Omics Big Data 2025

Final Exam

9:00AM June 9, 2025 – 11:59 PM June 15, 2025

This is a take-home exam. However, you are required to do this exam independently. You are not allowed to discuss with any other people. You need to submit an electric version (answers as well as program codes or scripts via the course website) to course canvas site.

If your result file is very big, please include it as a supplementary file or you can specify a path to your result files in our course server in Pi2.0. Please don't use the login node (pilogin.hpc.sjtu.edu.cn) for computation. We have reserved computational nodes for this final exam. You can access these two nodes by using "srun -p cpu --reservation=bio8402_class -n 4 --pty /bin/bash" to allocate a node to run your code or using "sbatch --reservation=bio8402_class sbatch_script.sh" to run the commands listed in sbatch_script.sh with the task management system. Only a file with date before 11:59 PM June 15th, 2025 will be accepted.

Note: There are 7 question in total. **Please 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 is allowed to answer questions （答题用英文或中文均可）.

1. **Motif finding for the Telomere-to-Telomere human genome.**

A Telomere-to-Telomere human genome (T2T human genome) was released recently. We want to find the motifs in this genome, which show much higher frequencies (for example, 3 times higher) than expected (the background frequencies). The first step to find the motifs is to find the most frequent M (for example M=10) k-mers (DNA fragments with length k bases) in this T2T human genome for different k values (k=2, 3, …, 20). Comparing to the background frequency (each base is treated as independent identical distribution), which k-mers have the top M enrichment scores (observed_frequency/expected_frequency) for each k (k=2, 3, 4, 5, …, 20)? Is M=2 is a proper value for this problem? If you use the human reference genome as the input, will the results be different?

The T2T human genome (chm13v2.0.fa) and the human reference genome (GRCh38.p14) can be downloaded from UCSC genome browser, NCBI website or you can access them under the directory **/lustre/share/class/BIO8402/exam/chr/ in our course server.**

2. **Evaluation of the 188 novel genes found in the Han Chinese pan-genome.**

Please create a pipeline to evaluate the quality of the novel genes in the Han Chinese pan-genome (but not included in the reference human genome with sequence identity percentage of 90% and with at least 90% of each protein coding region aligned). You need to answer what percentage of these novel genes can be aligned to the reference human genome with at least 30%, 50%, or 90% of the coding region of each gene. You can use any alignment tool that is comfortable for you to use, such as BLAST. Please explain why you choose this alignment tool. Or you can develop your own program to do the alignment.

The 188 non-redundant genes from 275 Han Chinese genomes missing in GRCh38 primary assembly, patch sequences and alternative loci (including their genomics sequences, their annotations, corresponding transcript sequences and protein sequences) are available at the bottom of this page: http://cgm.sjtu.edu.cn/hupan/download.php.

You can also use the T2T human genome as the reference for the analysis mentioned above. The T2T human genome (chm13v2.0.fa) and the human reference genome (GRCh38.p14) can be downloaded from UCSC genome browser, NCBI website or you can access them under the directory **/lustre/share/class/BIO8402/exam/chr/ in our course server.**

3. Planning a big omics research project.

Dr. Z is planning a project to investigate the impact of the host human genome and gut microbiomes on a therapeutic food intervention. A cohort of 200,000 patients is going to be collected. At the first stage of the project, 200 patients will have their genomes sequenced (with 30x coverage), together with RNA-seq data (~30Gbps each) before and after a 3-month therapeutic food intervention. All these 100 patients will also have their gut microbiomes sequenced (at least 10GB per sample) before and after the 3 months' therapeutic food intervention. Including these 200 patients, a total number of 2,000 patients will have their 16S rRNA genes sequenced (at least 30Mbps per sample) for their gut microbiota 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 or whatever sequencing platforms you know for sequencing. For the 16S rRNA sequencing, 500 samples can be merged into one run of illumine sequencing. For the cost estimation, please give

any reasonable estimation. For example, you can pick cloud computing or the Pi2.0 supercomputer from SJTU to do the data analysis. You can find current market price from any computing resource providers such as cn.aliyun.com or www.huaweicloud.com. If you use Pi2.0, the price can be set to ~0.02￥/(core*hour) for computing and ~50￥/(TB*year) for storage.

If Dr. Z wants to do the sequencing in a scale 100 times bigger in the second stage of the project above, what will be the cost in terms of time and money? Give your answers for these questions based on the best of your knowledge.

4. **Determining sequencing depth.**

The sequencing error of a platform is about 5/1000. We want to use this sequencing platform to identify a mutation that happens about once in a million replicates. 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 depth is high enough.

5. **Identifying novel proteins or peptides by combining transcriptome and proteome data in cancers**

Given that a dataset was collected from a liver cancer cohort, in which the transcriptome (RNA-Seq) and proteome data were generated from 100 tumor samples and the matched tissue adjacent the carcinoma. Please propose a data analysis workflow to identify the novel proteins or peptides through combing the transcriptome and the proteome data.

6. **Relationship between protein-coding abundance and gene length, as well as other sequence characteristics.**

It was reported that short proteins are expressed with low abundances. Other sequence features also may impact protein abundance, such as like GC content, base bias and et al. To clarify these points, please study the relationships between the protein-coding gene abundances and their lengths and other sequence related features by a statistical analysis of transcriptomic or proteomic data. The omics datasets and species can be freely selected.

7. **Integration between single-cell transcriptome and bulk RNA-seq.**

Describe the key challenges and methodologies for integrating single-cell RNA sequencing (scRNA-seq) and bulk RNA-seq data, emphasizing their complementary roles in biological research. Include specific examples of integration strategies, computational tools, and applications in fields such as cancer biology or developmental biology.