Course in Practical Analysis of Microarray Data

Introduction to R

Computational Exercises

January 2003 Wolfgang Huber

1.) Installing R. Check whether R is installed on your computer. If not, download it from cran.r-project.org and install it.

2.) Reading data files. In the folder `data/alizadeh`, you find a file `lc7b048rex.DAT`.

- **a.** Open it in a text editor.
- **b.** Read it into a data frame (use the function `read.delim`)
- **c.** Look at the contents of the table (use the functions `dim`, `colnames`, and subsetting)
- **d.** *Optional:* Have a look at the `importWizard` in the package `tkWidgets`.

3.) Simple plots.

- **a.** Make a histogram of the values in the column CH1I.
- **b.** Produce scatterplots of CHI1 versus CH2I, once with linear axis scaling, once with double-logarithmic.
- **c.** Find out how to decorate the plot with your own axis labels and plot title, and how to change the plot symbols.
- **d.** Save the plots as PDF, and as Windows metafiles. Copy and paste them into MS-Office applications.

4.) Calibration and variance stabilization.

- **a.** Subtract the background intensities CH1B, CH2B from the foreground intensities CH1I, CH2I.
- **b.** Use the function `vsn` from the package with the same name to calibrate and transform the data. Plot the result.
- **c.** You may have a look at the "vignette" - try the command `openVignette("vsn")` from the package `Biobase`.

5.) Reading multiple data files. In the folder `data/alizadeh`, you find a file `samples.txt`.

- **a.** Read it into a data frame (use the function `read.delim` with the `as.is=T` option)
- **b.** Create 4 matrices of dimensions 9216×8 that contain, respectively, CH1I, CH1B, CH2I, and CH2B intensities of the 9216 spots on the 8 slides.
- **c.** Save the matrices into an XDR file.
- **d.** Note: the bioconductor packages marrayInput and affy offer more comfortable methods for reading and managing data from a series of microarrays.

6.) Different normalization methods. In the following, we are going to identify genes that appear to be differentially transcribed between the 4 CLL samples and the 4 DLCL samples. For this, we will apply a number of different normalization strategies to the data and compare their results.

- **a.** Download the packages Biobase, marrayClasses, marrayNorm, and multtest from http://www.bioconductor.org and install them.
- **b.** Create a 3D array of dimensions $9216 \times 8 \times 3$ that contains, for all spots, the value of M (that is, the log-ratio or the generalized log-ratio), for the 8 slides and the following 3 different normalization methods:
  1. vsn (affine normalization and variance stabilization)
  2. maNorm with global median location normalization
  3. maNorm with loess for intensity- or A-dependent location normalization using the 'loess' smoother
- **c.** Save the array into an XDR file.

7.) Qualitatively compare the results. Look at scatterplots of the values of M from the same slide, calculated with different normalization methods. Do the values generally agree? How do they differ?

8.) Testing for differential transcription. Now we are ready to calculate test statistics and to select genes. *Note:* The number of replicates (4 versus 4) that we are considering here is very small and no significant conclusions about individual genes or individual samples will be derived from that. The full data set contains many more chips. Here we restrict ourselves to a few of them in order to keep things simple for the purpose of this course.

- **a.** Look at the function `t.test` from the package ctest (which is part of the base libraries), and at `mt.teststat` from the package multtest.
- **b.** For each gene, and for each of the normalization methods, calculate the t-statistic for the CLL-to-DLBL class distinction. Store the result in a 9216 x 3 matrix. Which of the functions `t.test`, `mt.teststat` calculates faster? Look at the histogram of t-values that they produce; you may find extreme values like '3e38' in there. Where do they come from?
- **c.** How do the t-statistics agree between the different normalization methods?

9.) t-thresholds. Designate as *differentially expressed* those clones for which the absolute value of t is larger than a certain threshold. What are the values of this threshold for our data, if we want to have a clone list length $10, 20, 50, 100, \ldots$?

10.) Permutations.

- **a.** How many ways are there to split a set of 8 objects into two groups of 4 and 4 ? Use the function `nchoosek` from the file `nchoosek.R` to generate a numerical representation of these splits.
- **b.** Prepare a matrix with 8 rows, corresponding to the 8 samples, and as many columns as there are splits. Set the matrix elements to 0 and 1, such that each column of the matrix represents a split.

11.) False discovery rate (FDR). Now we want to apply these permutations to the data to estimate the FDR. Do the following for each normalization method:

- **a.** For each of the splits, calculate the corresponding t-statistics for all genes.
- **b.** For each of the splits, and for each of the above choices for clone list lengths and thresholds, calculate the number of clones that have an absolute t-value greater or equal to the threshold.
- **c.** Calculate the median of these numbers across the splits. Divide this by the clone list length to obtain an estimate of the FDR.

12.) In the directory ../data/Shipp, you find a number of Affymetrix CEL files.

- **a.** Using the package `affy`, load them into an `AffyBatch`.
- **b.** Look at the spatial distribution of intensities on the chips.
- **c.** Normalize the data and calculate probe set summary values.
- **d.** Have a look at the vignettes (use `openVignette("affy")`).