**CS645 Project 1**

**Understanding Ebola Virus Outbreak**

Team members: Jaswanth Datt Meka, Srinivasa Rao Katta, Lihong Yin

**Abstract**

Ebola virus is a very dangerous [virus](https://simple.wikipedia.org/wiki/Virus) which causes a hemorrhagic fever called Ebola virus disease causing the victim to bleed a lot, inside and outside their body. It has been very difficult to find cure for this disease .The recent outbreak of Ebola in 2014 is far more deadly than the previous outbreaks. It has been reported that this strain of virus has mutated itself from the earlier known genomic sequences. In this project, we compared 10 EBOV genome sequences from 2014 patients with 1976 Zaire or Sudan EBOV genome. Smith-Waterman algorithm was employed for pairwise alignment. Our results showed the Ebola virus from 2014 outbreak is mostly derived from 1976 Zaire EBOV. More mutations occurred on coding region than noncoding region. The genome sequence alignment of EBOV may provide useful epidemiological information for future disease prevention.

**Introduction**

The Ebola virus (EBOV) causes an acute, serious illness which is often fatal if untreated. Ebola virus disease (EVD) first appeared in 1976 in 2 simultaneous outbreaks, one in Sudan, and the other in Democratic Republic of Congo. Since its discovery, five species of EBOV have been isolated: *Zaire* (ZEBOV), *Sudan, Côted’Ivoire, Reston* (REBOV) and the proposed ‘*Bundibugyo*’ EBOV. *Zaire*, *Sudan and Bundibugyo* EBOVs have been associated with large outbreaks in Africa with high human case fatalities (25–90%). The current outbreak in West Africa, (first cases notified in March 2014), is the largest and most complex Ebola outbreak since the Ebola virus was first discovered in 1976. There have been more cases and deaths in this outbreak than all others combined. It has also spread between countries starting in Guinea then spreading across land borders to Sierra Leone and Liberia, by air to Nigeria and USA, and by land to Senegal and Mali.

From the 2014 outbreak, 99 EBOV genomes were sequenced from 78 confirmed EVD patients from Sierra Leone. Investigation the virus genomic sequence will help us to better understand the epidemiological information from tracing contacts. In this study, we randomly selected 10 sequences from 2014, aligned with 1976 Zaire and Sudan EBOV genome. The method we used is Smith-Waterman pairwise alignment. The algorithm performs the local sequence alignment, that is, for determining similar regions between two strings or [nucleotide](http://en.wikipedia.org/wiki/Nucleotide_sequences) or [protein sequences](http://en.wikipedia.org/wiki/Protein_sequence). Instead of looking at the [total](http://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm) sequence, the Smith–Waterman algorithm compares segments of all possible lengths and [optimizes](http://en.wikipedia.org/wiki/Mathematical_optimization) the similarity measure.

By alignment of 2014 EBOV sequences with Zaire and Sudan virus genome, we aim to find out the ancestor of 2014 Ebola virus, whether they are stem from Zaire virus or Sudan virus. Our results implied the virus from 2014 outbreak is mostly related to 1976 Zaire virus.

**Materials and methods**

We have done the comparison of genome sequences to study the mutations in the Ebola virus that have occurred with earlier outbreaks. The data for the analysis is taken from the source provided by The National Center for Biotechnology Information (ncbi). We compared the alignment of EBOV[4] 2014 with that of Zaire[5] and Sudan[6].To study the effect, we have randomly considered 10 patient records( KR105205.1, KM233118.1, KM233069.1, KM233042.1, KM233066.1, KM233090.1, KR105238.1, KR105250.1, KR105258.1, KR105330.1). We have compared each of the Coding Sequence (CDS) with the Genome from that in Zaire and Sudan.Each patient record has 9 coding sequences ­ *NP, VP35, VP40, GP, VP30, VP24* and *L.*There were a total of 180 comparisons of sequences­90 each for Zaire and Sudan respectively.

Statistical analysis was performed by using two-sample paired nonparametric Wilcoxon rank test.

**Implementation**

To observe the mutations in the genome sequences we have used Smith­Waterman sequence alignment, which is a local alignment. Since the comparing coding sequences (CDS) are less in length, we used this local alignment to compare with the gene sequences of 1976 in Sudan and Zaire. We have used the pairwise alignment in the R tool to obtain the matches. The score for gap opening is taken as negative 2. The gaps, mismatch for the coding sequences based on the coding snippet below.

for (i in 1:length(seq1AlignedV\_sudan))

{

if (seq1AlignedV\_sudan[i]!=seq2AlignedV\_sudan[i])

{

if(seq1AlignedV\_sudan[i]=="­" || seq2AlignedV\_sudan[i]=="­" )

{

gapcount\_sudan=gapcount\_sudan+1

}

else

{

mismatch\_sudan = mismatch\_sudan+1

}

}

}

Using the statistical tool R, we have obtained the mismatches and gap count for each CDS sequences as shown in Table S1 (Appendices).

**Results and Discussion**

The Ebola virus genome from Zaire 1976 (Genebank accession No: AF086833.2) is made of cRNA. The genome contains seven genes: *NP, VP35, VP40, GP, VP30, VP24* and *L*. However, there are 9 coding sequences in the genome, which is due to cotranscriptional editing of the *GP* gene. So there are nine proteins encoded from this virus genome, which are nucleoprotein (NP), polymerase complex protein (VP35), matrix protein (VP40), structural glycoprotein (GP), the secreted glycoprotein (sGP) and the small soluble glycoprotein (ssGP), polymerase complex protein (VP30), membrane-associated protein (VP24) and polymerase (L). During transcription, a poly-A tail is added to the 3’ end of the messenger RNA (mRNA). This process is called polyadenylation. A poly-A tail is between 100 and 250 residues long. The poly-A tail makes the RNA molecule more stable and prevents its degradation. Additionally, the poly-A tail allows the mature messenger RNA molecule to be exported from the nucleus and translated into a protein by ribosomes in the cytoplasm.

To detect the difference of 2014 outbreak sequence aligned with 1976 Zaire and Sudan sequence, first we determined the sample size using Minitab software. When we set the difference of mismatches and gaps to 1000, standard deviation of error to 500, power to 0.90, it turns out we need at least 7 samples (Appendices).

So we randomly selected 10 Ebola sequences from 2014 which are KM233090, KM233042, KM233066, KM233069, KM233118, KR105205, KR105258, KR105330, KR105250, KR105238. We found all of the 10 virus genome from 2014 contain the same genes as in Zaire and Sudan genome. Using Smith-Waterman pairwise alignment, the number of mismatches and gaps were obtained for each sequence aligned with Zaire or Sudan sequence (Table 1).

Table 1. The number of total mismatches and gaps aligned with Zaire or Sudan genome

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Zaire |  | Sudan |  |
|  | mismatch | gap | mismatch | gap |
| KM233090.1 | 562 | 0 | 3070 | 6911 |
| KM233066.1 | 562 | 0 | 3071 | 6890 |
| KM233042.1 | 563 | 0 | 3063 | 6932 |
| KM233069.1 | 563 | 0 | 3067 | 6932 |
| KM233118.1 | 553 | 0 | 3068 | 6795 |
| KR105205.1 | 562 | 0 | 3071 | 6925 |
| KR105258.1 | 564 | 0 | 3068 | 6924 |
| KR105330.1 | 568 | 0 | 3063 | 6932 |
| KR105250.1 | 568 | 0 | 3063 | 6932 |
| KR105238.1 | 568 | 0 | 3063 | 6932 |
| mean | 563.3 | 0 | 3066.7 | 6910.5 |
| SD | 4.45 | 0.00 | 3.43 | 42.77 |

To check if there is statistical difference for the nucleotide changes (mismatch + gap) between the two sets of alignment, we first choose to use 2-sample paired T-test. In order for this test to be applicable, it needs to meet the assumption that the population difference is normally distributed. We use SAS software to check the normality. The null hypothesis is: the data is normal. The result showed that all the p-value are less than 0.01, which rejects the null hypothesis. (Appendices, Table S2, Figure S1). So we conclude that the data is not normally distributed. When the data is not normal, the t-test is not the most powerful test. Non-parametric testing is applicable if the assumption about normality of the population is not met. We then used two-sample Wilcoxon rank test by using R (Appendices).

The null hypothesis for nonparametric test is: The population distributions of x and y are the same. We got p-value=0.002897, which is less than 0.05, reject null hypothesis. We have enough evidence to conclude that there is significant difference for the number of mismatches and gaps from 2014 sequence when aligned with Zaire and Sudan genome. Since the number of mutations is less in aligning with Zaire sequence, it suggests the 2014 outbreak Ebola virus is most likely derived from 1976 Zaire virus.

From the result of alignment with Zaire virus genome, we found more mutations occurred in coding sequences than in intergenic noncoding region (p-value < 0.05) (Figure 1). The average mismatches on CDS is 382.6, while noncoding region has 180.7 mismatches.

Figure 1. Comparison of the number of mutations on coding and noncoding region aligned with Zaire genome.

When analyzing the mutations from the 9 CDS of the virus, we found the number of mutations varies among different coding region. The result showed mutations happened more often on gene L, NP and GP (Table 2, Figure 2). Also, within the same coding region, the number of mutations from different patient is not always consistent (Table 2).

Table 2. Selected 2014 virus sequence aligned with Zaire genome.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | NP | VP35 | VP40 | GP | sGP | ssGP | VP30 | VP24 | L |
| KM233090.1 | 63 | 12 | 15 | 59 | 20 | 15 | 12 | 17 | 169 |
| KM233066.1 | 63 | 12 | 15 | 59 | 20 | 15 | 12 | 17 | 169 |
| KM233042.1 | 63 | 12 | 14 | 59 | 20 | 15 | 12 | 17 | 170 |
| KM233069.1 | 63 | 12 | 14 | 59 | 20 | 15 | 12 | 17 | 170 |
| KM233118.1 | 63 | 12 | 14 | 59 | 20 | 15 | 12 | 17 | 169 |
| KR105205.1 | 63 | 12 | 14 | 59 | 20 | 15 | 12 | 17 | 170 |
| KR105258.1 | 64 | 12 | 14 | 59 | 20 | 15 | 12 | 17 | 170 |
| KR105330.1 | 63 | 13 | 14 | 59 | 20 | 15 | 13 | 17 | 171 |
| KR105250.1 | 63 | 12 | 14 | 59 | 20 | 15 | 12 | 17 | 171 |
| KR105238.1 | 63 | 12 | 14 | 59 | 20 | 15 | 13 | 17 | 171 |
| mean | 63.1 | 12.1 | 14.2 | 59 | 20 | 15 | 12.2 | 17 | 170 |
| SD | 0.32 | 0.32 | 0.42 | 0.00 | 0.00 | 0.00 | 0.42 | 0.00 | 0.82 |

Figure 2. Comparison of the number of mutations on different coding region aligned with Zaire genome.

We also found that there was one gap for GP and one gap for ssGP for each of the 10 sequences selected from 2014. The gap for GP gene is due to frameshift. GP gene is critical in the EBOV life cycle, as it is solely responsible for attachment, fusion and entry of target cells. Moreover, GP is responsible for critical pathogenic differences among viral species. The *GP* gene encodes seven consecutive adenosine nucleotides within a predicted hairpin loop. While the addition of an extra adenosine leads to the production of GP, deletion of one or addition of two adenosines during the transcriptional stutter leads to the expression of a third product, ssGP (schematic figure showed below).



For the mutations from coding region, we haven’t analyzed whether they are synonymous or nonsynonymous substitutions. More analytical work needs to be done for further understanding the biological meaning of the genomic changes.

**Conclusion**

We aligned 10 Ebola virus sequences from 2014 outbreak with the EBOV genome either from Zaire or from Sudan 1976. From our study, we found the 2014 Ebola virus was most related to Zaire virus by statistical analysis. By comparing 2014 sequences with Zaire Ebola virus genome sequence, we found more mutations occurred on coding region than on noncoding region. Also, the number of mutations varies among different coding region. The gene L polymerase has more mutations than other genes. Our study showed the 2014 EBOV carries a number of genetic changes. The observed intrahost and interhost mutations will provide important insight about virus transmission and epidemiology.

**References**

1. Shimyn Slomovic et al. Polyadenylation of ribosomal RNA in human cells. [*Nucleic Acids Res*.](http://www.ncbi.nlm.nih.gov/pubmed/16738135) 2006, 34 (10):2966-75.

2. Stephen K. Gire et al. Genomic surveillance elucidates Ebola virus origin and transmission

during the 2014 outbreak. *Science*, 2014, 345 (6202): 1369-1372.

3. Jeffrey E Lee et al. *Ebolavirus* glycoprotein structure and mechanism of entry. *Future Virol*. 2009 ; 4(6): 621–635.

**Appendices**

[Minitab output]

Power and Sample Size

One-way ANOVA

α = 0.05 Assumed standard deviation = 500

Factors: 1 Number of levels: 2

Maximum Sample Target

Difference Size Power Actual Power

1000 7 0.9 0.929070

[R output]

> x=c(562,562,563,563,553,562,564,568,568,568)

> y=c(9984,9961,9995,9999,9863,9996,9992,9995,9995,9995)

> wilcox.test(x, y, paired=TRUE, alternative = "less")

Wilcoxon signed rank test with continuity correction

data: x and y

V = 0, p-value = 0.002897

alternative hypothesis: true location shift is less than 0

Table S1. mismatches and gap count for each CDS sequences and complete sequence aligned with Sudan virus genome

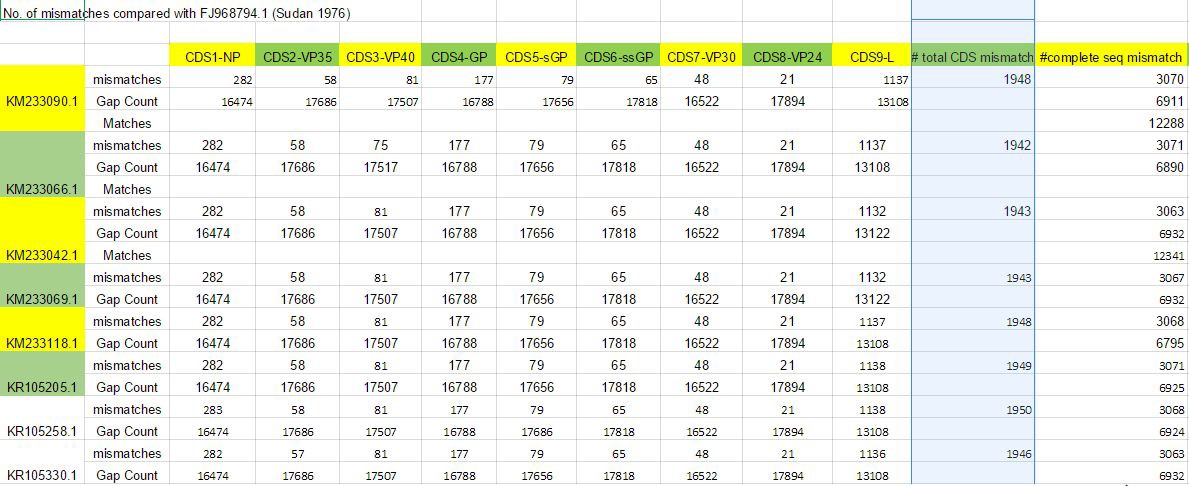
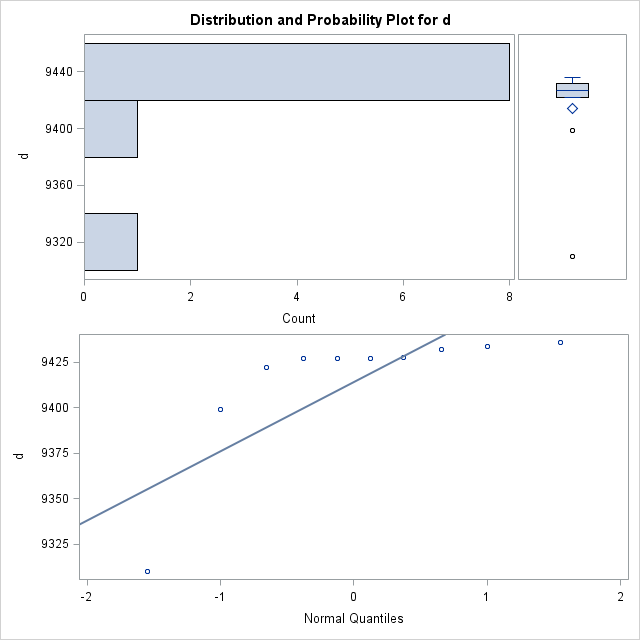


Table S2. Check for normal distribution of population

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tests for Normality** | | | | |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.543452 | **Pr < W** | <0.0001 |
| **Kolmogorov-Smirnov** | **D** | 0.361906 | **Pr > D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 0.389767 | **Pr > W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 2.052181 | **Pr > A-Sq** | <0.0050 |

Figure S1. Plot checking normality



**Contribution of team members:**

The team members contributed equally for this project.

Lihong helped to figure out reading sequence data into R, aligned part of the sequences, do data analysis, drafted introduction, results and discussion, conclusion, and do final revision.

Jaswanth help with R for 4 patients Gene banks, mismatch and gap count for all four records, materials and methods and abstract documentation.

Srinivas help with coding R for extraction mismatch and gap count for all of six patients Gene banks records, documentation.