Team-II Comparative Genomics

Homework Assignment [50 points]

Due: Thursday, April 19th, 1:25pm

Instructions

1. Please **do not** run anything related to this assignment on the server
2. Please keep your answers brief and to the point
3. Submit your completed homework by emailing the final pdf saved as **LastName_FirstName_CG** to **prachitiprabhu@gatech.edu**

(I) General Comparative Genomics [15 points]

1. List 3 differences between the Whole Genome method and the Phylogeny method.

2. Is the difference in genetic sequence the only factor that can confer the resistant/sensitive/heteroresistant phenotype? Please elaborate.

3. Does sequence conservation imply 100% identity? Give an example to support your answer.

4. You have 20 clinical isolates of *Klebsiella pneumoniae* and you find that they are all resistant to fosfomycin.
   You treat each isolate with a FosA inhibitor and then plate on media containing fosfomycin and find that you still get colony growth in 10 out of the 20 isolates.
   You sequence all 20 isolates, assemble and annotate their genome. You find that none of the 20 isolates have any plasmids or phage integrations.
   What is the simplest and most likely cause of fosfomycin resistance in the isolates that were not affected by the fosfomycin inhibitor? Explain the workflow you would use to identify the cause. (10)

(II) Whole Genome Approach [15 points]

5. How can you measure distance between genomes? Describe at least two different methods. What are the benefits and drawbacks of each method?

6. A distance was computed between every pair of 150 genomes. The results is saved in a form of a distance matrix in the file *dist.txt*. Cluster the genomes based on this distance matrix.
   a. Use any hierarchical clustering tool/algorithm and visualize the result as a dendrogram.
   b. Briefly describe the clustering algorithm and parameters used.

c. How many distinct genome clusters seem to be there in this dataset based on the dendrogram?

(Hint: You can use R's hclust() function. Be sure to treat the input as a distance matrix, not feature matrix)

(III) Pan/Core Genome Approach [5 points]

7. What is the difference between core genome, pan genome, and variable genome?

8. Explain the rationale for Pan/Core genome analysis approach?

(IV) Phylogeny Approach [15 points]

9. MLST
    a. What are housekeeping genes? And how can they help us with sequence typing?

    b. Explain a situation where using MLST would be ideal and a situation where using MLST would not be the best approach.

10. SNP Analysis

    For this question, you will be need the documentation for kSNP.

    a. The first step of this tool involves k-mer selection. How does the software detect k-mers by default? List a command which would alter this step.

    b. Pretend you ran kSNP for your whole population, but were handed an additional 10 genomes. What would you need to do in order to add these to your existing run?

    c. When annotations are added along with the SNP analysis, how can you interpret the splits in the branches of the tree?

11. 16S Analysis
    a. Why was this method traditionally used? Why is the field moving away from it?