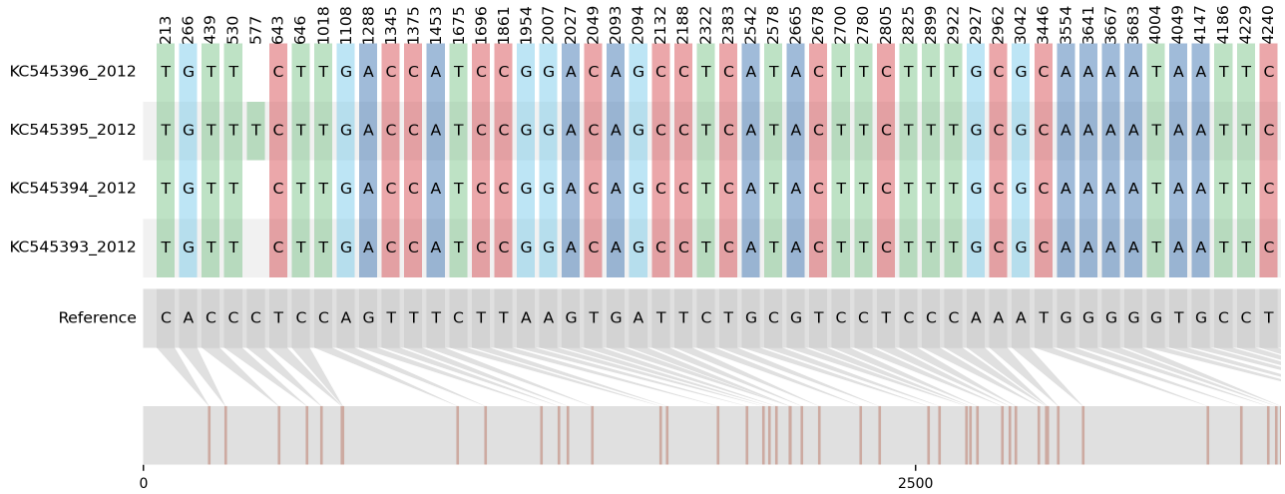


Fig 3: Nucleotide alignment showing variant sites in the four sequences relative to the reference genome. The 4 sequences represent isolates collected from the DRC outbreak in 2012 (NCBI accession numbers: KC545393, KC545394, KC545395, KC545396). The reference sequence is an isolate from the ancestral outbreak (Gene-bank accession number: FJ217161)

Phylogenetic Analysis to determine the genetic relatedness between the two Epidemics

When the tree in Fig 4 below was rooted at mid-point, two separate and unique clusters were identified from these two epidemics. Phylogenetic reconstruction demonstrates that the 4 sequences isolated from the outbreak in DRC cluster uniquely and distant from those of the 2007 outbreak in Uganda. This signify a separate variant and basing on our analysis, we identified approximately 8,622 mutations from the 4 DRC sequences, which is almost double the number of mutations identified from the ancestral outbreak in Uganda (5,740)

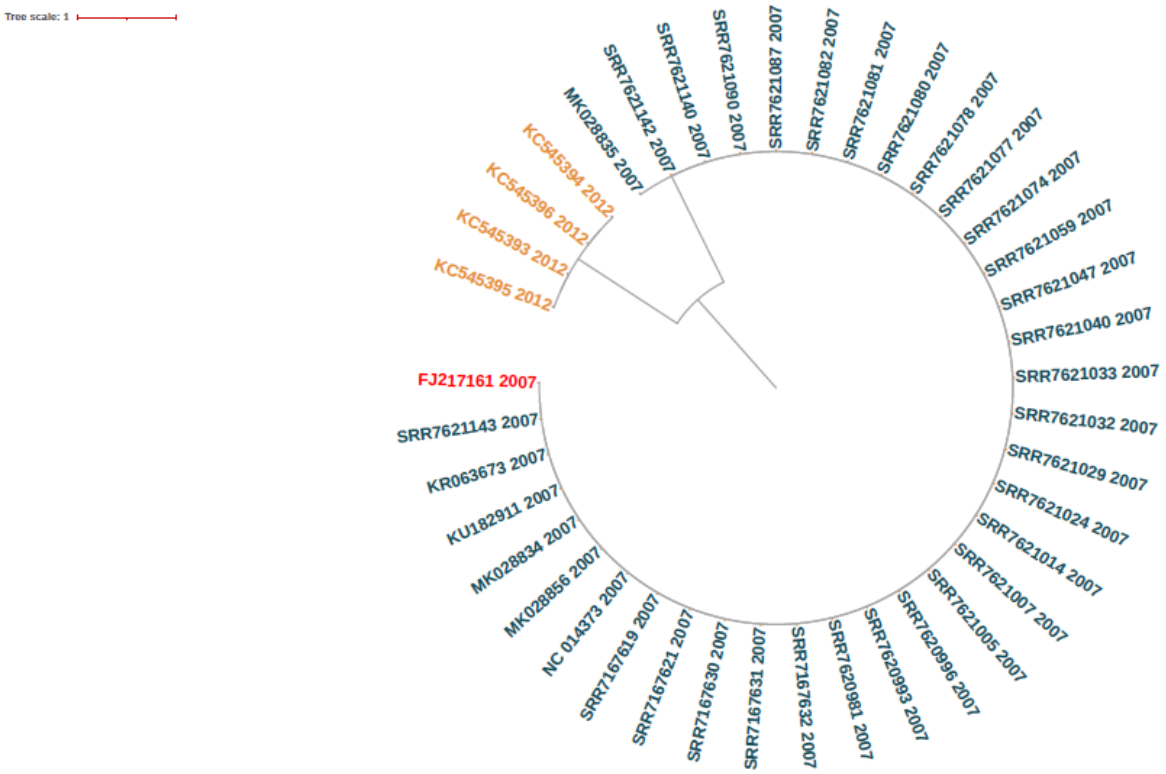


Fig 4: Maximum likelihood phylogenetic tree showing the two unique clusters identified from the two epidemics that occurred in Uganda (2007) and the Democratic Republic of Congo (2012). The reference sequence used was an isolate from the ancestral outbreak having a Gene-bank accession number of FJ217161

Discussion

We generally aimed to determine the genetic diversity of *Bundibugyo Ebolavirus* from the epidemics that occurred in Uganda in the year 2007 and the DRC in the year 2012. Our specific goal was to identify the unique variants in isolates collected from these two epidemics. This was to ultimately enable us determine the genetic relatedness between two epidemics.

Our analysis identified in total, 14,362 unique genomic variants from the two epidemics. The high impacts variants in the Uganda genomes were mainly frame shifts, stop gained and missense mutations. A frameshift is a genetic variant that changes the way codons are read during the process of creating an amino acid sequence (65). This variant is due to an insertion or deletion of a nucleotide (66). This is of significance because cells read a gene in groups of three bases. The three bases here correspond to one of 20 different amino acids that is used to build a protein (67). Therefore, if a mutation disrupts this reading frame, then the sequence of DNA that follows the mutation will be read incorrectly (68). With a stop gained variant, the mutation leads to changes of at least one base of a codon, hence a premature stop codon (69). This results in a premature stop of translation of messenger RNA in to a protein hence a non-functional or unstable protein (52). In addition, a missense mutation is a genetic change where a single pair of substitution alters the genetic code leading to production of a new amino acid (70). Most of these variant effects occurred within the L-gene region of the *Bundibugyo Ebolavirus*. Since the L-gene is the most conserved region and is a target for primer designs (71), the findings from our analysis is in line with a previous study conducted after the 2007 outbreak (11). Where sequence analysis of the PCR fragment from the virus L-gene revealed the initial failure of real-time RT-PCR assays, since the viral sequences were divergent from the four already known strains of Ebolavirus (11). Therefore, these alterations of the genetic code or disruption of one reading frame could have resulted in to the formation of a new strain of Ebolavirus and hence our findings supports this past study. The high frequency of frameshift variants is suggestive of a new strain (52), the *Bundibugyo Ebolavirus,* which was a novel strain that first emerged in the year 2007 in Uganda (11).

On the other hand, 8,622 nucleotide variant sites were identified from the DRC genomes. The variants had a modifier effect on the genome. This effect was predetermined by the type of variant identified in these isolates (52). Modifier variants are genes that alter the phenotypic outcomes and results in to altered effects or impacts (72). The phenotypic outcomes here could include; dominance, expression and penetrance (73). Naturally, viruses accumulate mutations over time

which may arise from adaptations in response to environmental changes or immune responses of the host reservoirs (74). Sometimes viruses transmit and persists after fixing beneficial mutations that

would allow it to better exploit it's host or other new hosts (75). This scenario could explain the re-emergence of the *Bundibugyo Ebolavirus*. That is, after the ancestral outbreak was declared over in the year 2007, this virus might have undergone an evolutionary change over a period of five years in a certain natural reservoir, generating variations in it's genome. This resulted into a separate and unique variant that was responsible for the 2012 outbreak in the DRC (76). The ultimate high frequency of modifier effects on the genome is an indicator and possibly explains, the divergence or formation of a new variant that was unique from the ancestral type (25)(26).

Phylogenetic trees help in our understanding of the evolutionary relationships between groups (77). In our context, we used it to determine the genetic relatedness between the epidemics that occurred in Uganda in the year 2007 and the DRC in the year 2012. Phylogenetic reconstruction in Fig 4 demonstrates that the 4 sequences from the 2012 outbreak in DRC cluster together and are similar but distantly related from those of the ancestral outbreak (78). This signify a new variant and basing on our analysis, we identified approximately 8,622 mutations from the 4 DRC sequences, which is almost double the number of mutations identified from the ancestral outbreak in Uganda (5,740) (79). This is indicative of viral evolution over the period of five years (80). In other words, the frequency of SNPs or mutations occurrence in a genome under the conditions of a survivor organism is reduced by a big magnitude compared to that from a host reservoir (81). This is because the virus under goes a period of latency in a human survivor (82). Therefore, these two separate variants indicate that the 2012 outbreak in DRC was a new introduction or an independent spillover event from a certain animal reservoir, rather a human transmission from a contact with a past survivor.

Our study however had limitations such as; limited sampling which led to less sequence data generated from the 2012 outbreak. Some patient demographics were unknown, this hindered our understanding in to the Epidemiology and Molecular findings. This led to uncertainty in drawing conclusions on the genetic diversity of *Bundibugyo Ebolavirus* from the 2012 outbreak in the DRC. For example; in variant analysis and phylogenetic estimations. The availability of more or complete genomes from the DRC outbreak in 2012 would improve the study of transmission dynamics between these two epidemics as well as identification of multiple key SNP's that can promote the study of *Bundibugyo Ebolavirus* pathogenesis.

In conclusion, our analysis provided further insights into the genetic diversity of Bundibugyo Ebolavirus from the two epidemics. Variant characterization can be used in the fight against Bundibugyo Ebolavirus and the development of effective treatments or vaccines. This is because key SNPs have been identified and can be used for further research about the pathogenesis of *Bundibugyo Ebolavirus.* The findings from our study has also provided knowledge on the likely origin or how the 2012 outbreak in the DRC was initiated. Phylogenetic reconstruction identified two unique variants. This signified an independent spillover event from a natural reservoir, rather a continuation from the ancestral outbreak that initiated the resurgence in the DRC in the year 2012. Therefore, the two epidemics are not genetically related.

Abbreviations

BDBV: Bundibugyo Ebolavirus, RT-PCR: Reverse Transcription Polymerase Chain Reaction, DRC: Democratic Republic of Congo, SNP: Single Nucleotide Polymorphism, IgM: Immunoglobulin M, IgG: Immunoglobulin G, UVRI: Uganda Virus Research Institute

Acknowledgments

I would like to extend my sincere gratitude to the department of Immunology and Molecular Biology, College of Health Sciences in Makerere University, for the training leading to the award of a Master of Science in Bioinformatics. Special thanks to the department of Arbovirology, Emerging and Re-emerging Infectious Diseases at the Uganda Virus Research Institute. They offered financial support to facilitate my studies. Finally, I would like to extend my appreciation to the MRC/UVRI and LSHTM Uganda Research Unit for an offer of a Manuscript Mentorship Programme that eventually facilitated the submission of this master's research project work for publication.

Supporting Information

S1 Appendix: List of the Gene-bank identifiers for the sequences that were used in our analysis

S2 Appendix: The frequency of unique variants which had high impact on the genomes

Ethical Clearance

This research project was approved by the School of Biomedical Sciences Research and Ethics Committee (SBSREC). This is an institutional review board found within the College of Health Sciences in Makerere University. The protocol number was SBS-2021-64

Data Availability

There was no funding for this project. We used publicly available sequence data that was retrieved from the National Center for Biotechnology and Information (NCBI). Below are the links.

Raw sequence data: https://www.ncbi.nlm.nih.gov/sra/?term=Bundibugyo+Ebolavirus+in+Uganda

Assembled genomes: https://www.ncbi.nlm.nih.gov/nuccore/?term=Bundibugyo+Ebolavirus

References

1. Rugarabamu S, Mboera L, Rweyemamu M, Mwanyika G, Lutwama J, Paweska J, et al. Forty-two years of responding to Ebola virus outbreaks in Sub-Saharan Africa: A review. BMJ Glob Heal. 2020;5(3):1–10.

2. Majid MU, Tahir MS, Ali Q, Rao AQ, Rashid B, Ali A, et al. Nature and history of Ebola virus: an overview. Arch Neurosci. 2016;3(3):e35027.

3. Dietzel E, Schudt G, Krähling V, Matrosovich M, Becker S. Functional Characterization of Adaptive Mutations during the West African Ebola Virus Outbreak. J Virol. 2017;91(2).

4. Carroll SA, Towner JS, Sealy TK, McMullan LK, Khristova ML, Burt FJ, et al. Molecular evolution of viruses of the family Filoviridae based on 97 whole-genome sequences. J Virol. 2013;87(5):2608–16.

5. Goldstein T, Anthony SJ, Gbakima A, Bird BH, Bangura J, Tremeau-Bravard A, et al. The discovery of Bombali virus adds further support for bats as hosts of ebolaviruses. Nat Microbiol. 2018;3(10):1084–9.

6. Chippaux JP. Outbreaks of Ebola virus disease in Africa: The beginnings of a tragic saga. J Venom Anim Toxins Incl Trop Dis. 2014;20(1):1–14.

7. Chippaux J-P. Outbreaks of Ebola virus disease in Africa: the beginnings of a tragic saga. J Venom Anim Toxins Incl Trop Dis. 2014;20(1):44.

8. Albariño CG, Shoemaker T, Khristova ML, Wamala JF, Muyembe JJ, Balinandi S, et al. Genomic analysis of filoviruses associated with four viral hemorrhagic fever outbreaks in Uganda and the Democratic Republic of the Congo in 2012. Virology [Internet]. 2013;442(2):97–100. Available from: http://dx.doi.org/10.1016/j.virol.2013.04.014

9. Li X, Zai J, Liu H, Feng Y, Li F, Wei J, et al. The 2014 Ebola virus outbreak in West Africa highlights no evidence of rapid evolution or adaptation to humans. Sci Rep [Internet]. 2016;6(October):1–9. Available from: http://dx.doi.org/10.1038/srep35822

10. Shoemaker T, MacNeil A, Balinandi S, Campbell S, Wamala JF, McMullan LK, et al. Reemerging Sudan ebola virus disease in Uganda, 2011. Emerg Infect Dis. 2012;18(9):1480.

11. Towner JS, Sealy TK, Khristova ML, Albariño CG, Conlan S, Reeder SA, et al. Newly discovered Ebola virus associated with hemorrhagic fever outbreak in Uganda. PLoS Pathog. 2008;4(11):3–8.

12. Nsio J, Kapetshi J, Makiala S, Raymond F, Tshapenda G, Boucher N, et al. 2017 Outbreak of Ebola Virus Disease in Northern Democratic Republic of Congo. J Infect Dis. 2020;221(5):701–6.

13. Wamala JF, Lukwago L, Malimbo M, Nguku P, Yoti Z, Musenero M, et al. Ebola hemorrhagic fever associated with novel virus strain, Uganda, 2007-2008. Emerg Infect Dis. 2010;16(7):1087–92.

14. Towner JS, Sealy TK, Khristova ML, Albariño CG, Conlan S, Reeder SA, et al. Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. PLoS Pathog. 2008;4(11):e1000212.

15. Carneiro J, Pereira F. EbolaID: an online database of informative genomic regions for Ebola identification and treatment. PLoS Negl Trop Dis. 2016;10(7):e0004757.

16. Bowen ETW, Platt GS, Lloyd G, Raymond RT, Simpson DIH. A comparative study of strains of Ebola virus isolated from southern Sudan and northern Zaire in 1976. J Med Virol. 1980;6(2):129–38.

17. Vivalya BM, Piripiri AL, Mbeva JBK. The resurgence of Ebola disease outbreak in North-Kivu: viewpoint of the health system in the aftermath of the outbreak in the Democratic Republic of Congo. PAMJ-One Heal. 2021;5(5).

18. Judson S, Prescott J, Munster V. Understanding ebola virus transmission. Viruses. 2015;7(2):511–21.

19. Rewar S, Mirdha D. Transmission of Ebola virus disease: an overview. Ann Glob Heal. 2014;80(6):444–51.

20. Ogawa H, Miyamoto H, Nakayama E, Yoshida R, Nakamura I, Sawa H, et al. Seroepidemiological prevalence of multiple species of filoviruses in fruit bats (Eidolon helvum) migrating in Africa. J Infect Dis. 2015;212(suppl_2):S101–8.

21. Changula K, Kajihara M, Mori-Kajihara A, Eto Y, Miyamoto H, Yoshida R, et al. Seroprevalence of filovirus infection of Rousettus aegyptiacus bats in Zambia. J Infect Dis. 2018;218(suppl_5):S312–7.

22. Johnson CK, Hitchens PL, Evans TS, Goldstein T, Thomas K, Clements A, et al. Spillover and pandemic properties of zoonotic viruses with high host plasticity. Sci Rep. 2015;5(1):1–8.

23. Wood JLN, Leach M, Waldman L, MacGregor H, Fooks AR, Jones KE, et al. A framework for the study of zoonotic disease emergence and its drivers: spillover of bat pathogens as a case study. Philos Trans R Soc B Biol Sci. 2012;367(1604):2881–92.

24. Kock RA, Begovoeva M, Ansumana R, Suluku R. Searching for the source of Ebola: the elusive factors driving its spillover into humans during the West African outbreak of 2013–2016. OIE Sci Tech Rev. 2019;38(1):113–7.

25. Hulseberg CE, Kumar R, Di Paola N, Larson P, Nagle ER, Richardson J, et al. Molecular analysis of the 2012 Bundibugyo virus disease outbreak. Cell Reports Med. 2021;2(8):100351.

26. Kratz T, Roddy P, Tshomba Oloma A, Jeffs B, Pou Ciruelo D, de la Rosa O, et al. Ebola virus disease outbreak in Isiro, Democratic Republic of the Congo, 2012: signs and symptoms, management and outcomes. PLoS One. 2015;10(6):e0129333.

27. Fairhead J, Leach M, Millimouno D. Spillover or endemic? Reconsidering the origins of Ebola virus disease outbreaks by revisiting local accounts in light of new evidence from Guinea. BMJ Glob Heal. 2021;6(4):e005783.

28. Keita AK, Düx A, Diallo H, Calvignac-Spencer S, Sow MS, Keita MB, et al. Resurgence of Ebola virus in guinea after 5 years calls for careful attention to survivors without creating further stigmatization. Virological. 2021;

29. Marí Saéz A, Weiss S, Nowak K, Lapeyre V, Zimmermann F, Düx A, et al. Investigating the zoonotic origin of the West African Ebola epidemic. EMBO Mol Med. 2015;7(1):17–23.

30. Spengler JR, Ervin ED, Towner JS, Rollin PE, Nichol ST. Perspectives on West Africa Ebola virus disease outbreak, 2013–2016. Emerg Infect Dis. 2016;22(6):956.

31. Kupferschmidt K. New Ebola outbreak likely sparked by a person infected 5 years ago. Science (80-). 2021;

32. Oluwagbemi O, Awe O. A comparative computational genomics of Ebola Virus Disease strains: In-silico Insight for Ebola control. Informatics Med Unlocked. 2018;12:106–19.

33. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, et al. GenBank. Nucleic Acids Res. 2018;46(D1):D41–7.

34. Leinonen R, Sugawara H, Shumway M, Collaboration INSD. The sequence read archive. Nucleic Acids Res. 2010;39(suppl_1):D19–21.

35. Kodama Y, Shumway M, Leinonen R. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40(D1):D54–6.

36. Mizrachi I. Genbank: the nucleotide sequence database. NCBI Handb [Internet], Updat. 2007;22.

37. Schmid MW. Rcount: User Guide. 2014;

38. Mboowa G, Sserwadda I, Aruhomukama D. Genomics and bioinformatics capacity in Africa: no continent is left behind. Genome. 2021;64(5):503–13.

39. Tong Y-G, Shi W-F, Liu D, Qian J, Liang L, Bo X-C, et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. Nature. 2015;524(7563):93–6.

40. Andrews S. Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data. URL https//www bioinformatics babraham ac uk/projects/fastqc/[Google Sch. 2010;

41. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8.

42. Krueger F, Andrews SR. Quality control, trimming and alignment of Bisulfite-Seq data (Prot 57). Dep Med Hematol Oncol Domagkstr. 2012;3(48149):1–13.

43. Hansen NF. Variant calling from next generation sequence data. In: Statistical Genomics. Springer; 2016. p. 209–24.

44. Garrison E, Marth G. FreeBayes. Marth Lab. 2010;

45. Mohammed KS, Kibinge N, Prins P, Agoti CN, Cotten M, Nokes DJ, et al. Evaluating the performance of tools used to call minority variants from whole genome short-read data. Wellcome open Res. 2018;3.

46. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92.

47. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. 2012;

48. Pertea G, Pertea M. GFF utilities: GffRead and GffCompare. F1000Research. 2020;9.

49. Consortium A gambiae 1000 G. Genetic diversity of the African malaria vector Anopheles gambiae. Nature. 2017;552(7683):96.

50. Van Borm S, Vanneste K, Fu Q, Maes D, Schoos A, Vallaey E, et al. Increased viral read counts and metagenomic full genome characterization of porcine astrovirus 4 and Posavirus 1 in sows in a swine farm with unexplained neonatal piglet diarrhea. Virus Genes. 2020;56(6):696–704.

51. Pakistan HI V. Public health round-up. Bull World Heal Organ. 2019;97:517–8.

52. Bindayna KM, Crinion S. Variant analysis of SARS-CoV-2 genomes in the Middle East. Microb Pathog. 2021;153:104741.

53. Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. Gene reports. 2020;19:100682.

54. Mishra D, Khandelwal G. Command-Line Tools in Linux for Handling Large Data Files. In: Bioinformatics: Sequences, Structures, Phylogeny. Springer; 2018. p. 375–92.

55. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes de novo assembler. Curr Protoc Bioinforma. 2020;70(1):e102.

56. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19(6):1117–23.

57. Liu Y, Schmidt B, Maskell DL. Parallelized short read assembly of large genomes using de Bruijn graphs. BMC Bioinformatics. 2011;12(1):1–10.

58. Paszkiewicz K, Studholme DJ. De novo assembly of short sequence reads. Brief Bioinform. 2010;11(5):457–72.

59. Boetzer M, Henkel C V, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27(4):578–9.

60. Nadalin F, Vezzi F, Policriti A. GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics. 2012;13(14):1–16.

61. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014;30(22):3276–8.

62. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.

63. Guindon S, Delsuc F, Dufayard J-F, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. In: Bioinformatics for DNA sequence analysis. Springer; 2009. p. 113–37.

64. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47(W1):W256–9.

65. Chen J, Wu J-S, Mize T, Moreno M, Hamid M, Servin F, et al. A frameshift variant in the CHST9 gene identified by family-based whole genome sequencing is associated with schizophrenia in Chinese population. Sci Rep. 2019;9(1):1–9.

66. Rausell A, Mohammadi P, McLaren PJ, Bartha I, Xenarios I, Fellay J, et al. Analysis of stop-gain and frameshift variants in human innate immunity genes. PLoS Comput Biol. 2014;10(7):e1003757.

67. Berg JM. Amino Acids Are Encoded by Groups of Three Bases Starting from a Fixed Point. 1970.

68. Yourno J, Heath S. Nature of the hisD3018 frameshift mutation in Salmonella typhimurium. J Bacteriol. 1969;100(1):460–8.

69. Cirulli ET, Heinzen EL, Dietrich FS, Shianna K V, Singh A, Maia JM, et al. A whole-genome analysis of premature termination codons. Genomics. 2011;98(5):337–42.

70. Gorlov IP, Pikielny CW, Frost HR, Her SC, Cole MD, Strohbehn SD, et al. Gene characteristics predicting missense, nonsense and frameshift mutations in tumor samples. BMC Bioinformatics. 2018;19(1):1–14.

71. Hammou RA, Kasmi Y, Khataby K, Laasri FE, Boughribil S, Ennaji MM. Roles of VP35, VP40 and VP24 Proteins of Ebola Virus in Pathogenic and Replication Mechanisms. CRTOMIR P, Ebola, Croácia Intechopen. 2016;101–17.

72. Davidson BA, Hassan S, Garcia EJ, Tayebi N, Sidransky E. Exploring genetic modifiers of Gaucher disease: The next horizon. Hum Mutat. 2018;39(12):1739–51.

73. Nadeau JH. Modifier genes in mice and humans. Nat Rev Genet. 2001;2(3):165–74.

74. Carroll SA, Towner JS, Sealy TK, McMullan LK, Khristova ML, Burt FJ, et al. Molecular Evolution of Viruses of the Family Filoviridae Based on 97 Whole-Genome Sequences. J Virol. 2013;87(5):2608–16.

75. Agudelo-Romero P, Carbonell P, Perez-Amador MA, Elena SF. Virus adaptation by manipulation of host's gene expression. PLoS One. 2008;3(6):e2397.

76. Munson MA, Banerjee A, Watson TF, Wade WG. Molecular analysis of the microflora associated with dental caries. J Clin Microbiol. 2004;42(7):3023–9.

77. Ladner JT, Wiley MR, Mate S, Dudas G, Prieto K, Lovett S, et al. Evolution and Spread of Ebola Virus in Liberia, 2014-2015. Cell Host Microbe [Internet]. 2015;18(6):659–69. Available from: http://dx.doi.org/10.1016/j.chom.2015.11.008

78. Pereira-Gomez M, Lopez-Tort F, Fajardo A, Cristina J. An evolutionary insight into emerging Ebolavirus strains isolated in Africa. J Med Virol. 2020;92(8):988–95.

79. Bosworth A, Rickett NY, Dong X, Ng LFP, García-Dorival I, Matthews DA, et al. Analysis of an Ebola virus disease survivor whose host and viral markers were predictive of death indicates the effectiveness of medical countermeasures and supportive care. Genome Med. 2021;13(1):1–18.

80. Emanuel J, Marzi A, Feldmann H. Filoviruses: ecology, molecular biology, and evolution. Adv Virus Res. 2018;100:189–221.

81. Kritsky AA, Keita S, Magassouba N, Krasnov YM, Safronov VA, Naidenova E V, et al. Ebola virus disease outbreak in the Republic of Guinea 2021: hypotheses of origin. bioRxiv. 2021; 82. Boseley S. Pauline Cafferkey: dedicated nurse and reluctant Ebola hero. Lancet. 2016;388(10043):455.

Supporting Information

S1 Appendix: List of the Gene-bank identifiers for the sequences that were used in our analysis

Gene-bank ID's for the Uganda outbreak in 2007		Gene-bank ID's for DRC outbreak in 2012
SRR7621005 SRR7621077	SRR7621144	KC545393
SRR7621007 SRR7621078	SRR7167631	KC545394
SRR7621014 SRR7621080	SRR7620993	KC545395
SRR7621024 SRR7621081	SRR7620996	KC545396
SRR7621029 SRR7621082	NC_014373	
SRR7621032 SRR7621087	MK028856	
SRR7621040 SRR7621090	MK028834	
SRR7621047 SRR7621140	KU182911	
SRR7621059 SRR7621142	KR063673	
SRR7621074 SRR7621143	MK028835	

S2 Appendix: The frequency of unique variants which had high impact on the genomes

Gene-bank Identifiers	Position on the Reference	Reference bases	Variants	Variant Type	Genomic Region
SRR7167619	6899	TAAAAAAACTT	TAAAAAAAACTT	frame-shift	GP-gene
	11930	G	T	stop_gained	L-gene
	11952	GAAAAAATTTTG	GAAAAAAATTTTG, AAAAAAATTTTG	frame-shift	L-gene
	16385	AAC	ACAC	frame-shift	L-gene
SRR7167620	7389	T	A,C (T7389A,C)	stop_lost	GP-gene
	7390	G	A,C (G7389A,C)	stop_lost	GP-gene
	7533	T	G (T7533G)	stop_lost	GP-gene
SRR7167621	11930	G	T (G11930T)	stop_gained	L-gene
SRR7621024	706	T	A	stop_gained	NP-gene
	1566	ATC	AC	frame-shift	VP35
	6899	TAAAAAAACTT	TAAAAAAAACTT	frame-shift	GP-gene
	7786	CATC	CC	frame-shift	GP-gene
	10764	CAACT	CACT	frame-shift	VP24
	13046	TGGGAT	TGGAT	frame-shift	L-gene
	14530	GCGTAG	ACGTCAG	frame-shift and synonymous	L-gene
	17734	T	A	stop_gained	L-gene
SRR7167630	7389	T	A,C,G	stop_lost	GP-gene
	7390	G	C	stop_lost	GP-gene

SRR7167631	6533	T	G	stop_gained	GP-gene
SRR7620981	3246	G	T	stop_gained	VP35
	8731	TTTTGTGTGA	TCTTTGTGTGA	frame-shift	VP30
	10587	TCTACT	ACTAGCT	frame-shift and Missense	VP24
	11952	GAAAAAATTTTG	GAAAAAAATTTTG	frame-shift	L-gene
SRR7620993	11952	GAAAAAATTTTG	GAAAAAAATTTTG, AAAAAAATTTTG	frameshift	L-gene
	17705	TAT	TT	frameshift	L-gene
SRR7620996	784	GAAAAAGGAAGGT	GAAAAGGAAGGT	frameshift	NP-gene
	1964	GAAAAAAATGAT	GAAAAAAAATGAT	frameshift	VP35
SRR7621082	1856	AATA	AA	frameshift	NP-gene
	1947	CGGCT	CCT	frameshift	NP-gene
SRR7621087	11952	GAAAAAATTTTG	GAAAAAAATTTTG, AAAAAAATTTTG	frameshift	L-gene
	17705	TAT	TT	frameshift	L-gene
SRR7621140	1322	C	T	stop_gained	NP-gene
	5293	CAAAAAATG	CAAAAAAATG	frameshift	VP40
	11027	CAAAAAACCCG	CAAAAAAAACCCG	frameshift	VP24
	11952	GAAAAAATTTTG	GAAAAAAAATTTTG	frameshift	L-gene
SRR7621143	1584	TAAAAGAC	TAAAGAC, AAAAAGAC	frameshift	NP-gene
	1658	A	T	stop_gained	NP-gene
	1964	GAAAAAAATGAT	GAAAAAAAATGAT	frameshift	NP-gene
	6278	G	A	stop_gained	GP-gene
	7107	C	T	stop_gained	GP-gene
	10513	TAAAAACT	TAAAAAACT	frameshift	VP24
	11027	CAAAAAACCCG	CAAAAAAACCCG	frameshift	VP24
	12895	TAAGAGG	TAGAGG	frameshift	L-gene
	14820	G	A	stop_gained	L-gene
	15478	TGGGGGGCA	TGGGGGCA	frameshift	L-gene
	16900	TAAAAAGT	TAAAAAAGT	frameshift	L-gene
	17145	CAGAGCATAGCATCGAGGCAGAAA	CTCGA	frameshift and missense	L-gene
	17882	C	T	stop_gained	L-gene
SRR7621144	12344	GAAGAT	GAGAT	frameshift	L-gene
SRR7621047	1070	C	T	stop_gained	NP-gene
	6548	TTCA	TA	frameshift	GP-gene
	6899	TAAAAAAACTT	TAAAAAAAACTT	frameshift	GP-gene
	6930	A	G	stop_lost and splice_region	GP-gene
	7718	CAAGC	CAGC	frameshift	GP-gene
	11027	CAAAAAACCCG	CAAAAAAACCCG	frameshift	VP24
	13324	TAAAGC	TAAGC	frameshift	L-gene
	14533	TAGA	TA	frameshift	L-gene
	15296	GTA	GA	frameshift	L-gene
	17268	TTAC	TC	frameshift	L-gene
SRR7621005	1856	AATA	AA	frameshift	NP-gene
	1947	CGGCT	CCT	frameshift	NP-gene

	9106	C	A	stop_gained	VP30
SRR7621007	1964	GAAAAAAATGAT	GAAAAAAAATGAT	frameshift	NP-gene
	4731	C	T	stop_gained	VP40
	11951	GG	GAA	frameshift and missense	L-gene
	13409	AGTG	AG	frameshift	L-gene
	14577	T	A	stop_gained	L-gene
	15008	G	T	stop_gained	L-gene
	16035	ATTTTATG	ATTTATG	frameshift	L-gene
SRR7621074	11027	CAAAAAACCCG	CAAAAAAACCCG	frameshift	VP24
	13672	TGGTC	TGTC	frameshift	L-gene
	15485	C	T	stop_gained	L-gene
	17705	TAT	TT	frameshift	L-gene
SRR7621080	11952	GAAAAAATTTTG	GAAAAAAATTTTG,AAAAAATTTTG	frameshift	L-gene
	17705	TAT	TT	frameshift	L-gene

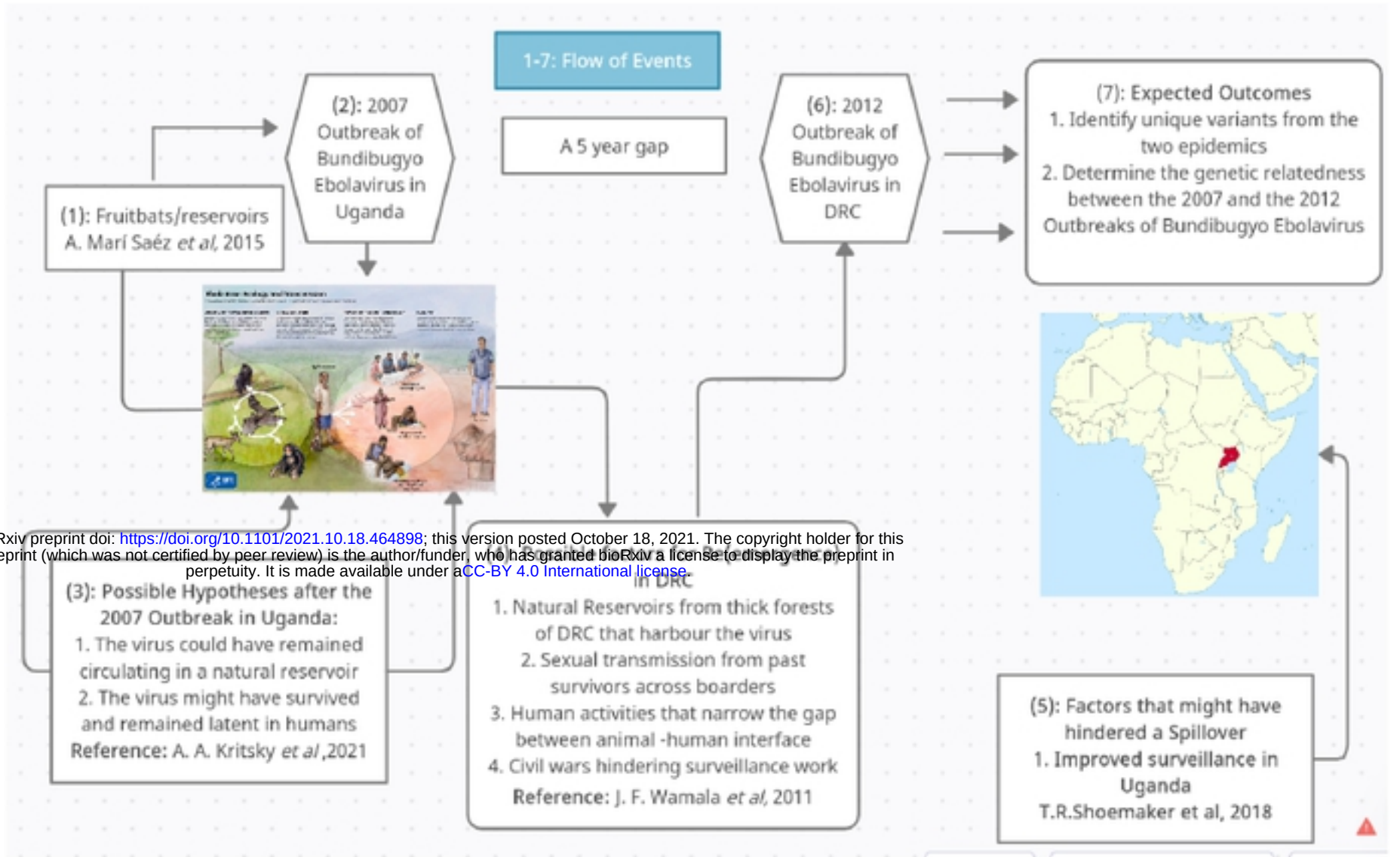
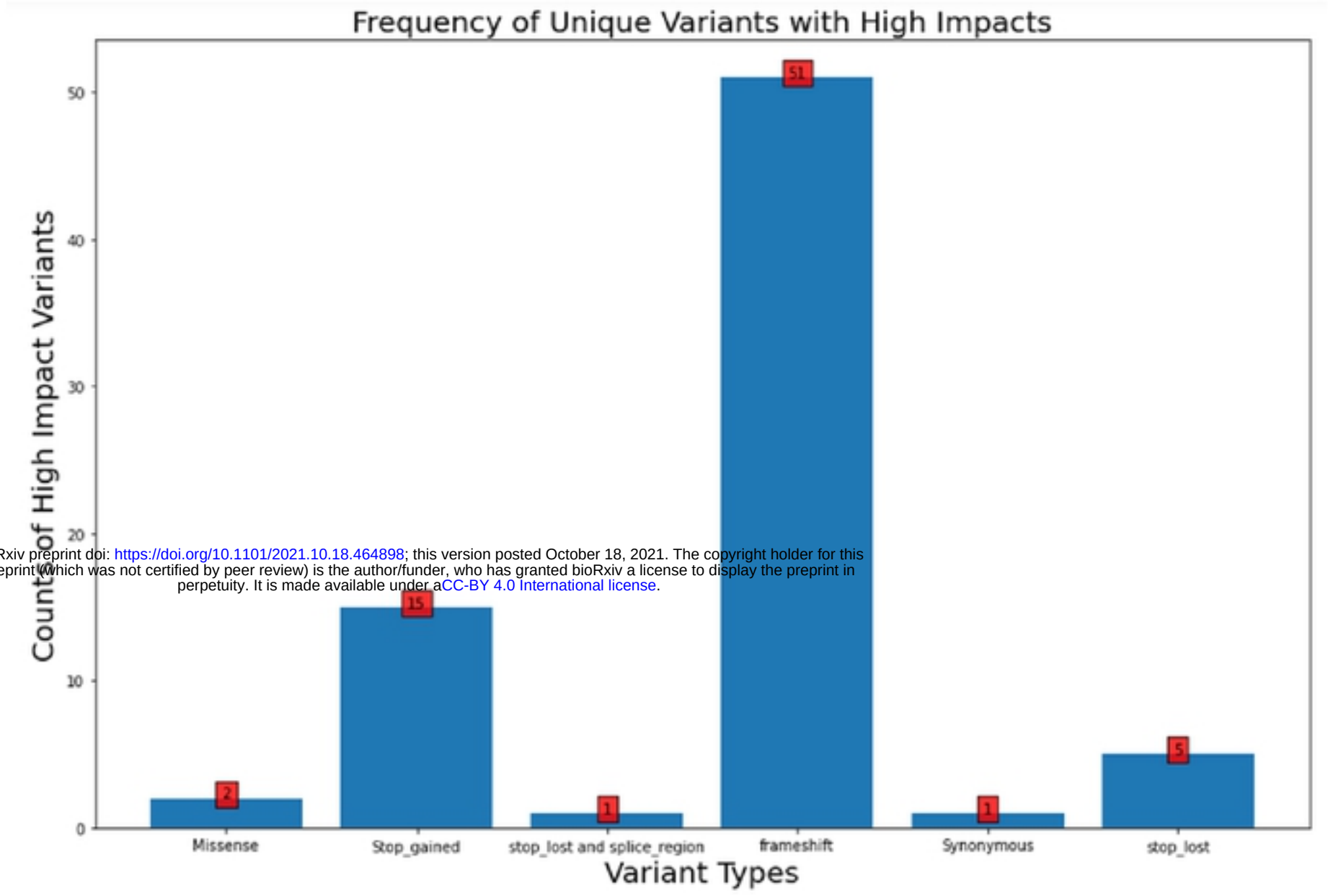


Fig 1: The working hypotheses for the resurgence in DRC in the year 2012

Fig 2: A Bar-plot showing the frequency of unique variants with high impact on the genomes of Bundibugyo Ebolavirus isolated from the 2007 outbreak in Uganda



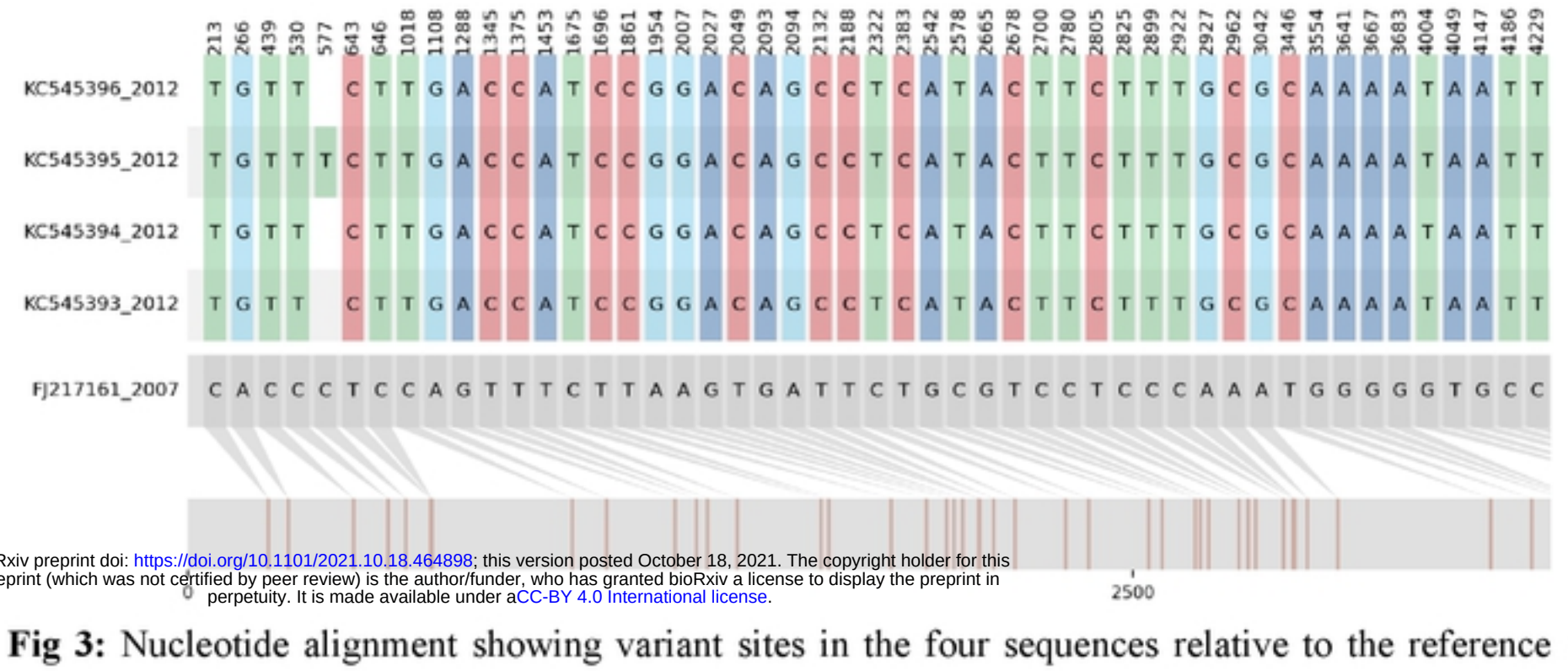


Fig 3: Nucleotide alignment showing variant sites in the four sequences relative to the reference genome. The 4 sequences represent isolates collected from the DRC outbreak in 2012 (NCBI accession numbers: KC545393, KC545394, KC545395, KC545396). The reference sequence is an isolate from the ancestral outbreak (Gene-bank accession number: FJ217161)

Fig 4: Maximum likelihood phylogenetic tree showing the two unique clusters identified from the two epidemics that occurred in Uganda (2007) and the Democratic Republic of Congo (2012). The reference sequence used was an isolate from the ancestral outbreak having a Gene-bank accession number of FJ217161