Bioconductor’s DEDS package

Yuanyuan Xiao\textsuperscript{1} and Yee Hwa Yang\textsuperscript{2}

March 20, 2008

Departments of \textsuperscript{1}Biopharmaceutical Sciences and \textsuperscript{2}Medicine
University of California, San Francisco
yxiao@itsa.ucsf.edu
jean@biostat.ucsf.edu

Contents

1 Overview 1

2 Case study 1: Affymetrix Spike-in Experiment 2
  2.1 Data ......................................................... 2
  2.2 The \texttt{deds.stat.linkC and \texttt{deds.stat functions}} .................................. 3
  2.3 The plotting functions \texttt{pairs.DEDS and \texttt{hist.deds}} ..................................... 4

3 Case study 2: ApoA1 Experiment 5
  3.1 Data .......................................................... 5
  3.2 Application of the \texttt{deds.stat.linkC function} ............................................. 5

4 Statistical functions 6
  4.1 The \texttt{comp.t and other related functions} .................................................. 6
  4.2 The \texttt{comp.stat function} ................................................................. 6

1 Overview

This document provides a tutorial for the \texttt{DEDS} package for assessment of differential expression (DE) in microarray data.

Introduction to DEDS. There are numerous statistics in the microarray literature that rank genes in evidence of DE, to name a few, fold change (FC), \textit{t} statistic, SAM (Tusher et al. (2001)). selecting a best statistic or ordering statistics in terms of merit has been problematic. No characterizations of microarray data that indicate desirability of a specific choice exist and, likewise, no comparisons across a sufficiently wide range of benchmark datasets have been undertaken. To avoid making fairly arbitrary choices when deciding which ranking statistic to use and to borrow strength across related measures, we apply a novel ranking scheme that assesses DE via distance synthesis
(DEDS) of different related measures. Further details on the packages are given in Yang et al. (2004).

**Functionalities in DEDS.** The DEDS package implements the DEDS procedure and several common statistics, such as, $FC$, $t$ statistics, SAM, $F$ statistics, B statistics (Lönnstedt and Speed (2001)) and moderated $F$ and $t$ statistics (Smyth et al. (2003)), for the analysis of DE in microarrays.

**Case study.** We demonstrate the functionality of the DEDS package using two microarray experiments: Affymetrix spike-in (Irizarry et al. (2003)) and ApoA1 (Dudoit et al. (2002b)).

**Related packages in Bioconductor.** The Bioconductor packages marrayClasses, marrayInput and marrayNorm provide functions for reading and normalizing spotted microarray data. The package affy provides functions for reading and normalizing Affymetrix microarray data.

**Help files.** As with any R package, detailed information on functions, classes and methods can be obtained in the help files. For instance, to view the help file for the function comp.FC in a browser, use `help.start()` followed by `?comp.FC`.

## 2 Case study 1: Affymetrix Spike-in Experiment

We demonstrate the functionality of this package using gene expression data from the Affymetrix spike-in experiment. To load the dataset, use `data(affySpikeIn)`, and to view a description of the experiments and data, type `?affySpikeIn`.

```r
> library(DEDS)
> data(affySpikeIn)
```

### 2.1 Data

The spike-in experiment represents a portion of the data used by Affymetrix to develop their MAS 5.0 preprocessing algorithm. The whole dataset features 14 human genes spiked-in at a series of 14 known concentrations ($0,2^{-2},2^{-1},\ldots,2^{10}$ pM) according to a Latin square design among 12612 null genes. Each “row” of the Latin square (given spike-in gene at a given concentration) was replicated (typically 3 times, two rows 12 times, 59 arrays in total for the whole dataset). Further details are available at [http://www.affymetrix.com/analysis/download_center2.affx](http://www.affymetrix.com/analysis/download_center2.affx) Here we showcase a portion of this dataset that presents a two-group comparison problem with 12 replicates in each group. Therefore, `affySpikeIn` contains the gene expression data for the 24 samples and 12,626 genes retained after RMA probe level summaries. The dataset includes

- `affySpikeIn`: a 12,626 × 24 matrix of expression levels;
- `affySpikeIn.gnames`: a vector of gene identifiers of length 12,626;
- `affySpikeIn.L`: a vector of class labels (0 for class 1, 1 for class 2).
- `spikegene`: a vector that shows that location and identities of the 14 spiked genes.
> dim(affySpikeIn)
[1] 12626  24
> affySpikeIn$L
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1
> spikedgene

37777_at  684_at  1597_at  38734_at  39058_at  36311_at  36889_at  1024_at
  7843  12244    658   8810   9137   6363   6946    28
36202_at  36085_at  40322_at  407_at  1091_at  1708_at
  6253   6134  10414  10795    102    780

2.2 The ded.s.stat.linkC and ded.s.stat functions

The ded.s.stat.linkC and ded.s.stat functions are the main functions that carry out the DEDS procedure. The former wraps around a C function and is therefore quicker than the latter; the latter does the computation solely in R and is slower but is more flexible in fine-tuning parameters for statistical measures. The user is recommended to use ded.s.stat.linkC for efficiency purpose. We describe the most important arguments in the function ded.s.stat.linkC below (see also ?ded.s.stat.linkC):

- **X**: A matrix, in the case of gene expression data, rows correspond to N genes and columns to p mRNA samples.

- **L**: A vector of integers corresponding to observation (column) class labels. For k classes, the labels must be integers between 0 and k − 1.

- **B**: The number of permutations.

- **tests**: A character vector specifying the statistics for synthesis of DEDS. test could be any of the following: t (t statistics), F (F statistics), FC (FC), SAM, moderated t statistics, moderated F statistics and B (B statistics). As a default, DEDS synthesizes t statistics, FC and SAM.

- **tail**: A character string specifying the type of rejection region; choices include abs, higher and lower.

- **adj**: A character string specifying the type of multiple testing adjustment; choices include fdr for returning q values controlling False Discovery Rate (FDR; see Benjamini and Hochberg (1995)) and adjp for adjusted p values (see Dudoit et al. (2002a)) controlling family wise type I error rate.

- **nsig**: If adj = fdr, nsig specifies the number of top differentially expressed genes whose q values will be calculated; we recommend setting nsig < N, as the computation of q values will be extensive. q values for the rest of genes will be approximated to 1. If adj = adjp, the calculation of the adjusted p values will be for the whole dataset.
We apply \texttt{deds.stat.linkC} on the \texttt{affySpikeIn} dataset using 400 permutations and evaluating the \textit{q} values for the top 100 genes. Here, as a default, DEDS synthesizes \textit{t} statistics, FC and SAM. The information of the top 20 genes can be printed out using the function \texttt{topgenes}; note that the rankings of genes by DEDS balance among the three measures it synthesizes, \textit{t} statistics, FC and SAM.

\begin{verbatim}
> deds.affy <- deds.stat.linkC(affySpikeIn, affySpikeIn.L, B = 400, + nsig = 100)
> topgenes(deds.affy, number = 20, genelist = affySpikeIn.gnames)
\end{verbatim}

<table>
<thead>
<tr>
<th>Name</th>
<th>geneOrder</th>
<th>DEDS</th>
<th>t</th>
<th>fc</th>
<th>sam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>684_at</td>
<td>12244</td>
<td>0.00000000</td>
<td>171.858505</td>
<td>7.2235484 82.349174</td>
</tr>
<tr>
<td>2</td>
<td>36085_at</td>
<td>6134</td>
<td>0.00000000</td>
<td>-19.250088</td>
<td>-0.7648563 -8.954151</td>
</tr>
<tr>
<td>3</td>
<td>36202_at</td>
<td>6253</td>
<td>0.00000000</td>
<td>-15.757862</td>
<td>-0.8603835 -8.579226</td>
</tr>
<tr>
<td>4</td>
<td>33818_at</td>
<td>3845</td>
<td>0.00000000</td>
<td>-16.560453</td>
<td>-0.6846256 -7.866766</td>
</tr>
<tr>
<td>5</td>
<td>1091_at</td>
<td>102</td>
<td>0.00000000</td>
<td>-19.858982</td>
<td>-0.4745293 -6.819760</td>
</tr>
<tr>
<td>6</td>
<td>36311_at</td>
<td>6363</td>
<td>0.00000000</td>
<td>-14.013592</td>
<td>-0.6919303 -7.278708</td>
</tr>
<tr>
<td>7</td>
<td>546_at</td>
<td>12106</td>
<td>0.00000000</td>
<td>-10.181562</td>
<td>-0.8456221 -6.568405</td>
</tr>
<tr>
<td>8</td>
<td>1024_at</td>
<td>28</td>
<td>0.00000000</td>
<td>-11.241181</td>
<td>-0.6650023 -6.342763</td>
</tr>
<tr>
<td>9</td>
<td>40322_at</td>
<td>10414</td>
<td>0.00000000</td>
<td>-12.258781</td>
<td>-0.5484858 -6.065384</td>
</tr>
<tr>
<td>10</td>
<td>38734_at</td>
<td>8810</td>
<td>0.00000000</td>
<td>-10.129580</td>
<td>-0.5154920 -5.337663</td>
</tr>
<tr>
<td>11</td>
<td>32660_at</td>
<td>2675</td>
<td>0.00000000</td>
<td>-13.974358</td>
<td>-0.3281164 -4.743865</td>
</tr>
<tr>
<td>12</td>
<td>1552_i_at</td>
<td>609</td>
<td>0.00000000</td>
<td>-6.499473</td>
<td>-0.6759853 -4.515819</td>
</tr>
<tr>
<td>13</td>
<td>36889_at</td>
<td>6946</td>
<td>0.00000000</td>
<td>-8.767055</td>
<td>-0.5016880 -4.874979</td>
</tr>
<tr>
<td>14</td>
<td>38254_at</td>
<td>8325</td>
<td>0.03571429</td>
<td>8.293945</td>
<td>0.3783841 4.144027</td>
</tr>
<tr>
<td>15</td>
<td>37777_at</td>
<td>7843</td>
<td>0.03571429</td>
<td>-8.696619</td>
<td>-0.2633858 -3.466853</td>
</tr>
<tr>
<td>16</td>
<td>39058_at</td>
<td>9137</td>
<td>0.03571429</td>
<td>-7.176071</td>
<td>-0.2977104 -3.415163</td>
</tr>
<tr>
<td>17</td>
<td>1032_at</td>
<td>37</td>
<td>0.03571429</td>
<td>4.684491</td>
<td>0.4109406 3.080278</td>
</tr>
<tr>
<td>18</td>
<td>AFFX-YEL021w/URA3_at</td>
<td>12625</td>
<td>0.03571429</td>
<td>-2.130391</td>
<td>-0.6669140 -1.859075</td>
</tr>
<tr>
<td>19</td>
<td>1708_at</td>
<td>780</td>
<td>0.03571429</td>
<td>-5.415472</td>
<td>-0.3216362 -3.060909</td>
</tr>
<tr>
<td>20</td>
<td>407_at</td>
<td>10795</td>
<td>0.03571429</td>
<td>-6.421535</td>
<td>-0.2048489 -2.637487</td>
</tr>
</tbody>
</table>

2.3 The plotting functions \texttt{pairs.DEDS} and \texttt{hist.deds}

We next illustrate the usage of the function \texttt{pairs.DEDS}, which is a S3 method for \texttt{pairs}. It displays a scatter matrix plot for individual statistics that DEDS synthesizes and highlights the top genes according to a user specified threshold (\texttt{thresh}); see also \texttt{?pairs.DEDS}. Plots on the diagonal panels are QQ-plots as the default, but can be set as \texttt{histogram}, \texttt{boxplot}, \texttt{density} or \texttt{none}. To display only qq-plots, the function \texttt{qqnorm.DEDS} can be use.

\begin{verbatim}
> pairs(deds.affy, subset = c(2:12626), thresh = 0.01, legend = F)
> qqnorm(deds.affy, subset = c(2:12626), thresh = 0.01)
\end{verbatim}
3 Case study 2: ApoA1 Experiment

The next example we demonstrate is a set of cDNA microarray data from a study of a mouse model with very low HDL cholesterol levels described in [Dudoit et al. (2002b)]. To load the dataset, use `data(ApoA1)`, and to view a description of the experiments and data, type `?ApoA1`.

> `data(ApoA1)`

3.1 Data

The goal of the ApoA1 experiment is to identify DE genes in apolipoprotein A1 (apo A1) knock-out mice. The treatment group consists of eight knock-out mice and the control group consists of eight normal mice. The dataset includes

- ApoA1: a 6,384 × 16 matrix of expression levels;
- ApoA1.L: a vector of class labels (0 for the control group, 1 for the treatment group);

> `dim(ApoA1)`

[1] 6384 16

> `ApoA1.L`

[1] 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1

3.2 Application of the `deds.stat.linkC` function

We apply `deds.stat.linkC` on the `ApoA1` dataset using 400 permutations and evaluating the adjusted p. Here, as a default, DEDS synthesizes t statistics, FC and SAM.


> `sum(deds.ApoA1$p <= 0.01)`

[1] 8

> `sum(deds.ApoA1$p <= 0.05)`

[1] 11

> `topgenes(deds.ApoA1, number = 9)`

<table>
<thead>
<tr>
<th>geneOrder</th>
<th>DEDS</th>
<th>t</th>
<th>fc</th>
<th>sam</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2149</td>
<td>0.0000</td>
<td>16.500626</td>
<td>3.2440281</td>
</tr>
<tr>
<td>2</td>
<td>540</td>
<td>0.0000</td>
<td>8.784189</td>
<td>2.9692008</td>
</tr>
<tr>
<td>3</td>
<td>5356</td>
<td>0.0000</td>
<td>9.256464</td>
<td>1.7848039</td>
</tr>
<tr>
<td>4</td>
<td>1739</td>
<td>0.0000</td>
<td>9.789542</td>
<td>0.9976147</td>
</tr>
<tr>
<td>5</td>
<td>2537</td>
<td>0.0000</td>
<td>7.842692</td>
<td>0.9840616</td>
</tr>
<tr>
<td>6</td>
<td>4139</td>
<td>0.0000</td>
<td>7.880206</td>
<td>0.9779046</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0.0025</td>
<td>-6.656223</td>
<td>-0.8907498</td>
</tr>
<tr>
<td>8</td>
<td>4941</td>
<td>0.0100</td>
<td>5.909812</td>
<td>0.9214721</td>
</tr>
<tr>
<td>9</td>
<td>1204</td>
<td>0.0200</td>
<td>-5.156277</td>
<td>-0.9961325</td>
</tr>
</tbody>
</table>
> pairs(deds.ApoA1, legend = F)

4 Statistical functions

4.1 The comp.t and other related functions

The DEDS package provides the following functions, comp.FC, comp.t, comp.SAM, comp.F, comp.B, comp.modt and comp.modF for the computation of FC, t-statistics, SAM, F statistics, B statistics, moderated t- and F- statistics respectively.

There are two steps in applying the above functions to obtain corresponding statistics:

1. Create the statistic function.
2. Apply the function to the microarray expression matrix.

We illustrate the usage with comp.t and other functions follow the same rules. The function comp.t has three arguments (see also ?comp.t):

- L: A vector of integers corresponding to observation (column) class labels. For \( k \) classes, the labels must be integers between 0 and \( k - 1 \).
- mu: A number indicating the true value of the mean (or difference in means if two sample statistics are calculated; default set at 0).
- var.equal: a logical variable indicating whether to treat the two variances as being equal.

comp.t returns a function of one argument with bindings for L, mu and var.equal. This function accepts a microarray data matrix as its single argument, when evaluated, computes \( t \) statistic for each row of the matrix.

> t <- comp.t(L = affySpikeIn.L)
> t.affy <- t(affySpikeIn)

4.2 The comp.stat function

A simple wrapper function comp.stat is provided for users interested in applying a standard set of statistical measures using default parameters. The most important arguments for comp.stat are elaborated below:

- X: A matrix, with rows correspond to genes and columns to mRNA samples.
- L: A vector of integers corresponding to observation (column) class labels. For \( k \) classes, the labels must be integers between 0 and \( k - 1 \).
- test: A character string specifying the statistic to be applied.
  - \texttt{t} – \( t \) statistics;
  - \texttt{fc} – FC;
  - \texttt{sam} – SAM statistics;
Figure 1: Pairs plots for the ApoA1 data
To compute $t$ statistics on the `affySpikeIn` data, instead of using `comp.t`, the users can also use `comp.stat` by specifying the `test` as `t`. However, `comp.stat` computes $t$ statistics assuming unequal variance (if it is a two-sample comparison); if the user desires to use an equal variance option, the function `comp.t` has to be applied instead.

```r
> t.affy <- comp.stat(affySpikeIn, affySpikeIn.L, test = "t")
```

References


