Discovering Host Specific Adaptive Mutations in Pathogenic Organisms

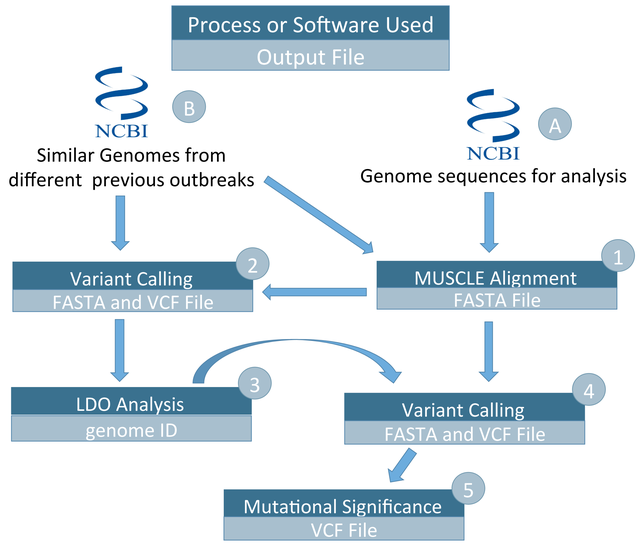
Kevin Chow, Luke Schuster, MengChi Tsai, Sarah Yeo
Mentored by: Justin Chu

Introduction

During outbreaks, mutations in pathogens can increase their fitness in a specific host. Important mutations are selected by the host environment and increase in frequency during the progress of an outbreak. We present an analysis pipeline that identifies these mutations and determines their potential significance. Our pipeline borrows from the concept of using recurrent mutations to find significant mutations often used in cancer analysis pipelines. We tested and validated our pipeline on data collected from the 2014 Ebola outbreak.

Objective: To develop an analysis pipeline that discovers important mutations that increase fitness of pathogen within a host.

Analysis Pipeline

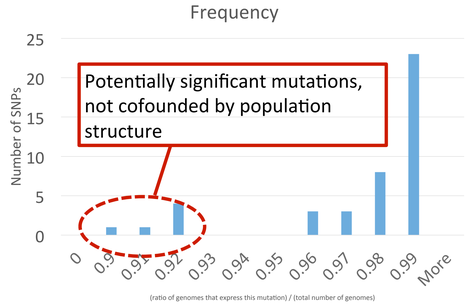


1. MUSCLE (multiple sequence alignment software) juxtaposes and pads the various genomes for analysis later on in the pipeline
2. Locates base pair differences between each genome in sample A and sample B. It does this by using each genome from set B as a reference against the FASTA file from step one.
3. The genome from set A which is consistently the least divergent from the genomes in set B is chosen as the "oldest" genome in set A.
4. Mutations that occurred during the outbreak of sample A are determined by comparing the FASTA file from step 1 with "oldest" genome found in step 3
5. The mutational differences found in step 4 are analyzed for potential mutational significance by analyzing percentages of the population that possess a found mutation.

Mutational Significance

The resulting data from the variant call in step five is displayed below. The "oldest" genome used a the reference is not actually the root of the phylogenetic tree but is simply the closest genome to that root from the available data set. Thus, the bimodal distribution bellow suggests that everything past 96% is actually because of mutations in the reference genome used.

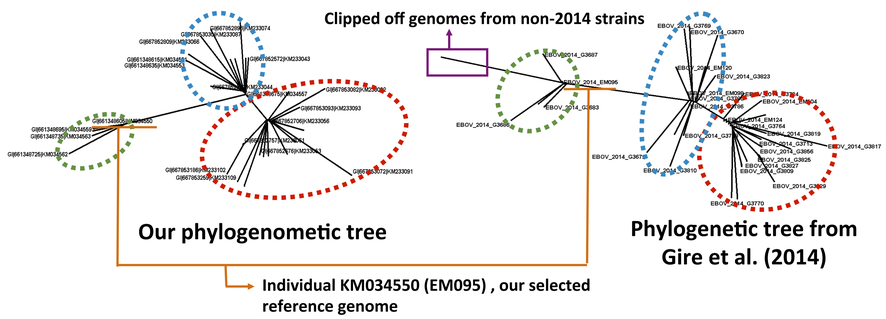
This data also collectively suggests that very few advantageous mutations occurred during the outbreak itself.



Validation of Methods

1. The strong similarity between our phylogenetic tree and a previously generated tree (Gire *et al.* 2014) suggests that our method for comparing and relating genomes must be reasonably accurate (fig. 1). The tree by Gire *et al.* contains many other Ebola strains (clipped off for visualization purposes).

Figure 1. Comparison of our phylogenetic tree and Gire's phylogenetic tree



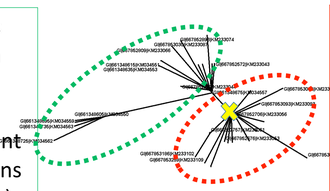
2. The "most distant" genome found from the 2014 outbreak, which was used as a reference to determine beneficial mutations that occurred during the outbreak, existed in close proximity on the phylogenetic tree to the start of the branch of the non-2014 strains used in Gire *et al.* thus suggesting that the methodology design to find our genome was both effective and accurate.

3. 97.13% precision of our pipeline in finding fixed mutations found by Gire et al. between the 2014 outbreak and other Ebola genomes, suggests accuracy in the alignment and variant calling process.

Phylogenetic Analysis

We also generated a phylogenetic tree from our data set, which can be used to identify potential confounding factors when we try to determine the relation between recurrent mutations.

e.g. If mutation A is found to be in both the red and green branch, it is more likely to be important (but only if mutations arise independently).



e.g. If genome A can only be found in red branch, it might be derived from the same ancestor (yellow cross) and not necessary important (founder effect).

Conclusion

Our pipeline can automatically process raw genome files from a database and generate results that can be beneficial to further pathogen research. The mutational significance test can help us determine the potential relationship between genotypes and phenotypes, while the phylogenetic data can help eliminate any confounding factors.

Future Work

1. Comparison of actual virulence of the various Ebola strains of the 2014 outbreak, determined through wet lab analysis, with the estimated virulence through mutational significance analysis.
2. Analysis of the resulting differences in amino acid patterns and protein structure of the Ebola virus, and estimations as to the functional differences these located mutations have had on the Ebola virus at a biochemical level.
3. Validation of method using data collected at different stages of an outbreak (early outbreak vs late outbreak).

References

- Danecek, P. et al. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156-2158.
- De Maio, N., Schlötterer, C., & Kosiol, C. (2013). Linking Great Apes Genome Evolution across Time Scales Using Polymorphism-Aware Phylogenometic Models. Molecular Biology and Evolution, 30(10), 2249-2262.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), 1792-1797.
- Gire, S. K. et al. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science, 345, 1369-1372.
- Huson, D. H., & Scornavacca, C. (2012). Dendroscope 3 - An interactive viewer for rooted phylogenometic trees and networks. Systematic Biology, 61(6), 1061-1067.
- Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenometic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics, 30(9), 1312-1313.
- Steve Haddock. (2013, April 17). Some Molecular Biology Scripts. Retrieved March 7, 2015, from Steve Haddock's Scripts: http://www.mbari.org/staff/haddock/scripts/