BIOF 520: Evolutionary Tools for Infectious Disease Analysis

Diana Lin[1,2]

[1]**Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada**
[2]**Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, Canada**

WORKFLOW

To determine a potential cause of the infection based on the sequence, nucleotide BLAST (1, 2) was run on the sequence, against the non-redundant nucleotide (nr/nt) database. Of all the BLAST hits, the top BLAST hit was AF372422.1: Zika virus envelope protein (E), partial cds, with an E-value of 0.0 and a 98.163% identity to the query sequence. Most of the BLAST results were either envelope genes, or polyprotein genes. A literature review on the structure of the Zika virus genome (3) revealed that the polyprotein gene encodes for the envelope protein.

Next, all published sequences from the virus (Zika virus), and the genomic region (envelope) were retrieved from Genbank (4) by querying "Zika virus envelope". These sequences were then filtered for sequence length, where only sequences with less than or equal to 1000 nucleotides in length remained. Additionally, the sequences were filtered to only include those whose primary organism was "Zika virus", thereby excluding synthetic constructs, leaving 107 sequences left. The metadata (country of origin) for these sequences was downloaded separately as a Genbank file and parsed using this bash script.

Then, a multiple sequence alignment was generated using MUSCLE (5, 6). This multiple sequence alignment was then fed into Simple Phylogeny (6), where a neighbour-joining tree was constructed with default parameters. This phylogenetic tree was outputted in Newick format, by default.

Finally, to view the tree, the Newick file was loaded into R and visualized and colour-coded by country using ggtree (7) in this R script, shown in **Figure 1-2**.

RESULTS

The phylogenetic tree shows that there are sequences from 21 different countries, where most of the sequences are from Brazil, Senegal, Russia, and the Ivory Coast, as shown in **Figure 3.**

In most cases, the sequences from the same country tend to cluster together as part of the same clade or group, and the same occurs for sequences from the same continent.

The unknown sequence from the patient aligned best with AF372422.1, a sequence also of unknown origin. This sequence is indicated in the phylogenetic tree in **Figure 1-2** with the grey text within the green box. It is in the same small clade as a sequence from Uganda, and in the same large clade with sequences from Nigeria, Gabon, Senegal, Burkina Faso, and the Ivory Coast.

Through this phylogenetic analysis, it is not unreasonable to conclude that the source of this pathogen is from Africa, perhaps even more specifically from Uganda or its neighbouring countries. However, in this particular case, it is known that the patient had recently travelled to Brazil to attend a large sporting tournament, where he may have contracted the virus and passed it onto his family members. With this knowledge, it is even more reasonable to conclude that the patient's strain of Zika virus was transmitted from a fellow traveller from Africa (possibly from Uganda), who was attending the same sporting tournament. This is a valid assumption, as large sporting events, such as the FIFA World Cup or the Olympics, often attract athletes and spectators from all over the world.

This type of analysis aids in tracking transmission across countries, and finding the source of such pathogens during outbreaks.

The concept of 'Original Antigenic Sin', the fact that immune systems favour immunological memory when exposed to a second similar antigen, may be important to predicting the outcome of a patient infected with the unknown sequence. If a patient had exposure to any of the other strains within the same clade, the patient may fare worse than if the patient had no exposure at all. Because of this concept, it should become more standard in healthcare to, at least, partially sequence, if not entirely sequence each strain of virus that the patient has contracted, and keep a detailed record in an electronic health record. Without that knowledge in the patient history, the rationale behind the patient's worsening condition could be completely missed.

The idea of 'Original Antigenic Sin' is also an important concept when it comes to creating vaccines for these large outbreaks that often mutate very fast, with various strains. The optimal vaccine would have to be general enough to be effective against multiple strains, but not too specific that the immune system would be trapped by the first immune response. Ideally, the immune system would produce general antibodies that would be effective against the antigens of multiple strains. A phylogenetic analysis of these multiple strains would definitely facilitate vaccine development.

FIGURES

Figure 1. The phylogenetic tree of Zika virus envelope protein genes from various countries. The green box highlights the clade that contains the sequence most similar to the patient's sequence, AF372422.1. Due to the large size of the tree and the low image resolution below, PDF and PNG files have also been provided.

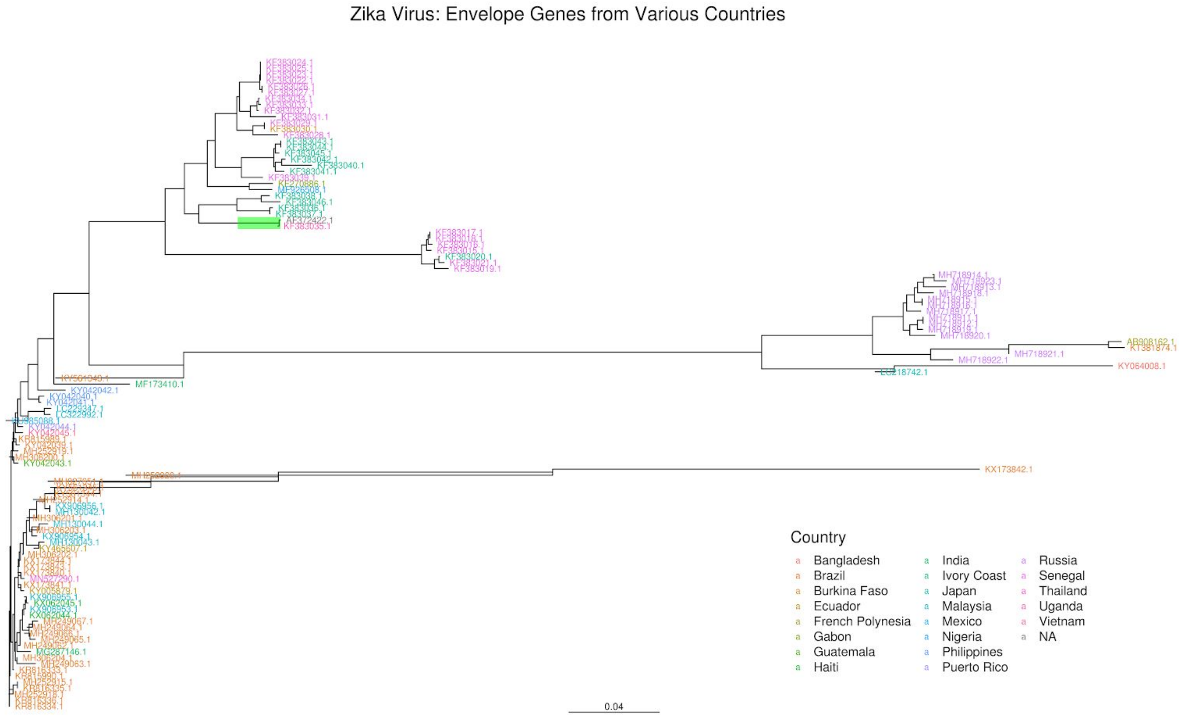


Figure 2. The corresponding cladogram to the phylogenetic tree in Figure 1. The green box highlights the clade that contains the sequence most similar to the patient's sequence, AF372422.1. Due to the large size of the tree and the low image resolution below, PDF and PNG files have also been provided.

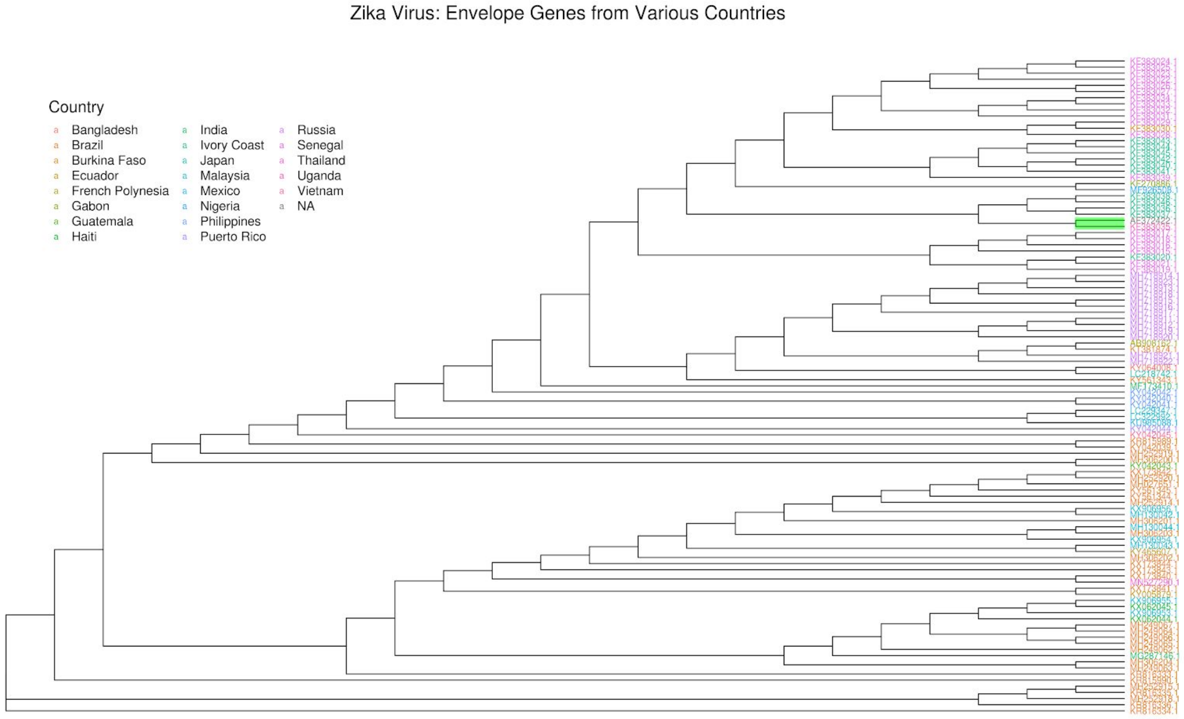
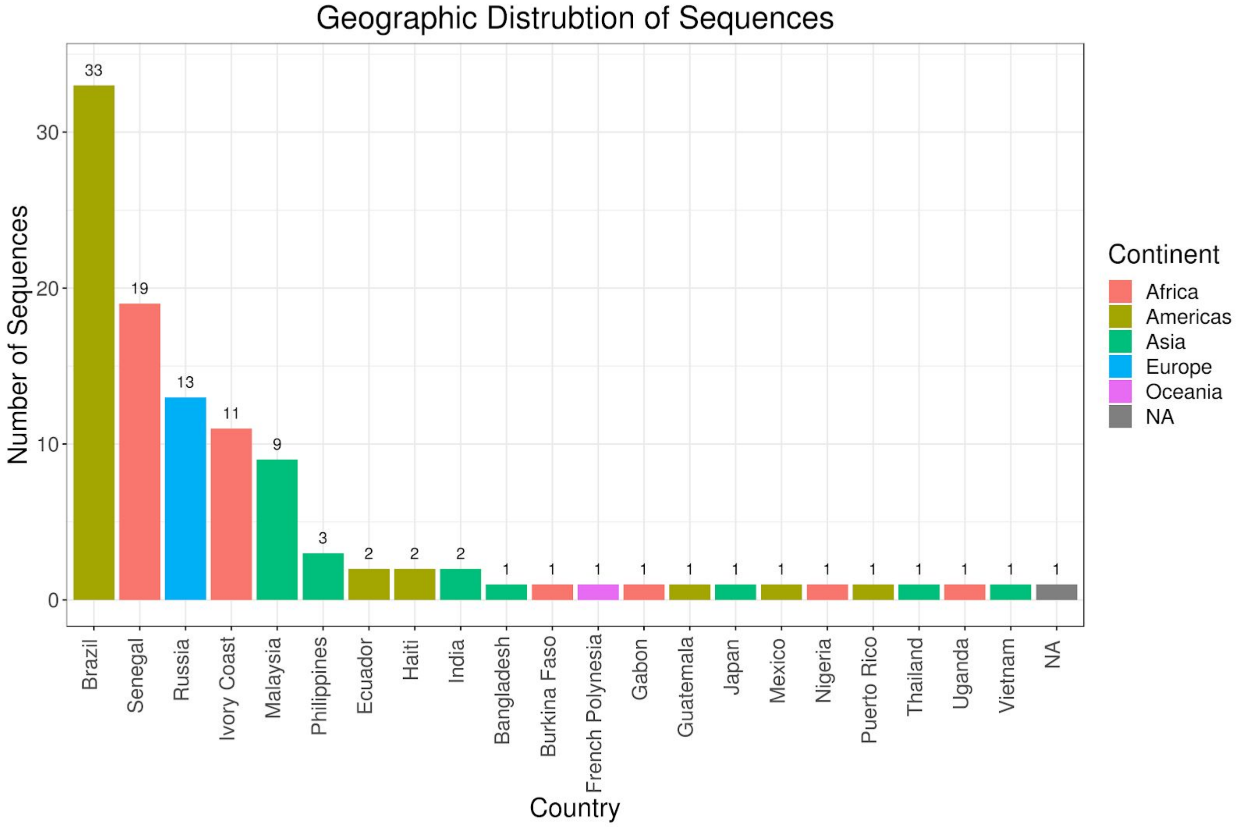


Figure 3. A histogram of the number of sequences from each country, coloured by continent.



REFERENCES

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.
2. Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214.
3. Sirohi D, Kuhn RJ. 2017. Zika Virus Structure, Maturation, and Receptors. J Infect Dis 216:S935–S944.
4. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016. GenBank. Nucleic Acids Res 44:D67–72.
5. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.
6. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47:W636–W641.
7. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution.