Omics Big Data 2019

Final Exam

12:01PM June 17, 2019 – 12:01 PM June 24, 2019

This is a take-home exam. However, you are required to do this exam independently. You are not allowed to discuss with other students in this class. The exam starts from 12:01PM June 17, 2019. You need to submit an electric version (answers as well as program codes or scripts via the course website) and a hard copy to room 4-221 biology building before 12:01 PM June 24, 2019.

If your result file is very big, please include it as a supplementary file in the electric copy or you can specify a path to your result file in our course server (202.120.45.100). Only a file with date before 12:01 PM June 24, 2019 will be accepted.

Note: There are seven problems. **You can pick three from the first four and two from the last three questions.** Each one will be 20 points.

Most questions are open. Students are encouraged to work on the exam as early as possible since it may take a long time to finish. Please try your best to finish as many questions as you can.

Either English or Chinese are allowed to answer questions （答题用英文或中文均可）.

1. **Determining sequencing platform and sequencing depth.**

   The sequencing error of platform A is about 15/100, and the read length can be as long as 100kb. The sequencing error of platform B is about 6/1000, and the read length can be about 150 bases. We want to use one or both of these two sequencing platforms to identify a mutation that happens about once in a million replicates. Which sequencing platform will you choose and what is the minimum sequencing depth we should do in order to determine with high confidence whether this mutation has happened or not? You can give your own criterion about the “high confidence”, and you need to show why this sequencing strategy or sequencing depth is enough.

2. **Motif finding.**

   Please find the top 20 most frequent k-mers (DNA fragments with length k bases) in the human genome (version GRCh38 or hg38) for each k (=3, …, 20). Compared to the background frequency (each base is treated as independent identical distribution), which k-mers have the top 20 enrichment scores (observed_frequency/expected_frequency) for each k (k=3, 4, 5, …, 20)? The human genome can be downloaded from genome.ucsc.edu or from NCBI website or you can access it (genome.fa) under directory /share/data/reference/hg38/ in our teaching server.

3. **Evaluation of the novel genes found in the rice pan-genome.**

   Please create a pipeline to evaluate the quality of the novel genes in the rice pan-genome (but not included in the reference rice genome). You need to answer what is the percentage of the novel proteins (with names started with "Un_maker_") in the pan-genome protein sequences that can be aligned to the reference rice genome with at least 30%, 50%, or 90% of the protein coding region length? For alignment tools, such as Blast or any other alignment tools that are comfortable for you to use, please explain why you choose this tool. Or you can develop your own program to do the alignment.

   The rice pan-genome and its annotations are available at http://cgm.sjtu.edu.cn/3kricedb/data_download.php. The reference rice genome is included in the pan-genome, i.e., chromosome 1-12. The protein sequences of the rice pan-genome are also provided in the above link.

4. **Planning a big omics research project.**

   Dr. Z is planning a project to investigate the impact of the host genome and gut microbiomes on a therapeutic food intervention. A cohort of 30,000 patients is going to be collected. At the first stage of the project, 100 patients will have their genomes sequenced (with 30x coverage), together with RNA-seq data (~30GB each) before and after a 3-month therapeutic food intervention. All these 100 patients will also have their gut microbiomes sequenced (at least 30GB per sample) before and after the 3 months' therapeutic food intervention. Including these 100 patients, a total number of 1,000 patients will have 16S rRNA sequenced (at least 30MB per sample) before and after the 3 months' therapeutic food intervention. Please give a strategy to cover the sequencing and data analysis stages. You need to show the strategy (with diagrams) and give your estimation about the time

and cost for the strategy. For example, you can pick the different sequencing platforms for genome, transcriptome and metagenome sequencing. For the 16S rRNA sequencing, 100 samples can be merged into one run of illumine sequencing. For the cost estimation, please give a reasonable estimation. For example, you can pick cloud computing or the Pi supercomputer from SJTU to do the data analysis. You can find current market price from computing resource providers such as cn.aliyun.com or www.huaweicloud.com.

If Dr. Z wants to do the sequencing in a scale 10 times bigger in the second stage of the project above, what strategy will you propose and what will be the cost in terms of time and money?

5. **Estimating the detectability of peptides .**

   Proteomics search engines identify peptides by comparing the theoretical and observed peaks in the MS/MS spectrum, and measure the fitness between theoretical and observed distribution of m/z of tryptic peptides.

   Using human proteome sequences, generate theoretical tryptically-digested peptides database with two misscleavages at most. Find out the percentage of detectable peptides, and evaluate the relationship between the detectability in MS/MS and the expression level, misscleavages, charge state and lengths of the candidate peptides. Graphs are encouraged to show the results.

6. **Correlation between gene expression and protein expression.**

   Of all proteome spectrum or RNA-Seq reads from human samples, only part of them can be mapped into human sequence. Estimate the percentage of unmatched reads or spectrums in human RNA-Seq or proteomics sample. How about the everage correlation between gene expression and protein expression? Which genes show the higher correlation, which not? Is there any rules?

7. **Where are the unexplained spectrums in shotgun proteomics from?**

   Many spectrums from real sample still cannot be explained by the known human protein. Please try to propose a possible reason, and test your hypothesis.

Some data for questions 5-7 are available at our teaching server under directory /share/data/proteomics/final-examination/, which include mgf and xml (searching results by X!tandem) files from two human samples and a fasta file of human proteome sequences.