MBB 291

Problem Set 6

Due Date: 25 November 2017, 11:59pm

General Instructions

- The answer sheet for this problem set should be submitted as a PDF file. Use $\langle last\ name \rangle$, $\langle first\ name \rangle$*.pdf* as the name of the file. For example, *De la Cruz, Juan.pdf*.
- If you have consulted references (books, journal articles, online materials, other people), cite them as footnotes to the specific item where you used the resource/s as reference.
- Submission of the problem set answers should be done via e-mail. Attach the PDF file, and write as the subject header of the e-mail: [MBB 291] < *Student Number* > − < *Last Name, First Name* > − Problem Set 6. For example, [MBB 291] 190800001 - De la Cruz, Juan - Problem Set 6. Send your answers to janmichaelyap@gmail.com.
- **You should receive a confirmation e-mail from me stating receipt of your deliverable within 24 hours upon your submission of the problem set**. If you have not received any, forward your previous submission using the same subject header once more.
- If you have any questions regarding an item (EXCEPT the answer and solution) in the problem set, do not hesitate to e-mail me to ask them. However, **questions regarding this problem set forwarded/received on or after 12:01am of 22 November 2017 will NOT be entertained**.

Questions

For this problem set, you will also need to download the *20171111 Problem 6 Dataset.zip* file, the link of which is provided in the website. The .zip file contains two folders: *Microarray* and *RNA-seq*. *Microarray* contains 6 CEL files which represents a subset of the GEO dataset with accession number GSE13735[1] containing expression data obtained from the roots of controlled and salt-stressed samples of FL478 rice cultivars. The *RNA-seq* folder meanwhile contains count and sample information data for a Drosophila dataset[2].

In addition, you will need to have all of the R packages that we used during the hands-on exercises.

1. *(NOTE: Prior to performing the analysis, download the ricecdf from the Bioconductor package first)*
   Perform data preprocessing and differential expression test on the microarray data in the dataset. For this part you need to normalize the dataset by subjecting it to MAS5 background correction, applying logarithmic (base 2) transformation, and performing z-score normalization prior to performing the actual differential expression test.
   (a) Using a *two-tailed t-test* ($\alpha = 0.01$), determine how many significantly differentially expressed genes are there in the dataset.
   (b) A gene is *down-regulated* if its (mean) expression value is lower when subjected to a condition than when it is in a controlled environment. To get this using a basic differential expression test, perform a *left-tailed t-test* ($\alpha = 0.01$). How many genes are down-regulated?

[1] Walia H, Wilson C, Ismail AM, Close TJ *et al.* Comparing genomic expression patterns across plant species reveals highly diverged transcriptional dynamics in response to salt stress. BMC Genomics 2009 Aug 25;10:398.

[2] Brooks, A.N., Yang, L., Duff, M.O., Hansen, K.D., Park, J.W., Dudoit, S., Brenner, S.E. and Graveley, B.R. (2011) Conservation of an rna regulatory map between drosophila and mammals. Genome Research, 21(2), 193-202.

(c) A gene is *up-regulated* if its (mean) expression value is higher when subjected to a condition than when it is in a controlled environment.To get this using a basic differential expression test, perform a *right-tailed t-test* ($\alpha = 0.01$). How many genes are up-regulated?

2. Perform a basic, automated weighted gene coexpression network analysis on the set of significantly differentially expressed genes in the microarray data using the pertinent functions in the *WGCNA* package. Choose the appropriate power value based on the power vs. R^2 plot for building the network. Apart from the chosen power value, use the following settings when running *blockwiseModules*:

   - TOMType = "unsigned"
   - minModuleSize = 30
   - reassignThreshold = 0
   - mergeCutHeight = 0.25
   - numericLabels = TRUE
   - pamRespectsDendro = FALSE

   (a) What value did you choose for power estimate?

   (b) How many clusters did the analysis churn out?

   (c) How many genes do each cluster contain?

3. Perform data preprocessing and basic differential expression test on the RNA-seq data in the dataset, according to the instructions set in the following link: http://combine-australia.github.io/RNAseq-R/09-applying-rnaseq-solutions.html (up to *Test for differential expression* section only)

   (a) Show the R statements executed to perform the following:

      i. Reading the count data into R
      ii. Reading the sample information data into R
      iii. Filtering out lowly expressed genes
      iv. Check library sizes
      v. Check boxplots of log2 cpm
      vi. Check MDSplots
      vii. Hierarchical clustering (using heatmap.2)
      viii. Normalzation
      ix. voom transformation of the data
      x. Fitting linear model

   (b) How many *down-regulated* genes are there if only the *untreated and treated* groups were to be compared?

   (c) How many *up-regulated* genes are there if only the *untreated and treated* groups were to be compared?