CSci 3003: Introduction to Computing in Biology

Lab Assignment #7

20 points
Assigned: 11/07/19
Due: Tuesday, 11/19/19 (before 11:55pm)

Lab 7: Statistical analysis of gene expression data in R

Goals of this lab:

- Practice using R to load/analyze data.
- Learn about statistical analyses of gene expression data.

This lab will focus on statistical analysis of gene expression data using both microarray and RNA-seq datasets. Specifically, we will investigate genes whose expression is different between estrogen receptor (ER) positive and ER negative breast cancer tumors. The ER status reflects whether a tumor will be responsive to hormone therapies and is often used as an important diagnostic in determining the course of treatment. In this lab, we will examine gene expression differences between breast cancer tumor samples using both microarray and RNA-Seq expression data.

Part I: Understanding the breast cancer study

This week's lab will focus on understanding two expression datasets, loading the data into R, and performing some manipulations on the data. Before we begin doing statistical analyses of the gene expression data, it is important that you take some time to understand the study. We will be examining expression data generated from microarrays as well as from RNA-Seq. These datasets were generated from two studies (cited at the end of this document), and the data have been uploaded to the NCBI Gene Expression Omnibus and the NCBI Short Read Archive. Examine the studies, GSE20711 and SRP005601, by clicking the following links:

- http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20711
- http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP005601

Answer the following questions:

a) What was the purpose of each study and what experiments did the authors do to answer the question they were interested in?

b) How many **gene expression** samples are in each dataset?

c) What two sets of tumor samples did the authors profile for gene expression?

d) How large are the raw data files for each study?

- hint: GSE20711: GSE20711_RAW.tar

- hint: SRP005601: see “Related SRA Data – Runs”

Part II: Loading and manipulating expression data in R

We have provided prepared data files for each experiment that can be downloaded from the Canvas Website.

Write a script that will do the following:

a) Load the microarray gene expression files into R matrices using the `read.table()` function. We will want to use a matrix to interact with the data, so create a matrix from data frame you just made using the `as.matrix()` function. In the matrix, we want each column name to be a tumor ID and each row name to be a gene name.

- BC_MicroArray.txt - Microarray expression data: 32,864 probes (genes) x 90 samples (the values are log-transformed intensities from an Affymetrix array)

- BC_MicroArray_status.txt - The status labels for the samples (ER+ or ER-)

b) Write a short series of commands that will check the structure of the data matrices you’ve just loaded. The microarray data should contain values for 32,864 probes (genes) by 90 samples.

c) Print the data for the *BRCA1* gene for all samples.

d) Print the values for the first 10 genes for the first ER+ sample (GSM519791) in the microarray dataset.

e) Compute the mean and standard deviation of the expression data for the *BRCA1* gene across all samples.

f) Compute the mean and standard deviation of the expression of all genes in the first ER+ sample (GSM519791).

g) Print the list of 10 genes with the highest expression values in order of decreasing expression based on the first ER+ sample (GSM519791).

h) Create a vector called `ERpos_samples` that contains the GSM names of all ER+ samples.

i) Create a vector called `ERneg_samples` that contains the names of all ER- samples.

j) Use the vectors created in Parts (h) and (i) to compute a new vector `expr_difference`, which contains the difference in the mean expression for each gene between the ER+ and ER- sample groups. Positive values in this vector should indicate where mean ER+ expression is greater than mean ER- expression, and negative values should indicate where ER- < ER+ expression (Hint: matrices can be indexed by name and there are functions that compute means on matrix rows and columns).

k) Using the `expr_difference` vector you just created, print the gene that has the largest positive difference (i.e. ER+ > ER-) and the largest negative difference (i.e. ER+ < ER-) in mean expression between the ER+ and ER- groups. (Hint: you may find the `names` function useful here).

Part III: Statistical analysis of microarray gene expression data

For this part of the lab, we will find a subset of the genes that are differentially expressed using the *t*-test. The purpose of the *t*-test is to statistically assess a gene's difference in expression across two groups of samples (in this case, ER+ and ER-), and answer the yes/no question "is this difference significant?"

To perform a *t*-test to determine if one gene's differential expression is significant, you must first calculate a *t*-statistic. For a single gene, this is:

$$t = \frac{\underline{X}_1 - \underline{X}_2}{\sqrt{\frac{s_{x_1}^2}{N_1} + \frac{s_{x_2}^2}{N_2}}}$$

where $\underline{X}_1$ is the mean of the ER+ expression levels for that gene, $\underline{X}_2$ is the mean ER- expression level, N_1 and N_2 are the number of samples per group (ER+/-), $s_{x_1}^2$ is the variance of the ER+ expression values for that gene and $s_{x_2}^2$ is the variance of the gene's ER- expression.

a) Create a vector of ***t*-statistics** measuring the difference in expression for each gene between the ER+ and ER- samples (this vector should be 32864 x 1 or 1 x 32864). (Hint: the functions `mean()`, `var()`, and `length() or dim()` will be useful).

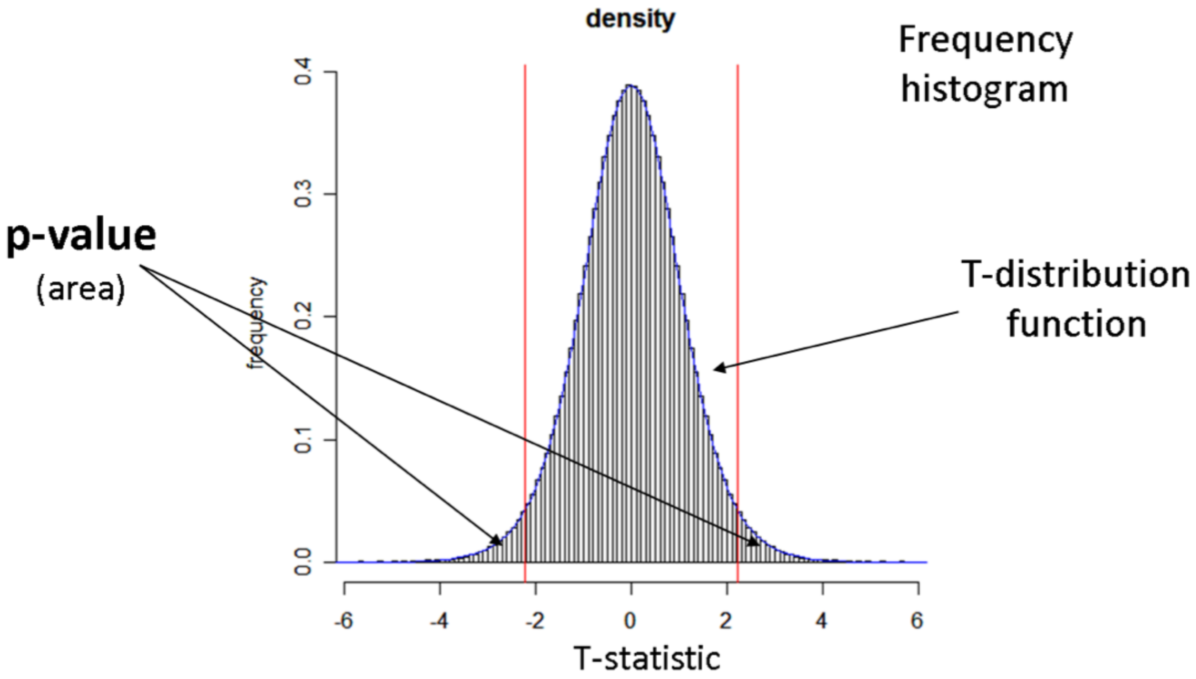
b) Use the `hist()` function to plot a histogram of *t*-statistics. (Hint: `hist(myvect,breaks=100))` will create a histogram of the values in `myvect` with 100 bins.

c) Your goal is to answer the yes/no question of whether or not each gene's expression value is significantly different from the random expectation in each sample. To do this, you will need to compute a p-value that reflects the significance of the *t*-statistics you just measured. A p-value represents the probability that the observed *t*-statistic was derived from samples where there was *no significant difference in the means of the two groups.* Thus, a small p-value reflects a case in which the gene is very likely differentially expressed between the two sets of samples. Typically, a cutoff of 0.05 is used; genes with less than a 0.05 p-value are called significantly differentially expressed.

Modify the code you created above to also compute a vector of p-values, one for each gene. (Hint: you will need to use the `pt()` function here, which computes the integral under the *t*-distribution for you; N = N_1+N_2) Because we want to test for positive and negative differences between the two groups at the same time, we will use a two-sided *t*-test, which computes the integral under the *t*-distribution for the ranges $-\infty \rightarrow -|t_{stat}|$ and $|t_{stat}| \rightarrow +\infty$. Remember, this is how we compute the probability of observing, under the *null hypothesis* that the gene is not differentially expressed, a value at least as extreme as the one you observed (on either side of the *t*-distribution). You can compute this probability using built-in R functions as follows:

```
pval = 2 * pt(-abs(tstat),N-2)
```

where N is the total number of tumor samples (both ER+ and ER-).
(`pval = pt(tstat,N-2)` is appropriate for performing a one-sided *t*-test)



d) Extra credit (2pts):

- Explain in detail why this works for computing the p-value for a one-sided test (1pt):
  `pval = pt(-abs(tstat),N-2);`
- Explain in detail why this works for computing the p-value for a two-sided test (1pt):
  `pval = 2 * pt(-abs(tstat),N-2);`

e) Using the results you computed for parts 1-3, answer the following questions by writing short sections of R code:

i.) Print a list of significantly differentially expressed genes (p-value < 0.05), sorted in order of their significance (from smallest to largest). Include the gene name, the *t*-

statistic, and the corresponding p-value in your list. Hint: either the `order` function, or the `sort` function with the `index.return=TRUE` option, will be useful.

ii.) How many genes are significantly differentially expressed at a p-value < 0.05? How many genes are significantly differentially expressed at a Bonferroni-corrected p-value < 0.05?

iii.) How many genes are significantly differentially expressed and more highly in ER+ tumors relative to ER- tumors?

iv.) How many genes are significantly differentially expressed and more lowly in ER+ tumors relative to ER- tumors?

v.) Select a gene from the top 10 most significantly differentially expressed and describe what's known about its function (http://www.genecards.org will be useful).

Part IV: Using R Packages to Process RNA-Seq Data

As we discussed in class, RNA-Seq technology is the current technology of choice for measuring gene expression profiles. Typical RNA-Seq datasets contain hundreds of millions of raw sequencing reads that need to be processed with specialized methods in order to perform statistical analysis. Fortunately, corresponding advances in computational and statistical approaches have paralleled this leap in experimental technology. A few well-established analysis "pipelines" are now responsible for processing most RNA-Seq data generated today.

Here, we have processed an RNA-Seq breast cancer data set comparing ER+ and ER- gene expression differences into raw read counts. This allows you to run a comparable differential expression test based on RNA-Seq data. In this portion of the lab, you will download and run an R package that will identify differentially expressed genes related to ER+/- status.

Open the provided script, `Part4.R`, and run the code. You will need to modify/add code for parts **b** and **f**. For the remaining parts, study the code that accomplishes each of the following:

a) Run the code to download and install DESeq2

b) Add code to read in RNA-Seq files:

- BC_RNAseq.txt – RNA-Seq expression count data (18517 genes X 7 samples)

- BC_RNAseq_status.txt – RNA-Seq expression sample status labels (ER+ or ER-)

c) Convert status table into a factor and create DESeq2 Count Data Set

d) Normalize for size effects and estimate dispersion of the data

e) Calculate differential expression

f) Print the most significant 100 genes based on their adjusted p-value using False Discovery Rate control. What is the corresponding FDR for these 100 genes?

g) Extra credit (2pts): how many of the top 500 most significantly differentially expressed genes overlap between the microarray and RNA-Seq analyses? Comment on the degree of overlap—is this what you expected? **Note**: you will need to clean the RNAseq-derived gene names before matching them to the microarray-derived gene names.

Submit to Canvas:

When you’re finished with the lab, make a report of any questions you answered plus any requested output, and gather the scripts that you modified. Submit your homework files using the online Canvas submit tool.

References:

“Gene expression profiling predicts clinical outcome of breast cancer.”
Nature 415, 530-536 (31 January 2002).
van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH.

“Integrated Analysis of Gene Expression, CpG Island Methylation, and Gene Copy Number in Breast Cancer Cells by Deep Sequencing”. PloS One 6(2): e17490 doi:10.1371/journal.pone.0017490 (2011)
Zhifu Sun, Yan W. Asmann, Krishna R. Kalari, Brian Bot, Jeanette E. Eckel-Passow, Tiffany R. Baker, Jennifer M. Carr, Irina Khrebtukova, Shujun Luo, Lu Zhang, Gary P. Schroth,
Edith A. Perez, E. Aubrey Thompson

“DNA methylation profiling reveals a predominant immune component in breast cancers.”EMBO Mol Med. 2011 Dec;3(12):726-41. doi: 10.1002/emmm.201100801. Epub 2011 Nov 16.
Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK, Haibe-Kains B, Defrance M, Michiels S, Volkmar M, Deplus R, Luciani J, Lallemand F, Larsimont D, Toussaint J, Haussy S, Rothé F, Rouas G, Metzger O, Majjaj S, Saini K, Putmans P, Hames G, van Baren N, Coulie PG, Piccart M, Sotiriou C, Fuks F.