

### **Assignment Goals**

- i. Resolve read mapping uncertainty in RNA-Seq quantification.
- ii. Get familiar with the workflow of machine-learning modeling.
- iii. Browse online databases such as Gene Expression Omnibus.

### **Submission Instructions**

- To turn in your assignment, please log in to the server **pluto.biostat.wisc.edu** using your **pluto** username and password.
- Copy all relevant files to the directory

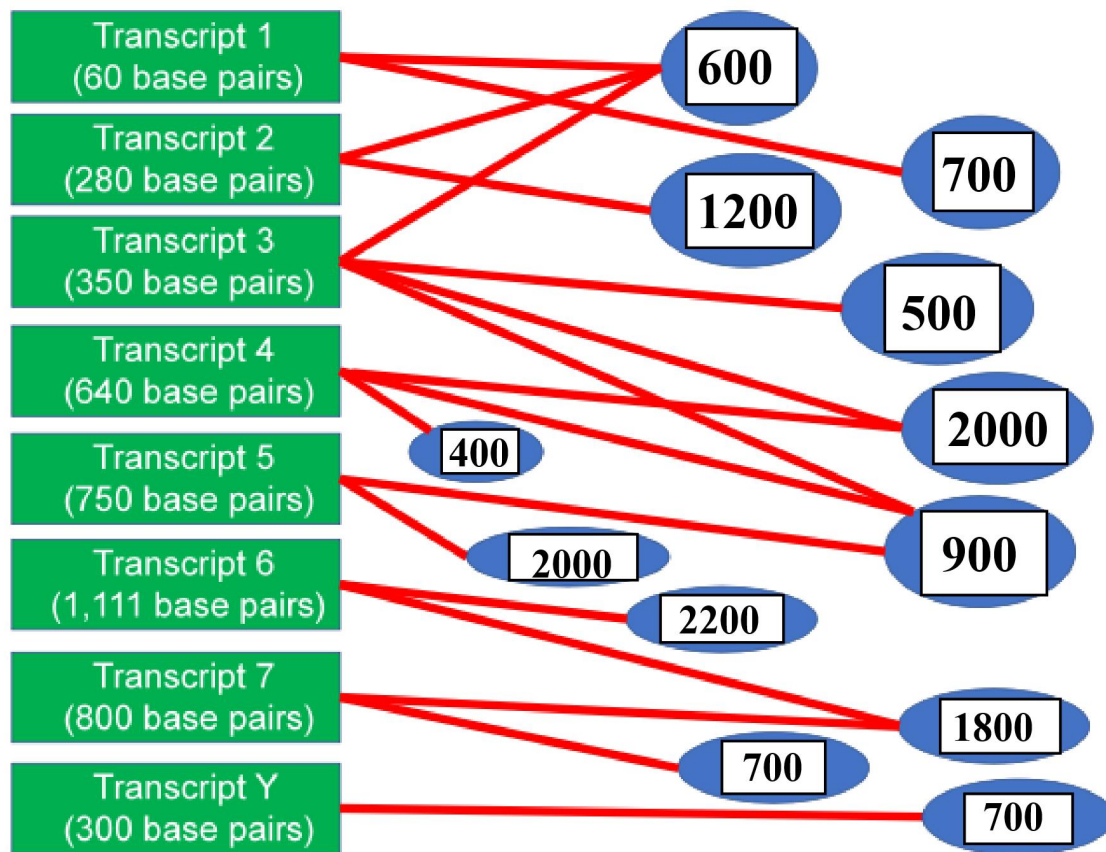
**`/medinfo/bmi776/bmi776-s25/hw1/<USERNAME>`**

where **<USERNAME>** is your pluto (biostat) username. Please submit all of your Python source code and test that it runs on the pluto server.

- For the rest of the assignment, compile all your answers in a single file and submit as **solution.pdf**.
- Write the number of late days you used at the top of **solution.pdf**.
- For the written portions of the assignment, show your work for partial credit.

### **Part 1: RNA-Seq Rescue Algorithm (35 points)**

The full RSEM algorithm is too complicated to execute manually, but we can use the RNA-Seq rescue method to approximate one iteration of expectation-maximization. The bipartite graph below contains two types of nodes: transcripts and read groups. The transcript nodes contain a transcript ID and the transcript length in base pairs (bp). The read nodes contain the read counts for a group of reads that all align to the same transcripts. The edges designate the transcripts to which each read group aligns.



(A) Use the rescue method to calculate the *relative* abundance for the 8 transcripts. (20 points)

(B) Transcript Y is an RNA spike-in. 900 copies of transcript Y were mixed into the experimental sample when preparing the sample for RNA-Seq, meaning its absolute abundance is 900. Use the relative abundance from (A) to calculate the *absolute* abundances for the other 7 transcripts, rounded to the nearest integer. (15 points)

**Part 2 Machine-Learning Modeling (45 points)**

It is important to better understand genetic mechanisms of critical illness (severe outcomes) in Covid-19 patients. In collaboration with biologists and front-line researchers, you as a bioinformatician want to build a predictive machine-learning model for Covid-19 severity based on changes in global gene expression of blood gene expression data of hospitalized human patients. You receive data such as that in this study (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157103>); please note that this dataset is just an example, for your reference, but you do not need to use this dataset to answer this problem. In general, being in the Intensive Care Unit (ICU) tends to mean the patient is undergoing severe health problems (whether by Covid-19 for infected patients or some other health cases for negative controls (groups 3 and 4). The key information is as follows: you have 4 groups of patients in the data:

**Group 1:** The 50 patients who are hospitalized with severe Covid-19 (in the Intensive Care Unit (ICU)).

**Group 2:** The 50 patients who are hospitalized with Covid-19 (but it is not severe)

**Group 3:** The 13 patients who are hospitalized in the Intensive Care Unit for some other severe condition (they do not have Covid-19 and have never been exposed).

**Group 4:** The 13 patients who are hospitalized for some other condition (they do not have Covid-19 and have never been exposed) that is not severe.

Your collaborators performed RNA-Seq and obtained measurements for 24,000 genes following exposure for these 126 different human samples, with five biological replicates for each gene.

In this data, you thus have access to whether a patient has been exposed to Covid-19 or not, as well as to different levels of severity and can ask a few different questions.

**(A)** What are some of the questions you could address using Differential Gene Expression Analysis (DGEA)? That is, how could you compare groups with one another or group them together to derive some meaningful statistical analysis from this multi-class data? Which of these questions would you like to focus on for parts B and C, and why (please pick 1 as your research question). Hint: the questions typically are classification-based. **(5 points)**

**(B) (Exploratory data analysis)** It is common practice to apply unsupervised learning methods (clustering, dimensionality reduction, etc.) on the measurement data in order to understand the intragroup variability among replicates and the intergroup variability among samples with different outcomes. Based on your research question of interest for part A, please outline *two* unsupervised learning methods that you think would serve this purpose. Please describe what you would expect these methods to uncover. **(15 points)**

**(C) (Learning a classifier)** Now that you have gained some insight from the data through unsupervised learning, you would like to proceed and build a support vector machine (SVM) to address this same research question you chose from part A (please choose 1). Given a gene expression profile of an individual based on the blood serum levels, the SVM will perform classification. Please describe a workflow for training and evaluating the SVM. Please beware of the high feature dimensionality relative to the sample size and the class imbalance problem **(15 points)**.

Please find another dataset of your choosing from Gene Expression Omnibus dataset browser (<https://www.ncbi.nlm.nih.gov/sites/GDSbrowser/>) and include the following information below. For instance, the screenshot below shows how you can analyze studies related to small cell lung cancer. **(10 points)**

The screenshot shows the NCBI GDS Browser interface. At the top, there is a search bar with the text "lung cancer" and buttons for "Search", "Clear", "Show All", and "Advanced Search". Below the search bar, a table lists 59 dataset records. The table has columns for "DataSet", "Title", "Organism(s)", "Platform", "Series", and "Samples". The dataset GDS4794 is highlighted in blue. Below the table, there is a detailed view of dataset GDS4794, including its title, summary, organism, platform, citation, reference series, sample count, and series published date. The detailed view also includes a "Cluster Analysis" section with a heatmap and a "Download" section with links to various file formats.

DataSet	Title	Organism(s)	Platform	Series	Samples
GDS5410	Non-small cell lung cancer H460-derived cancer stem cell differentiation	<i>Homo sapiens</i>	GPL4133	GSE54712	8
GDS5418	Glucose effect on steroid receptor coactivator 1 deficient-A549 lung cancer epithelial cell line	<i>Homo sapiens</i>	GPL570	GSE56843	8
GDS5648	Chronic high-calorie diet effect on Kras-driven lung tumors	<i>Mus musculus</i>	GPL6887	GSE56260	11
GDS5004	Adult alveolar type 2 and embryonic bipotent progenitor lung cells	<i>Mus musculus</i>	GPL1261	GSE49346	6
GDS5409	Akt inhibitor MK206 effect on influenza H1N1 infection of non-small cell lung cancer line	<i>Homo sapiens</i>	GPL10558	GSE54293	8
GDS5391	Protein tyrosine kinase PTK7 knockdown effect on lung adenocarcinoma cell lines	<i>Homo sapiens</i>	GPL6244	GSE50138	8
GDS5040	V-ets erythroblastosis virus E26 oncogene homolog 2 knockdown effect on lung cancer cells	<i>Homo sapiens</i>	GPL6244	GSE43459	6
GDS5067	Oligonucleotide effect on hypoxic alveolar adenocarcinoma cell line	<i>Homo sapiens</i>	GPL6244	GSE48134	15
GDS4794	Small cell lung cancers	<i>Homo sapiens</i>	GPL570	GSE43346	65
GDS4767	Breast cancer model: kailuo subnormalant	<i>Mus musculus</i>	GPL1361	GSE47811	12

**DataSet Record GDS4794: (Expression Profiles) | Data Analysis Tools | Sample Subjects**

**Title:** Small cell lung cancers

**Summary:** Analysis of 23 clinical small cell lung cancer (SCLC) samples from patients undergoing pulmonary resection and 42 normal tissue samples including the lung. SCLC is a lung cancer subtype with poor prognosis. Results provide insight into the molecular mechanisms underlying SCLC.

**Organism:** *Homo sapiens*

**Platform:** GPL570: [HG-U133\_Plus\_2] Affymetrix Human Genome U133 Plus 2.0 Array

**Citation:** Sato T, Kaneda A, Tsui S, Iagawa T et al. PRC2 overexpression and PRC2-target gene repression relating to poorer prognosis in small cell lung cancer. *Sci Rep* 2013;3:1811. PMID: 23714854

**Reference Series:** GSE43346

**Sample count:** 65

**Value type:** count

**Series published:** 2013/06/12

**Cluster Analysis**

**Download**

- DataSet full SOFT file
- DataSet SOFT file
- Series family SOFT file
- Series family MINIMAL file
- Annotation SOFT file

- Screenshot of the GEO download page
- Link to dataset
- Title of dataset

- Paraphrased summary of dataset
- Organism studied
- # of samples
- Type of data (e.g. bulk RNA-seq, etc.)
- Potential question(s) that could be addressed from this data (such as in (A))
- Other information