Ebola Virus Project [folder]

Priscilla Fujikawa | Tipparat Umrod
priscillayumiko@gmail.com | tipparat.umrod@utexas.edu

Final Report | Main Page | **Presentation**
Due December 11, 2014

Title: RNA sequences in *ebolavirus* species that facilitate viral transmission to humans
Topic: Analysis of different gene mutations in *ebolavirus* species using sequence alignment

Abstract Tipparat Umrod | Priscilla Fujikawa

Background Tipparat Umrod

The 2014 Ebola epidemic is the largest one in the history [1]. Ebola virus (EBOV) and Marburg virus (MARV) are the members of the *Filoviridae* family, both of which cause severe hemorrhagic fever in primates. EBOV consists of 5 species: *Reston ebolavirus, Sudan ebolavirus, Zaire ebolavirus, Bundibugyo ebolavirus, and Tai Forest ebolavirus* [3]. A virus must be able to target specific cells in order to infect a new host. Thus, this research focuses on the molecular characteristics of some species of the Ebolavirus genus that differ in other species and facilitate viral transmission to humans.

Objective Priscilla Fujikawa

The main objective of this project was to determine the genes that facilitate viral transmission to humans in the *Reston* species of the *Ebolavirus* genus. This was accomplished by identifying conserved and dissimilar RNA domain sequences among the various taxa in the *Filoviridae* family.

Results Tipparat Umrod | Priscilla Fujikawa

We discovered 7 conserved genes in the *Filoviridae* family: nucleoprotein (NP), polymerase cofactor (VP35), matrix VP40, virion envelope glycoprotein (GP), transcription factor (VP30), VP24, and polymerase (L). We annotated them in every sequence, and calculated gene length, pairwise identity, and identical sites. Not only does Reston have all the same genes present in other *ebolavirus* species, but the length of those genes also does not differ by a great amount within the different taxa.

The gene GP has shown to have the lowest percentage of identical sites in all three cases: within the *Filoviridae* family, within the *Ebolavirus* species, and in the comparison of only *Reston* and *Zaire.* In addition, the gene VP35 showed the lowest pairwise percent identity in the *Filoviridae* family.

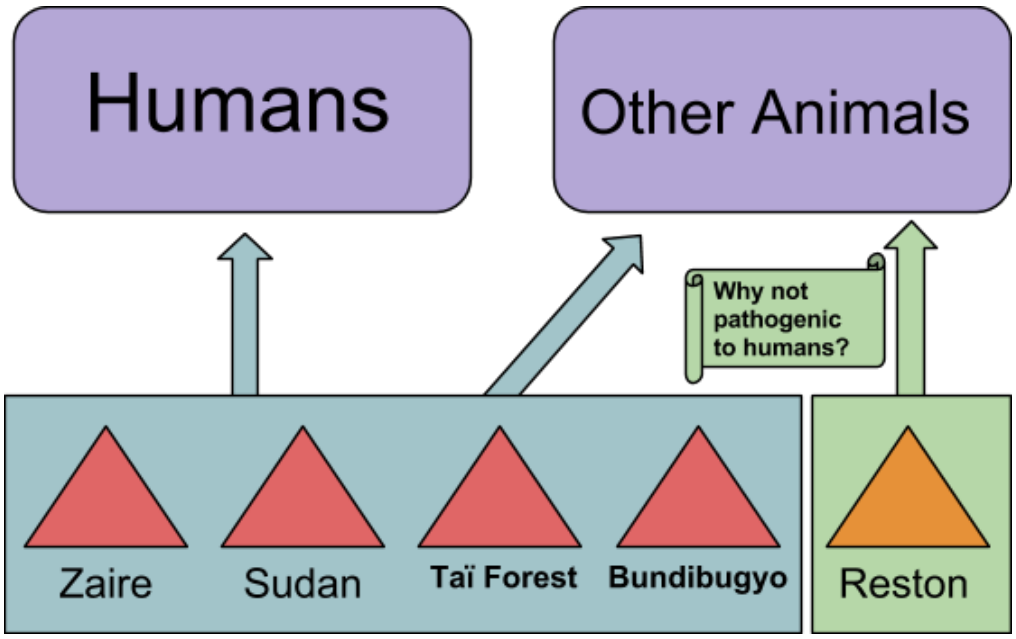
Conclusion Tipparat Umrod | Priscilla Fujikawa

Previously, after observing gene length data, we limited our scope to 3 potential genes that might help facilitate viral transmission to humans: VP35, VP40, and GP. That is due to their differences in average gene lengths in Reston and Zaire.

However, after calculating the pairwise % identity for VP35, we observed a significantly low percentage in the *Filoviridae* family, but no difference in *Ebolavirus*. On the other hand, the percentage of identical sites for GP genes is significantly low across the table, which includes the comparison within *Ebolavirus* and between *Zaire* and *Reston*.

GP is responsible for accepting the virus into the host genome, and it is present in the viral envelope. Therefore, after more detailed analysis, we were able to determine that GP is the strongest candidate for the gene that facilitates viral transmission to humans.

Graphical Abstract Priscilla Fujikawa



Background Tipparat Umrod

On March 24, 2014, the World Health Organization reported the outbreak of Ebola Virus Disease (EVD) in West Africa. Pasteur Institute in Lyon, France, suggested Zaire ebolavirus as the causing agent. The 2014 Ebola epidemic is the largest one in the history [1]. By October 27, 2014, there were 13,703 EVD cases, with 4,920 deaths [2]. The number of cases has continued to rise.

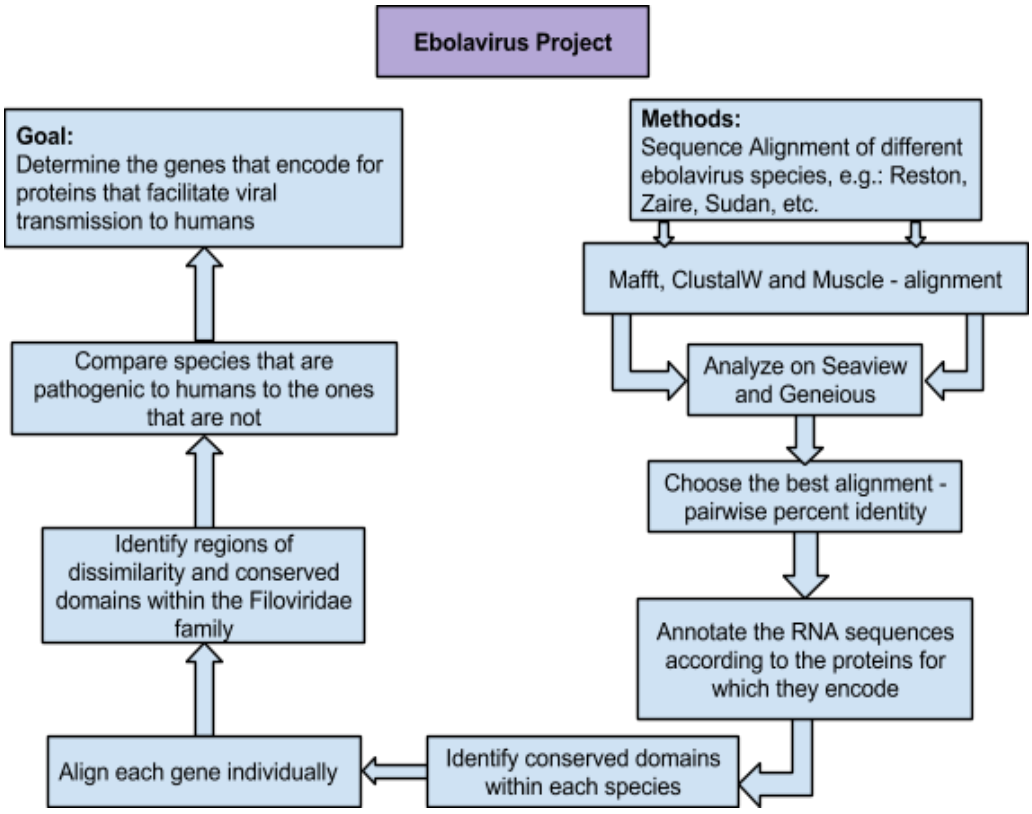
Ebola virus (EBOV), Marburg virus (MARV) and Cueva virus are members of the *Filoviridae* family in the *Mononegavirales* order. Both of the Ebola virus and Marburg virus cause severe hemorrhagic fever in primates. While those two virus are similar, the Cueva virus from the same family is distantly related. EBOV consists of 5 species: *Reston ebolavirus, Sudan ebolavirus, Zaire ebolavirus, Bundibugyo ebolavirus, and Tai Forest ebolavirus* [3]. *Reston ebolavirus* has caused the disease in nonhuman primates and not yet caused diseases in humans [4]. The viruses in the Filoviridae family have genomes of 18-19 kb long, and encode for 7 proteins: virion envelope glycoprotein (GP), nucleoprotein (NP), polymerase cofactor VP35, matrix VP40, transcription factor VP30, VP24, and polymerase (L) [5].

A virus must be able to infect the appropriate cells in order to infect a new host, which can be restricted at different levels. For example, receptor binding, entry or fusion, trafficking within the cell, genome replication, and gene expression. [6] This research focuses on the gene expression that facilitates viral infections to appropriate cells.

It is significant that we understand how viruses switch to a new host, especially the viral transmission to human, in order to prevent it at earlier stage. Identifying the genes that facilitate viral transmission to humans in the Zaire ebolavirus would be an approach to help us understand the mechanism of host switching in the Ebola virus.

Both *Ebola* virus and *Marburg* virus are on biosafety level 4, which is considered to be dangerous. Therefore, computational methods, like sequence alignments and sequence analysis are the preferred approach by researchers. This research also extends to the whole *Filoviridae* family, including the *Marburg* virus. Our purpose is to identify the conserved domains within the family with the objective of identifying the conserved domains that facilitates viral transmission to humans.

Methods Tipparat Umrod | Priscilla Fujikawa



Our methods consisted of 7 parts:

1. Data collection
2. Alignment programs and parameters
3. RNA sequence annotation
4. Quantitative analysis of the annotation
5. Alignment of each gene individually
6. Translation of RNA sequences into amino acids
7. Quantitative analysis of amino acids alignment

1. Data collection

Our 225 sequences were obtained from the NCBI GenBank website. See Appendix-1 for a full list of the accession numbers of all the sequences used in this research.

2. Multiple sequence alignment programs and parameters

We attempted to perform multiple sequence alignments on the online versions of Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/), Muscle (http://www.ebi.ac.uk/Tools/msa/muscle/), and Mafft (http://mafft.cbrc.jp/alignment/server/).

Mafft has various parameters available, such as progressive methods, iterative refinement methods, scoring matrices and gap penalties.

3. **RNA sequence annotation**

We annotated all 225 RNA sequences in the *Filoviridae* family, and as a result, 7 main genes were conserved across species: nucleoprotein (NP), polymerase cofactor (VP35), matrix protein (VP40), virion envelope glycoprotein (GP), transcription factor (VP30), VP24 and polymerase (L). This procedure was performed using the tools on Geneious 8.04 (http://www.geneious.com, Kearse et al., 2012), and the database used was NCBI.

See Appendix-2 and Appendix-3 for complete gene annotations.

4. **Quantitative analysis of the annotation**

We calculated the mean gene length of all 7 genes on each species; the pairwise percent identity with mean and standard deviations; and identical sites percentage, also with mean and standard deviations.

5. **Alignment of each gene individually**

We aligned each of the 7 genes separately in all species, in order to determine both regions of similarity and dissimilarity within each gene according to the species in which it is found.

6. **Translation of RNA sequences into amino acids**

This procedure was performed using the tools available on Geneious, in order to qualitatively assess the multiple sequence alignment.

7. **Quantitative analysis of amino acids alignment**

We performed a quantitative analysis of the amino acid sequence alignments by calculating the mean amino acid sequence length produced by each gene in every species, as well as the pairwise percent identity.

Results

Priscilla Fujikawa | Tipparat Umrod

Alignment Results

Priscilla Fujikawa

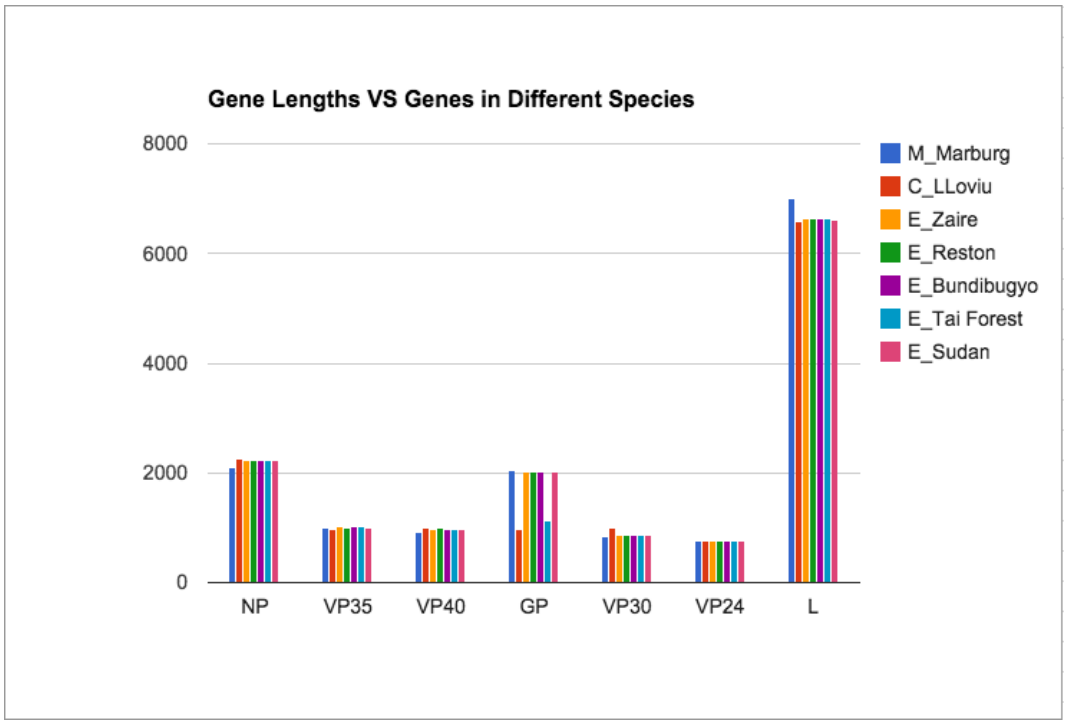
Due to our large data (225 sequences of approximately 18,500 base pairs), Clustal and Muscle did not accept the FASTA file containing all the sequences. Therefore, we used Mafft for sequence alignment.

After performing the alignment with various combinations of different parameter options, we decided to continue this research with the alignment that showed the highest

pairwise identity: 66.5%. Thus, the final alignment was performed using: G-INS-i; Gap: 1.00; 200 PAM.

Annotation Results Tipparat Umrod

The data consist of 3 genus from the Filoviridae family: Marburg(M), Cueva(C), and Ebola(E). While marburgvirus and cuevavirus have only one species, the ebolavirus has 5 species. As a result from the graphical gene annotation in Geneious, all the species have the starting of NP proteins relatively at the same position except Marburg that has a shorter and later starting position. VP30 gene of Marburg and of Cueva (Lloviu) have significantly different starting positions than that of all ebolavirus species. The L gene of all species are approximately the same length and at the same starting/ending positions. [Appendix-3]

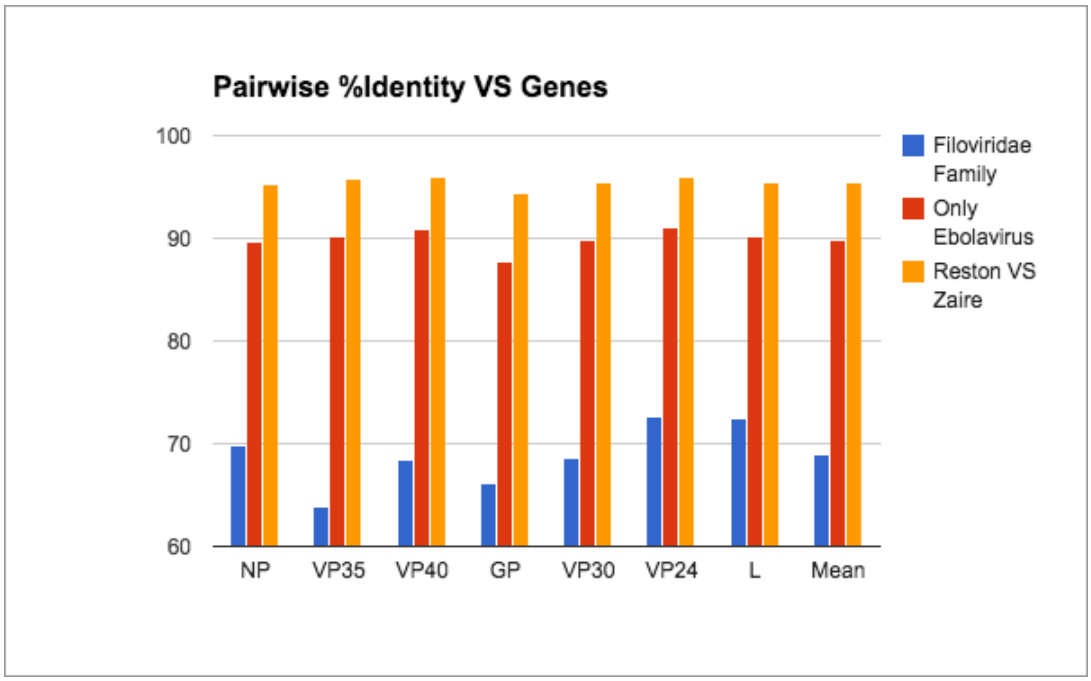


[Fig. Annotation 1: shows the gene lengths of each species in a bar graph to illustrate the differences]

Each gene in different species has relatively the same amount of nucleotides. The average NP gene length in *Marburg* (2088) and *Lloviu* (2250) are different from the ones from *ebolavirus* (2220). The average VP35 gene length in *Marburg* (990) is the same as *Sudan* (990) and *Reston* (990); however, *Zaire* (1023) and *Bundibugyo* (1026) are longer while *Lloviu* (963) is shorter. VP40 average gene length in Marburg(912) is the shortest and differed from other species by approximately 70 nucleotides. In addition, the VP40 gene length in *Reston* (996) differed from the rest of the *ebolavirus* (981) by 15 nucleotides. VP30 gene length in *Lloviu* is the longest and differed from the other species by approximately 120 nucleotides. VP24 gene length is relatively close across all of the species. L gene length in *Marburg* is the

longest and differed from the rest of the species by about 350 nucleotides. The GP gene varies in length but stay relatively the same across the species except the *Tai Forest* and *Lloviu*. [See Appendix-4]

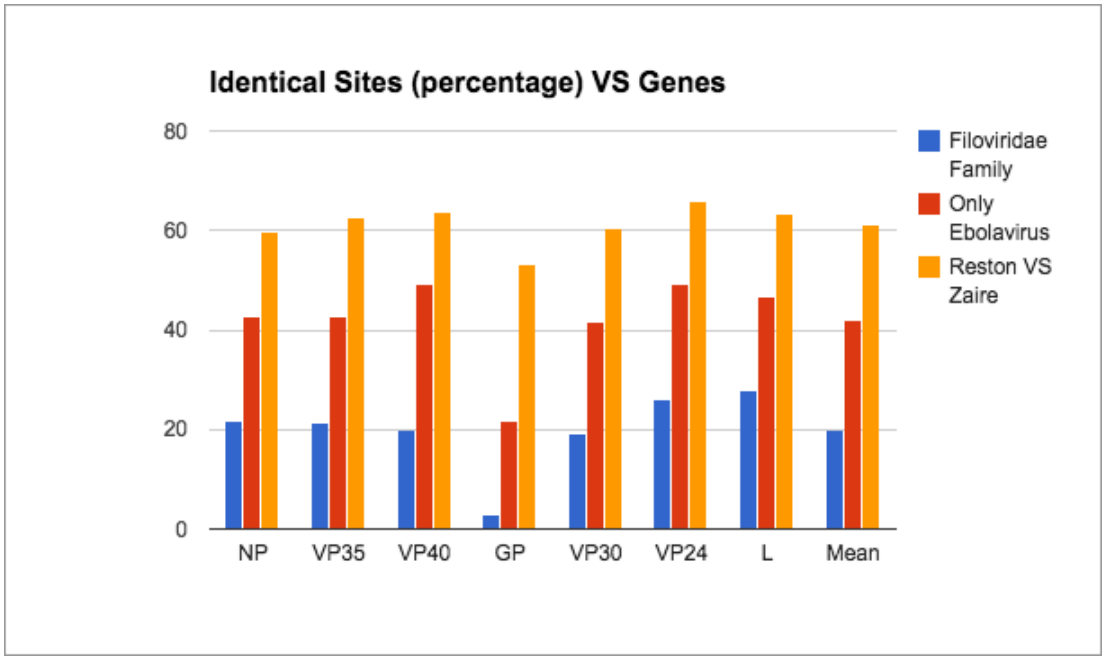
From the prior information above, Fig Annotation 1. that shows the differences in gene lengths of each species, the NP gene in Marburg is the shortest among all of the species. While the VP24 has relatively the same gene lengths, the average gene length in GP and L show significant differences. The average gene length in VP35, VP40 and VP30 are relatively the same but with various gene lengths.



[Fig. Annotation 2: shows the pairwise %identity of each gene in a bar graph to illustrate the differences]

According to Geneious user manual, the pairwise % identity was computed by "looking at all pairs of bases at the same column and scoring a hit (one) when they are identical, divided by the total number of pairs. Ambiguity characters are interpreted, meaning a nucleotide A versus a nucleotide R is considered to have 50% identity."

In Fig. Annotation 2, the mean for pairwise % identity of all the genes are as following: Filoviridae Family (68.9±3.19), Only Ebolavirus(89.9±1.09), and Reston VS Zaire(95.5±0.53). VP35 pairwise %identity when comparing with the whole family is significantly below average and its standard deviation. However, when comparing with only ebolavirus and Reston VS Zaire, the significance seems to vanish. [See Appendix-5 for numerical data]



[Fig. Annotation 3: shows the identical sites (in percentage) of each gene in a bar graph to illustrate the differences]

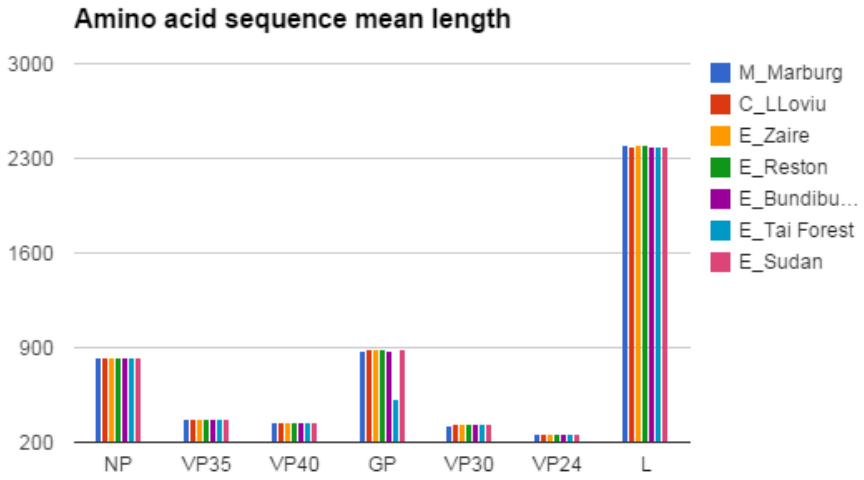
Identical sites was calculated by "considering only those columns in the alignment that have at least 2 nucleotides/amino acids/gaps that are not free end gaps and are not columns consisting entirely of gaps. A column not meeting this requirement is not even counted as non-identical for the percentage calculation. A column meeting this requirement is considered identical if it contains no internal gaps and all the nucleotides/amino acids are identical. Ambiguity characters are not interpreted, so a nucleotide column of A and R is not considered identical."

In Fig. Annotation 3, the mean for identical sites (in percentage) of all the genes are as following: *Filoviridae* Family (19.9±8.13), Only Ebolavirus(42.0±9.50), and Reston VS Zaire(61.4±4.10). GP identical sites when comparing the entire *Filoviridae* family, only ebolavirus, and Reston VS Zaire, are significantly below average and its standard deviation. [See Appendix-6 for numerical data]

Amino Acid Analysis Results **Priscilla Fujikawa**

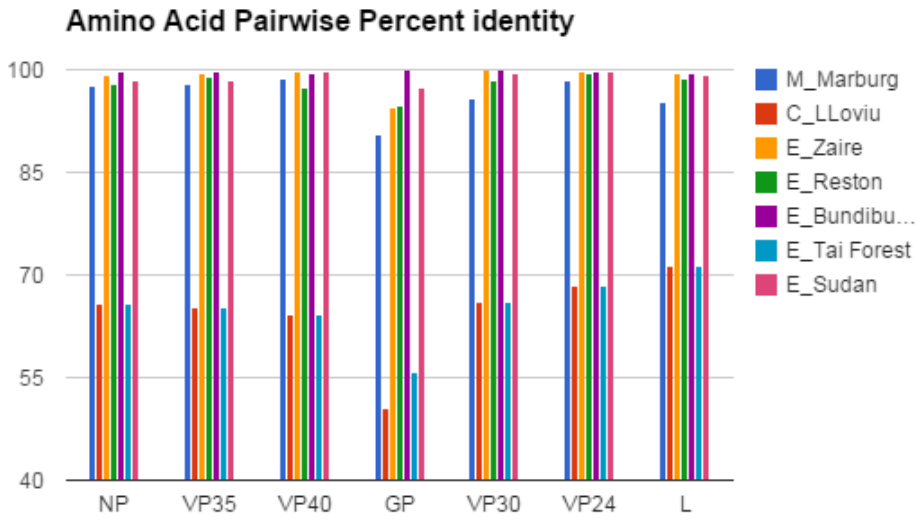
For a quantitative analysis of the alignment of the amino acid sequences, we calculated the amino acid sequences mean length; and the amino acid sequence pairwise percent identity. A table with all the data collected can be found on Appendices 7 and 8, as well as a sample alignment of the amino acid sequences of GP on Geneious [Appendix 9].

The results were summarized in the graphs below:



[Fig. Amino acid analysis 1: amino acid sequence mean length for each gene, categorized by species]

This graph shows that the mean amino acid sequence length is consistent across species, with the exception of *Taï Forest*. Moreover, L corresponds to the longest protein, while VP24 is the shortest.



[Fig. Amino acid analysis 2: amino acid pairwise percent identity for each gene, categorized by species]

This graph shows that in all cases, *Taï Forest* and *Bundibugyo* have the two lowest percent identity of approximately 70%, while all other species have a pairwise percent identity above 90%.

Conclusions

Tipparat Umrod | Priscilla Fujikawa

Originally, we expected the analysis of the genomes to yield the conclusion that there was some gene in *Reston* that was not present in any other species of the *Filoviridae* family. That is because *Reston* is the only species in the *Ebolavirus* genus that has not been shown to infect humans. However, we came to discover that:

1. All species had approximately the same number of base pairs (18,500) in their genome [see Appendix 10]
2. All species had the same seven genes (GP, NP, VP24, VP30, VP35, VP40, L)
3. Gene length was consistent across different species for all genes

Thus, *Reston* was not easily distinguishable from others in its family.

After calculating the percentage of identical sites for GP genes, we discovered that it is significantly low across the table. Moreover, GP is present in the viral envelope that attaches to the membrane of human cells. Therefore, it is a very strong candidate for the gene that facilitates viral transmission to humans.

Future studies should include a larger data set if possible, as well as secondary structure and covariation analysis studies of all seven genes.

Acknowledgments

Tipparat Umrod | Priscilla Fujikawa

The authors thank Dr. Robin Gutell and Ms. Anna Mengjie Yu for providing support; suggesting the most appropriate softwares; sharing articles and experience; and addressing our issues and concerns patiently. More importantly, we thank you for your time and effort in caringly help us succeed.

Appendices

Tipparat Umrod | Priscilla Fujikawa

Appendix - 1 List of accession numbers used in this research, categorized by species

Zaire ebolavirus	AF086833, AF272001, AF499101, AY354458, EU224440, HQ613402, HQ613403, JQ352763, KC242784, KC242785, KC242786, KC242787, KC242788, KC242789, KC242790, KC242791, KC242792, KC242793, KC242794, KC242795, KC242796, KC242797, KC242798, KC242799, KC242800, KC242801, KF827427, KJ660346, KJ660347, KJ660348, KM034549, KM034550, KM034551, KM034552, KM034553, KM034554, KM034555, KM034556, KM034557, KM034558, KM034559, KM034560, KM034561, KM034562, KM034563, KM233035, KM233036, KM233037,KM233038, KM233039, KM233040, KM233041, KM233042, KM233043, KM233044, KM233045, KM233046, KM233047, KM233048, KM233049, KM233050, KM233051, KM233052, KM233053, KM233054, KM233055, KM233056, KM233057, KM233058, KM233059, KM233060, KM233061, KM233062, KM233063, KM233064, KM233065, KM233066, KM233067, KM233068, KM233069, KM233070, KM233071, KM233072, KM233073, KM233074, KM233075, KM233076, KM233077, KM233078, KM233079, KM233080, KM233081, KM233082, KM233083, KM233084, KM233085, KM233086, KM233087, KM233088, KM233089, KM233090, KM233091, KM233092, KM233093, KM233094, KM233095, KM233096, KM233097, KM233098, KM233099, KM233100, KM233101,KM233102, KM233103, KM233104, KM233105, KM233106, KM233107, KM233108, KM233109, KM233110, KM233111, KM233112, KM233113, KM233114, KM233115, KM233116, KM233117,

	KM233118, KM655246 [130]
Sudan ebolavirus	KC545389, KC545390, KC545391, KC545392, KC589025, FJ968794, KC242783, EU338380, AY729654, JN638998, [10]
Reston ebolavirus	AB050936, AF522874, AY769362, FJ621583, FJ621584, FJ621585, JX477165, JX477166, [8]
Taï Forest ebolavirus	FJ217162
Bundibugyo ebolavirus	FJ217161, KC545393, KC545394, KC545395, KC545396, [5]
Lloviu virus	JF828358
Marburg marburgvirus	AY358025, AY430365, AY430366, DQ217792, DQ447649, DQ447650, DQ447651, DQ447652, DQ447653, DQ447654, DQ447655, DQ447656, DQ447657, DQ447658, DQ447659, DQ447660, EF446131, EF446132, EU500827, EU500828 FJ750953, FJ750954, FJ750955, FJ750956, FJ750957, FJ750958, FJ750959 GQ433351, GQ433352, GQ433353, JN408064, JX458825, JX458826, JX458827, JX458828, JX458829, JX458830, JX458831, JX458832, JX458833, JX458834, JX458835, JX458836, JX458837, JX458838, JX458839, JX458840, JX458841, JX458842, JX458843, JX458844, JX458845, JX458846, JX458847, JX458848, JX458849, JX458850, JX458851, JX458852, JX458853, JX458854, JX458855, JX458856, JX458857, JX458858, KC545387, KC545388, KM261523, Z12132, Z29337, [70]

Appendix - 2 Gene Annotations of 225 sequences

Appendix - 3 Color Coded Gene Annotation of candidate sequence from each species

Appendix - 4 Average Gene Length of Each Gene in Different Species

Species	Average Gene Length						
	NP	VP35	VP40	GP	VP30	VP24	L
M_Marburg	2088	990	912	2046	846	762	6996
C_LLoviu	2250	963	990	963	987	753	6591
E_Zaire	2220	1023	981	2030	870	756	6639
E_Reston	2220	990	996	2031	868	756	6639
E_Bundibugyo	2220	1026	981	2030	870	756	6633
E_Tai Forest	2220	1026	981	1122	870	756	6633
E_Sudan	2220	990	981	2031	870	756	6621

Appendix - 5 Pairwise %Identity of Each Gene

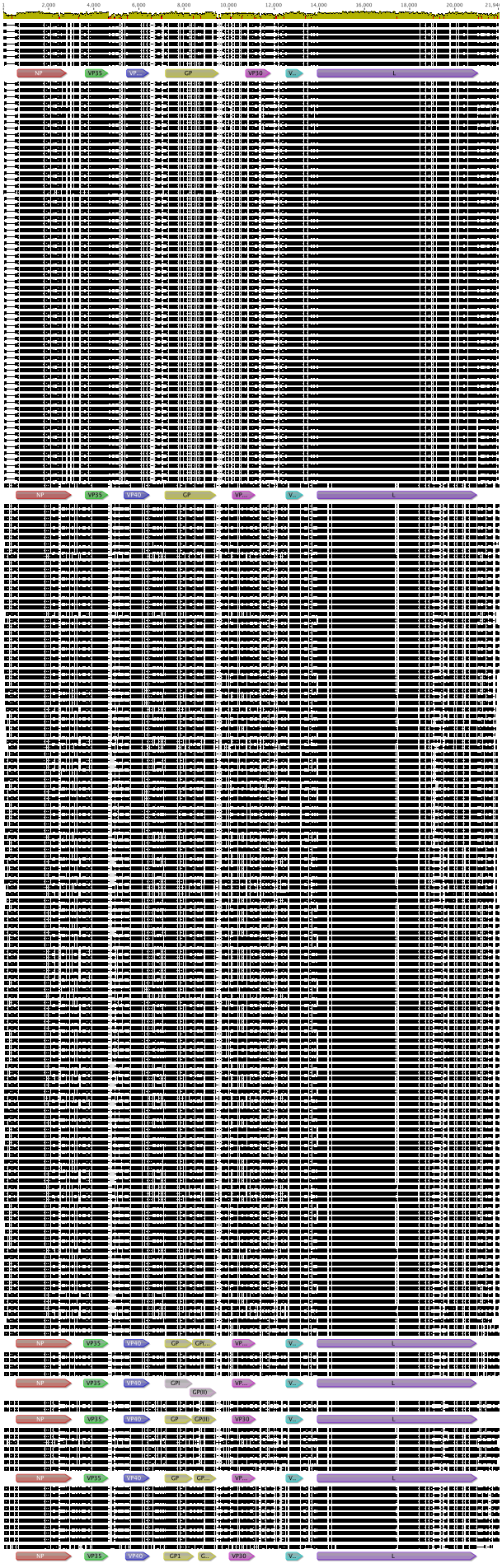
Species	Pairwise % Identity								
	NP	VP35	VP40	GP	VP30	VP24	L	Mean	Standard Deviation
Filoviridae Family	69.9	63.9	68.5	66.2	68.6	72.7	72.5	68.9	3.19
Only Ebolavirus	89.6	90.1	90.8	87.7	89.8	91.0	90.2	89.9	1.09
Reston and Zaire	95.2	95.8	95.9	94.4	95.5	95.9	95.5	95.5	0.53

Appendix - 6 Identical Sites (in percentage) of Each Gene

Species	Identical Sites (percentage)								
	NP	VP35	VP40	GP	VP30	VP24	L	Mean	Standard Deviation
Filoviridae Family	21.6	21.5	19.9	2.9	19.3	26.2	27.8	19.9	8.13
Only Ebolavirus	42.8	42.8	49.4	21.7	41.7	49.1	46.8	42.0	9.50
Reston and Zaire	59.8	62.8	63.7	53.3	60.6	66.0	63.3	61.4	4.10

Appendix-2 [zoom in to see detail]

Appendix-3



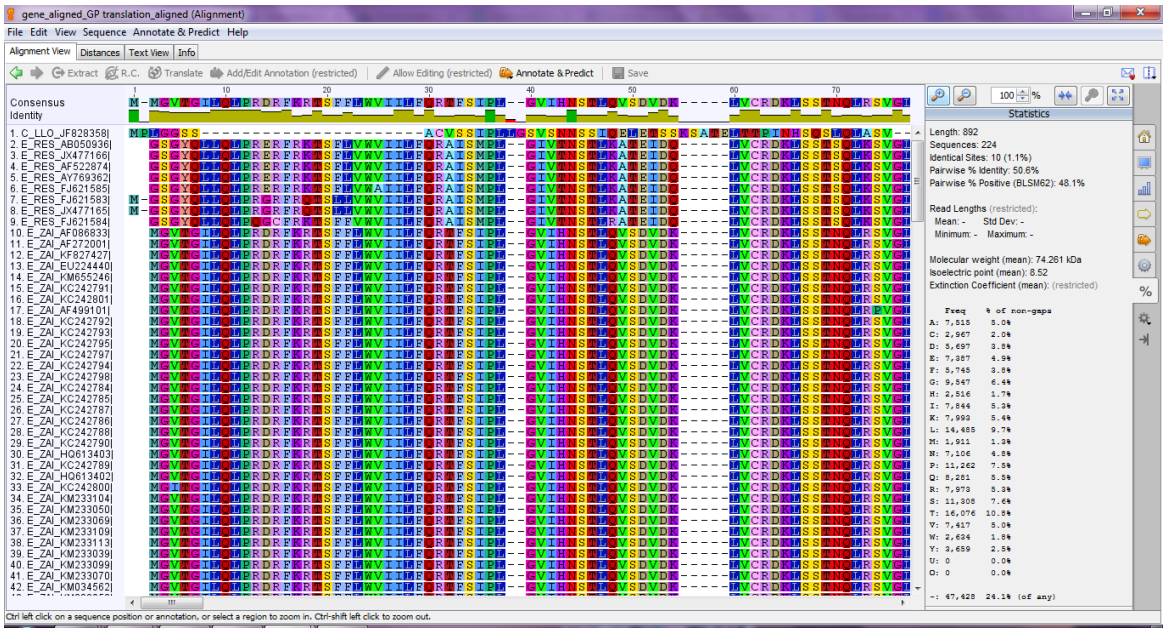
Appendix - 7 Amino acid sequence mean length

	Amino acid sequence mean length						
Species	NP	VP35	VP40	GP	VP30	VP24	L
M_Marburg	823	371	347	876	319	257	2,397
C_LLoviu	823	371	350	892	332	257	2,389
E_Zaire	825	371	352	892	335	257	2,397
E_Reston	825	371	352	884	335	257	2,395
E_Bundibugyo	825	371	347	874	335	257	2,391
E_Tai Forest	825	371	347	523	335	257	2,391
E_Sudan	825	371	347	890	335	257	2,392

Appendix - 8 Amino acid pairwise percent identity

	Pairwise percent identity						
Species	NP	VP35	VP40	GP	VP30	VP24	L
M_Marburg	97.6	97.9	98.8	90.4	95.9	98.4	95.3
C_LLoviu	65.7	65.2	64.2	50.6	66	68.5	71.3
E_Zaire	99.2	99.6	99.7	94.5	99.9	99.8	99.6
E_Reston	97.8	99	97.4	94.8	98.5	99.4	98.7
E_Bundibugyo	99.7	99.8	99.5	99.9	99.9	99.8	99.6
E_Tai Forest	65.7	65.2	64.2	55.8	66	68.5	71.3
E_Sudan	98.4	98.3	99.8	97.5	99.5	99.8	99.3

Appendix - 9 Multiple sequence alignment of GP after translation



Appendix - 10 Genome length of all species

	A	B	C	D	E
1	**Species**	**Shortest**	**Longest**	**Average**	
2	**Marburg**	19104	19112	19108	
3	**Zaire**	18613	18961	18787	
4	**Bundibugyo**	18939	18940	18939.5	
5	**Sudan**	18874	18875	18874.5	
6	**Reston**	18796	18935	18865.5	
7	**Lloviu**	18927	18927	18927	
8					
9		MIN	MAX		
10		18613	19112		
11					
12	**Conclusion: Zaire has the shortest sequence, and Marburg the longest one**				

References

Tipparat Umrod

[1] "2014 Ebola Outbreak in West Africa." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, 27 Oct. 2014. Web. 29 Oct. 2014.

[2] Ebola Response Roadmap Situation Report. Rep. no. 10. World Health Organization, 29 Oct. 2014. Web. 31 Oct. 2014. <http://apps.who.int/iris/bitstream/10665/137376/1/roadmapsitrep_29Oct2014_eng.pdf?ua=1>

[3] Kuhn JH, Becker S, Ebihara H, Geisbert TW, Johnson KM, Kawaoka Y, Lipkin WI, Negredo AI, Netesov SV, Nichol ST, Palacios G, Peters CJ, Tenorio A, Volchkov VE, Jahrling PB. Proposal for a revised taxonomy of the family Filoviridae: classification, names of taxa and viruses, and virus abbreviations. Arch Virol. 2010 Dec;155(12):2083-103. doi: 10.1007/s00705-010-0814-x. Epub 2010 Oct 30. PubMed PMID: 21046175; PubMed Central PMCID: PMC3074192.

[4] Miranda ME, Ksiazek TG, Retuya TJ, et al. Epidemiology of Ebola (subtype Reston) virus in the Philippines, 1996. J Infect Dis 1999;179:Suppl 1:S115-S119.

[5] "ViralZone: Ebolavirus." ViralZone: Ebolavirus. N.p., n.d. Web. 31 Oct. 2014. <http://viralzone.expasy.org/all_by_species/207.html>.

[6] Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, Calisher CH, Laughlin CA, Saif LJ, Daszak P. Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol Mol Biol Rev. 2008 Sep;72(3):457-70. doi: 10.1128/MMBR.00004-08. Review. PubMed PMID: 18772285; PubMed Central PMCID: PMC2546865.

[7] Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjies, P., & Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.*Bioinformatics*, *28*(12), 1647-1649.

[8] Ltd., Biomatters. "Geneious 7.0 Manual." (n.d.): n. pag. Web. 10 Dec. 2014. <http://www.unipos.net/download/Geneious7Manual.pdf>.