Team 1 Comparative Genomic Homework Assignment

All tools necessary for this assignment are already installed on the server.

Please email this assignment as LastName_FirstName_CP.pdf or .docx to you42@gatech.edu by midnight 4/19/2018 along with any additional files related to this assignment.

1. **Comparative Genomic** (20 points, 10 points each)
   a. Background: What is the difference in resolution between (traditional) MLST, (core genome) cgMLST, (whole genome) wgMLST, (pan genome) pgMLST?
   b. Bonus: What are the ESKAPE pathogens and why are they significant?

2. **GWAS** (40 points, 10 points each)
   Background information about GWAS can be found here: https://www.nature.com/articles/nrg.2016.132
   a. What are the main confounding factors in GWAS analysis? What are the main differences between human GWAS and microbial GWAS?
   b. What information can you obtain from Manhattan plots and quantile-quantile plot? How do you choose p-value to identify significant variants?
   c. What is pangenome? What is the purpose of generating the pangenome?
   d. Roary/Socary (export this path to your .bashrc file before usitn Scoary: /projects/data/team1_ComparativeGenomics/bin/Scoary)
      i. What information is given from the Roary output?
      ii. Try running Scoary: scoary -g [Roary output] -t [phenotype information], what information we can get from the Scoary output? **Use the attached Roary ouput and phenotype.txt**

3. **MLST** (30 points, 10 points each)
   a. What are the loci of the 7 housekeeping genes used in the classic MLST scheme for klebsiella-pneumoniae?
   b. Dr. Jordan's lab provides step by step manual for stringMLST (https://github.com/jordanlab/stringMLST).
      i. Follow the instruction from the manual, create database for klebsiella-pneumoniae located at **/projects/data/team1_ComparativeGenomics/bin/stringMLST/mlst_dbs/klebsiella-pneumoniae/**
      ii. predict ST from any samples located at **/projects/data/team1_ComparativeGenomics/Reads/upload**. *The answer should be lines of codes with output in form of table (1st row with title and 2nd line with profile corresponding to ST).
   c. What are the three main functions of STing?

4. **SNP Calling** (30 points, 10 points each)

   a. What is SNP calling? Can you tell the advantage and disadvantage of SNP calling comparing with the GWAS and MLST?

   b. Try to use samtools and bcftools to make variant calling with given GCF_002249975.1_ASM224997v1_genomic.fna and SRR3982229_sorted.bam(**path: /projects/data/team1_ComparativeGenomics/Assignment**)
   *NOTE:Ensure samtools and bcftools on your PATH. Please paste the command you used. Can you specify the parameters you choose?

   c. Try to use kSNP. Please paste the command you used. What kind of trees can kSNP generate. Can you simply explain their differences?