STAT115 Homework 2: Sample classification, RNA-seq analysis

(your name)

2017-02-11

"The difference between theory and practice is greater in practice than in theory".

Part I– High Level Microarray Analysis (Clustering and Classification)

The sample data is in file **"taylor2010_data.txt"** included in this homework. This dataset has expression profiles of 23974 genes in 27 normal samples, 129 primary cancer samples, 18 metastasized cancer samples, and 5 unknown samples. Assume the data has been normalized and summarized to expression index. The skeleton R code is in **HW2_1.R**. Please fill in missing R code for each question, and attach this R file with homework submission. You could fill in the code in the script and include the results and answer to the questions in the write-up, or filling the corresponding part in this R markdown as well as the results and answer. But make sure your code can be run through by the TAs on their computer.

Clustering (unsupervised learning)

1. Try hierarchical clustering average linkage on the samples (normal, primary, metastasized and unknown) using all genes, and plot the dendrogram.

Hint: use functions dist, hclust, plot

2. Do PCA biplot on the samples with all genes, and use 4 different colors to distinguish the 4 types of samples (normal, primary, metastasized and unknown). Do the samples from different groups look separable?

Hint: use the PCA ggplot R code, also function legend is useful. (http://docs.ggplot2.org/0.9.3.1/geom_point.html)

3. What % of variation in the data is captured in the first two principle components? How many principle components do we need to capture 85% of the variation in the data?

RHint: use function prcomp.

4. For the 174 samples with known type (normal, primary, metastasized), use LIMMA to find the differentially expressed genes with fold change threshold 1.5, and adjusted p-value threshold 0.05. How many differentially expressed genes are there?

5. Repeat the hierarchical clustering and PCA on all samples with the differentially expressed genes found in 3. Does the result look better? Why do you think it is the case?

6. Based on the PCA biplot, can you classify the 5 unknown samples? Put the PCA biplot in your HW write-up, and indicate which unknown sample should be classified into which known type (normal, primary, metastasized). Do you have different confidence for each unknown sample?

7. Also try k-means clustering on all samples using the differentially expressed genes.

Hint: function kmeans

8. For GRADUATE students: How do you determine the number of clusters for K-means clustering? Three clusters or four clusters, which one seems to work better?

Hint: http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

Classification (supervised learning)

9. Use differentially expressed genes in the known samples to perform LDA, and predict the unknown samples. Hint: use MASS library and function lda.

10. What is the "K" in K-Nearest-Neighbor (KNN) algorithm? What is the problem if K equals N (the number of data points)?

11. Run KNN (try K = 1, 3 and 5) on all the genes and all the samples, and predict the unknown samples based on the 174 labeled samples. Hint: use library class and function knn.

12. Run SVM (try a linear kernel) on all the genes and all the samples, and predict the unknown samples based on the 174 labeled samples. Hint: use library e1071 and function svm.

13. In K-fold cross-validation, what does "K" mean? Why it is necessary to shuffle the samples before doing K-fold cross-validation?

14. For GRADUATE students: Implement a 5-fold cross validation to estimate the classification error rate of KNN and SVM, based on the 174 samples with known label. Which of the method has lower estimated classification error rate?

Part II. RNA-seq analysis

For this RNA-seq analysis, we will use the data from: Xu H et al. NUP98 Fusion Proteins Interact with the NSL and MLL1 Complexes to Drive Leukemogenesis. Cancer Cell 2016. The raw sequencing data is at GEO with six FASTQ files: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75997

Normally for paired-end RNA-seq, each sample will have two separate FASTQ files, with line-by-line correspondence to the two reads from the same fragment. This data is a little unusual in that the two reads of the same fragment are written in the same file. We have split each FASTQ file into two, which is the standard format most RNA-seq mapping and analysis algorithms take. In order for students to test the different mapping algorithms quickly, we will only use 5M paired-end reads from 40H Sample The 5M fragments from this sample have been split into paired ends, and are available on Odyssey under /n/stat115/2017/HW2/Homework2/FastqData. We are going to use the RNA-seq workflow in BioConductor to analyze the RNA-seq data(http://www.bioconductor.org/help/workflows/rnaseqGene/).

1. Use STAR (Dobin et al, Bioinformatics 2012) to map the reads to the mouse reference genome, available on Odyssey at /n/stat115/2017/HW2/Homework2/STARIndex/

- For RNA-seq read mapping, a popular mapper is TopHat (Trapnell et al, Bioinformatics 2009): https://ccb.jhu.edu/software/tophat/index.shtml. However, if the data is paired-end sequencing with huge read count, STAR is much faster.
- For STAT115 students, please use STAR to map the 5M paired-end reads in the two FASTQ files for 4OH Sample 1.(4OH1_5M_a.fastq;4OH1_5M_b.fastq) How many fragments are mappable?
- For GRADUATE students, please use 5 million fragments from 4OH Sample 1 to test both TopHat(Index is at /n/stat115/2017/HW2/Homework2/bowtieIndex) and STAR for read mapping. Also compared paired-end read mapping (use both FASTQ files in the 5 M 4OH Sample 1) and single end read mapping (4OH1_5M_a.fastq).

- FOR GRADUATE students, for the above 4 scenarios (TopHat / STAR, PE / SE), evaluate the mapping on two metrics: # of mappable reads (for SE) or fragments (for PE), and time.

2. The above test will allow you to learn how to use the read mappers well. We have mapped the full dataset already and put the BAM files in /n/stat115/2017/HW2/Homework2/BamFile. Use HTSeq (Anders et al, Bioinformatics 2014) to count the reads for each gene from the BAM files. (Annotation File is on /n/stat115/2017/HW2/Homework2/)

3. Use DESeq2 (Love et al, Genome Biol 2014) to look at differential expression. How many RefSeq transcripts are up vs down-regulated by MLL1?

4. Can you generate some visual diagnostic figures to check your overall data quality?

Hint: clustering samples.

5. Use some GO analysis tool to look at whether there are enriched functions / processes / pathways that are regulated by MLL1?

6. For GRADUATE students: The exciting thing about bioinformatics is how fast things move. Since we started covering RNA-seq analysis, every year a new and potentially better algorithm comes along. This year is no exception. Recently a pair of new algorithms were developed for differential RNA-seq analysis from the TopHat / Cufflinks developers. They haven't been published yet, but have already received good feedback from early adopters. Instead of mapping to the whole genome, Kallisto only maps to the RefSeq transcriptome, so can go from FASTQ to read count on genes in a snap!

Try using Kallisto (http://pachterlab.github.io/kallisto/) on the full FASTQ data (2 FASTQ for each of the 6 samples, /n/stat115/2017/HW2/Homework2/FastqData.) to map the reads to the mouse RefSeq transcriptome. (/n/stat115/2017/HW2/Homework2/transcriptome)

7. For GRADUATE students: Run Sleuth (http://pachterlab.github.io/sleuth/) on the results from Kallisto for differential expression analysis

8. For GRADUATE students: Compare the results of Kallisto + Sleuth with STAR + DESeq2

- Which one is better, why (hint: features, ease of use, correctness of the differential genes from visual plots or GO enrichment)? It is possible that there is no right or wrong answer here, as long as you can justify your decision here.

- You can see the field is still very competitive and there are new tools coming out often. Whether they will be winner depends on solid statistics, good programming, other useful features, and user-friendly interface.

Part 3 Python exercise

Write a python program to merge two paired-end Fastq files (4OH1_5M_a_sub.fastq;4OH1_5M_b_sub.fastq in /n/stat115/2017/HW2/Homework2/FastqData) in to one.

1) Read in the two files into python
2) Distinguish quality ID and quality score, and sequence Id and sequence information
3) Match the sequence according to their sequence ID
4) Write out the results in csv files (remember to name your file with .csv suffix so that it can be opened directly by excel)

The results should look like below:

Sequace ID,sequence a,sequence b,sequence a quality,sequence b @SRR2994643.15099964 15099964 length=50, CTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT, AAAAAAAAAAAAAAAAAAAAAAAAAA CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG, BBB@ BGGGGGGGGGGGFGGGGGFGGGGGD;?/:9??9/999:9::>9?C?

Submission

Please submit your solution directly on the canvas website. Please provide your code (.Rmd) and a pdf file for your final write-up. Please pay attention to the clarity and cleanness of your homework. Page numbers and figure or table numbers are highly recommended for easier reference.

The teaching fellows will grade your homework and give the grades with feedback through canvas within one week after the due date. Some of the questions might not have a unique or optimal solution. TFs will grade those according to your creativity and effort on exploration, especially in the graduate-level questions.