Analysis of The Nucleotide Sequence Diversity of the Lassa Virus and Augmenting its Phylogenetic Tree

10.17975/sfj-2018-005

Sean Oddoye¹

1. THE S.T.E.M. Academy at Morristown High School

Abstract

Lassa Virus (LASV) is the etiological catalyst for Lassa fever, an acute hemorrhagic disease with a mortality rate of 15%. Many aspects of the Lassa virus are not understood, like the causation of deafness in $\frac{1}{3}$ of surviving patients or why symptoms are benign for 80% of those infected with the virus. Ambiguities like these suggest that there might exist some genomic heterogeneity among infecting viruses and demonstrate a need to quantify and analyze polymorphisms within LASV. Patterns that emerge from phylogenetic trees can be used to assess the structure of a population while also providing insights to the genetic makeup. The purpose of this investigation was to develop a more streamlined means of calculating nucleotide diversity within a subpopulation of Lassa virus strains and to augment a phylogenetic tree of the Lassa Virus glycoprotein precursor (GPC) segment. A total of 25 partial and complete data sequences of LASV strains were obtained from the Genbank Archives. During phase one of this investigation, the sequence data was inputted into MEGA analytical software and the sequence diversity was derived on a nucleotide level. Data from the individual strand sequences was used to augment a phylogenetic tree using Treeview X software. In phase two of this investigation, an algorithm was created using RStudio, with BSGenome and BioStrings extensions. The sequence diversity derived from the statistical analyses on MEGA was compared to that of the algorithm created. A p-value of 0.08 was found, which deviates from the accepted range of non-medical p-value of 0.00 to 0.05. It is suggested that future research focuses on creating a refurbished version of the algorithm to calculate a nucleotide diversity within a percent error of 5%.

Key Words

Lassa Virus, Nucleotide Diversity, Phylogenetic Tree, Sequence Alignment, Glycoprotein-Precursor

Introduction

Lassa fever is a significant health risk factor that continues to threaten the lives of over 300,000 people.² Vaccines are seldom created before an intimate knowledge of the disease's genetic or hereditary information are considered.⁴ Society has been aware of the Lassa virus for well over fifty years, and in that time, several remedies for the disease have been made, but a long-lasting vaccine has yet to be created. The augmentation of a molecular phylogenetic tree is critical in understanding the geographic evolution of the Lassa virus. Molecular phylogeny can offer a unique possibility to understand the reproductive process of the virus, and its geographic evolution patterns.¹ An analysis of the phylogeny of the Lassa virus could be employed to create a candidate vaccine. This is significant as linking the evolutionary lineages of the virus will aid in rationalizing incomplete data related to the heritability of the Lassa virus in an efficient manner.

Materials and Methods

Phylogenetic Construction

1. 25 nucleotide sequences of the LASV GPC were acquired and aligned using MEGA software.
2. Using MEGA, the sequence was reconstructed into the first draft of the maximum-likelihood, neighbor-joining phylogenetic tree.
 - a. Transitions and trans-versions were included.
 - b. Gaps, censored, or otherwise incomplete data were omitted from the phylogenetic tree.
 - c. Node IDs were displayed.
3. Using MEGA, a second draft of a maximum-likelihood, neighbor-joining phylogenetic tree was created.
 - a. Taxa were grouped and renamed.
 - b. Branch lengths were calculated and displayed.
 - c. Phylogenetic tree was changed to have a rectangular layout.
7. 25 nucleotide sequences were aligned a second time using ClustalW program.

8. Alignment exported to TreeViewX software.
9. Third draft of phylogenetic tree was constructed.

Algorithm Creation

1. Draft one of the algorithm was created using RStudio. An attempt to calculate sequence variability with Shannon's Entropy function (Figure 1) was made.
- b. No environments or extensions within R were used.
- c. First draft was deemed unsuccessful, as the principal investigator found that the Shannon's Entropy function is used to find the sequence variability of a set of sequences rather than the nucleotide diversity value. This draft was saved and discontinued.

$$H = - \sum p(x) \log p(x)$$

Figure 1: Shannon's entropy is a formula used for calculating the degree of variability in a sequence where H is the sequence variability and p(x) is the expected probability of nucleotide "x" appearing in a sequence (Nei et Li).

2. Second draft of the algorithm was created using RStudio. An attempt to calculate nucleotide diversity with the nucleotide diversity equation (Figure 2) was made.
- c. Biostrings and Multiple Sequence Alignment extensions were used in creating the second draft.
- d. Second draft was deemed unsuccessful, was saved and discontinued.

$$\pi = \sum_{ij} x_i x_j \pi_{ij} = 2 * \sum_{i=2}^n \sum_{j=1}^{i-1} x_i x_j \pi_{ij}$$

Figure 2: Two equations for calculation nucleotide diversity between two different sequences. The principal investigator used the first equation in creating the algorithm.³

3. Third draft of the algorithm was created using RStudio. A second attempt to calculate nucleotide diversity with the nucleotide diversity equation was made.
- d. BSgenome, Biostrings, Gdata, and MSA extensions were used in creating the third draft
- e. Third draft was generated with a somewhat feasible nucleotide diversity output when compared to the theoretical value, and had a p-value of 0.08.

Results

It was hypothesized that if a phylogenetic tree for the Lassa virus was augmented, then the geographic divergence, represented by the tree's branch length of the nucleotide sequences observed will be greater than that of another phylogenetic tree that was scientifically accepted in GenBank; Ultimately, the principal investigator failed to reject this hypothesis. The geographic divergence in the augmented phylogenetic tree was greater than that of a scientifically accepted tree in GenBank. It was also hypothesized that if an algorithm was created to calculate sequence diversity in a subpopulation of the LASV glycoprotein precursor segment, then the nucleotide diversity value would deviate from the accepted value by a value greater than 5%. Through an ANOVA statistical test, the theoretical nucleotide diversity value generated by the algorithm was found to be statistically significant. As a result, the principal investigator failed to reject the hypothesis.

Figure 3: An analysis of variance (ANOVA) table showing the variance, f-stat, and p-value when mean frequencies of nucleotides across 25 sequences were compared.

Analysis of Variance Table				
Source of Variation	Sum of Squares	Degree of Freedom	Variance	F-Statistic
Between Groups	497.5569	3	165.8523	15.9322
Within Groups	999.3456	96	10.4099	
Total	1496.9025	99 *		†

Discussion

This determination regarding the hypotheses for the phylogenetic tree happened to be the favorable outcome, implying that the methods used were reminiscent of the methods the researchers used in their phylogenetic reconstruction. Furthermore, the lack of evolutionary divergence between this phylogenetic tree and others allows researchers to compare the taxa between phylogenetic trees to highlight any genetic commonalities. Similar phylogenetic constructions and comparative analyses could allow for more detailed evaluations of the Lassa Virus to be made.

Failing to reject the null hypothesis for the algorithm suggests an unfavorable result, as it demonstrates that there were likely errors made when aligning the sequences, creating the nucleotide diversity algorithm, or running the statistical tests following the investigation. The error is believed to be the cause of the deviation in the p-value, due to an erroneous or inconsistent sequence alignment. The p-value was found to be 0.08 while the maximum accepted non-medical p-value is 0.05. There was a rather miniscule deviation in these values, so it stands to reason that a very miniscule error facilitated this change. It is unlikely that the error was the result of an unsatisfactory algorithm, as any small error would have either caused an error message to appear in the R-console or would have resulted in an inaccurate nucleotide

diversity value that deviates from the accepted value by a larger factor. Nor is it likely that an error was made in calculating the accepted nucleotide diversity or running the one-way ANOVA test to determine statistical significance, as these values were entirely computer-generated. It is more likely that a few minor errors were made in the creation of the sequence alignments since its creation required processing parts of the alignment and any mistake made in handling the sequence information is very unlikely to be detected.

Conclusion

The phylogenetic tree construction was successful and could possibly lend itself into a deeper analysis of the evolutionary patterns of the Lassa virus. However, the nucleotide diversity algorithm diverges from the accepted non-medical p-value of 0.05 and would not be an accepted means with which to calculate nucleotide diversity. If this investigation was continued in the future, conjoining the two separate phases of the study (i.e. the phylogenetic tree and the nucleotide diversity algorithm) into a single algorithm would be favourable to produce better results.

Acknowledgement

Thank you to the Young Science Achievers Program (USAP MHSSO 1617) and the Morris Educational Foundation (MEF SO 20162017) for providing the principal investigator with the funds to conduct this research, Dr. David Safronetz and infectious disease specialist at the University of Manitoba for providing the principal investigator with critical insights into the Lassa Fever and the Lassa Virus, Mr. Derek Stein also from the University of Manitoba for giving helpful suggestions regarding the phylogenetic tree creation, Ms. Vidhi Chitalia for her time and dedication to helping the principal investigator with the algorithm creations and Mrs. Erin Colfax the research teacher of the principal investigator for her expertise and assistance with nearly all aspects of the research study.

References

1. Broyles, Robyn. "What Is Phylogeny". Bright Hub Education. N.p., 2012. Web. 20 Oct. 2016.
2. Centers for Disease Control and Prevention. "Lassa Fever". CDC. N.p., 2015. Web. 22 Oct. 2016.
3. Nei, M, and W H Li. "Mathematical Model For Studying Genetic Variation In Terms Of Restriction Endonucleases.". Proceedings of the National Academy of Sciences of the United States of America 96.10 (1979): 1-5. Web. 18 Oct. 2016.
4. Offit, Paul. "How Are Vaccines Made?". CHOP. N.p., 2017. Web. 28 June 2016.

STEM Fellowship Journal Downloaded from journal.stemfellowship.org by 18.207.199.55 on 04/26/23

Figure 1: The image depicts the first 35 sites of all 25 sequences in the LASV GPC alignment.

Figure 3: The second draft of the phylogenetic tree shown with a 'rectangular' layout.

Appendix B

```
seq = sample(DNA_ALPHABET[1:4], size=10, replace=TRUE)
dnastring_one <- DNAString("ACTGACTGACTG")
dnastring_two <- DNAString("TGATAGCCAAAC")
dnastring_three <- DNAString("ACATCAGCTGGTT")
dnastring_four <- DNAString("CCTGTGTGAGCT")
dnastring_five <- DNAString("CGAGGCTGGGGA")
dnastring_six <- DNAString("TGACAGCTCTTT")
dnastring_seven <- DNAString("AAACATAGGCGG")
dnastring_eight <- DNAString("GGGCGTCGAAC")
dnastring_nine <- DNAString("GGTCGAGAAACT")
dnastring_ten <- DNAString("CATAGTACCTCG")

diff_per_site <- 3/4 * (1 - exp(-4 * delta))
delta <- 0.1 * years_evo
#the years_evo value can change to suit, so long as it is between 0 and 2.5
years_evo <- 0.1

#frequency function is shown below, don't delete and save for later.
alphabetFrequency(dnastring_one, baseOnly=TRUE, as.prob = TRUE)
alphabetFrequency(dnastring_two, baseOnly=TRUE, as.prob = TRUE)
alphabetFrequency(dnastring_three, baseOnly=TRUE, as.prob = TRUE)
alphabetFrequency(dnastring_four, baseOnly=TRUE, as.prob = TRUE)
alphabetFrequency(dnastring_five, baseOnly=TRUE, as.prob = TRUE)
alphabetFrequency(dnastring_six, baseOnly=TRUE, as.prob = TRUE)
alphabetFrequency(dnastring_seven, baseOnly=TRUE, as.prob = TRUE)
alphabetFrequency(dnastring_eight, baseOnly=TRUE, as.prob = TRUE)
alphabetFrequency(dnastring_nine, baseOnly=TRUE, as.prob = TRUE)
alphabetFrequency(dnastring_ten, baseOnly=TRUE, as.prob = TRUE)

P1_one_two <- sum()
P1_one_three
P1_one_four
P1_one_five
P1_two_three
P1_two_four
P1_two_five
P1_three_four
P1_three_five
P1_four_five
```

Figure 4: The first draft of the nucleotide diversity algorithm. The algorithm attempted to transpose the Shannon's Entropy function into readable lines of code in RStudio.

Data Table C: Phylogenetic Tree Descriptive Statistics		
Feature	Branch Length (RFD)	Pairwise Distance (DPS)
Maximum	3.49417	10.246
Minimum	0.00001	0.001*
Mean	1.60726	6.229
Mode	0.00001*	7.592
Median	2.13508	6.0615
Range	3.49416	10.245
Standard Deviation	1.38484	1.460

Figure 5: Descriptive statistics for the phylogenetic tree's branch length and pairwise distances comparing the distances between taxa.

Figure 6: A graphical representation of the average nucleotide frequencies for each sequence.