# Omics Big Data

Spring 2019

## Project 1, week 5

**Which genome annotation set is the best for the rice reference genome?**

The goal of this project is to evaluate the quality of the three genome annotation datasets for the rice genome.
There are three commonly used genome annotations for the reference rice genome. You can access them from
http://rice.plantbiology.msu.edu/ (known as MSU rice genome annotation, Release 7),
http://rapdb.dna.affrc.go.jp (known as RAP-DP) and
ftp://ftp.genomics.org.cn/pub/ricedb/rice_update_data/bgf_genes/irgsp/ (known as BGI rice database). The naming may be a little bit misleading, but we will stay with it for this project.

You can also find these annotation files from the server /share/home/ccwei/courses/2019/omics/proj1/gff.

We have collected 226 RNA-seq datasets for rice. You can access some of them at /share/home/ccwei/courses/2019/omics/proj1/rice_rna. The rice reference genome is also given to you at
/share/home/ccwei/courses/2019/omics/proj1/reference/IRGSP-1.0_genome.fasta for the MSU and RAP-DP annotation. The reference genome for BGI rice annotation is /share/home/ccwei/courses/2019/omics/proj1/reference/IRGSP4_genome.fasta.

You are expected to evaluate two rice genome annotation data by these RNA-seq data. Please pick at least 5 RNA-seq data, and align these RNA-seq data to the rice genome and check how many of the genes/transcripts can be validated by at least 2 RNA-seq datasets. You can use any sequencing alignment (or mapping ) tools to align these RNA-seq data back to the rice genome. You are expected to report the description of steps in your analysis pipeline, the reasons why you choose these tools for each step, together with the following results.

1. The total number of RNA-seq reads for your analysis;
2. The number of reads that can be aligned to the rice genome (with the criteria you use) for each RNA-seq dataset;
3. The number of reads that can't be aligned to the rice genome;
4. For all genome annotation sets,
   a) How many of the aligned reads locate inside an exon, i.e., not include exon junctions;
   b) How many of the aligned reads include exon junction;
   c) How many genes are validated with at least two exons with their junctions supported with at least k reads, where k=1, 5, 10 and 50;
   d) the percentages of exons in the rice genome that can be covered by at least k reads, where k=1, 5, 10, and 50;
5. (bonus)You are assigned with a rice chromosome. In order to get a full credit of this project, you need to have results on at least this chromosome. You will get some bonus if you finish

your analysis for more than one chromosome.

The assignment of chromosomes to students.

| Student ID | Assigned chromosome |
|---|---|
| 018080910001 | Chr7 |
| 018080910011 | Chr2 |
| 018080910019 | Chr3 |
| 018080910063 | Chr12 |
| 118080910007 | Chr4 |
| 118080910008 | Chr8 |
| 118080910029 | Chr9 |
| 118080910031 | Chr11 |
| 118080910053 | Chr5 |
| 118080910056 | Chr1 |
| 118080910064 | Chr6 |
| 118080910075 | Chr10 |
| 118080910108 | Chr7 |
| 118080990005 | Chr2 |
| 118080990009 | Chr8 |
| 118150990017 | Chr10 |
| 515760910050 | Chr6 |
| 515111910093 | Chr5 |

This is a real research project. The lecturer does not know the result either. Therefore, you need to give detailed information about each step in your analysis so that your result can be reproduced by others if it is needed.

Students are encouraged to form a team of two (at most two), and submit a report for the whole team instead of a report for each individual. However, you have to describe the contribution of each individual in the report if you choose to work as a team instead.

Note: You are expected to use the task management system OpenLava to submit all your computation jobs. Please don't run your job in the management node. A brief introduction for OpenLava system is in our server (not the course webpage) under the directory /share/home/ccwei/OpenLava. Please download it to your own computer and try to use the task management system for all the project.

**Turning in your project work**

**Submit your homework to** http://cgm.sjtu.edu.cn/test/obd/index.html. Please submit it before 10:00AM on April 24th, 2019. You are also required to submit a hard copy before the class start on April 24th, 2019.

Ask **TA Huimin Lu:** linuslu6@outlook.com in case you have any question about the homework submission webpage.

-------------------------------------------------------------cut-----here--------------------------------------------------

独立作业承诺：（请选择一个，　并签名）

Confirmation of independent homework: (please select one and sign you name)

1. 本人，__________，保证本次作业由自己独立完成。

   I, _________________, confirm that this homework was done by myself independently.

   签名

   Signature

   时间　　　年　月　日

   Date (Day/Month/Year)　　/　　/

或者

or

2. 本人，__________，保证本次作页和_______________同学讨论后，由自己独立完成。

   I, ______________, confirm that this homework was done by myself independently after discussing with _____________.

   讨论内容包括_____________________________________________________

   The contents of discussion include___________________________________________

   签名

   Signature

   时间　　　年　月　日

   Date (Day/Month/Year)　　/　　/