PCOT2: Principal Coordinates and Hotelling's T^2 for the analysis of microarray data

Sarah Song and Mik Black

October 18, 2010

1 Overview

pcot2 is an R-package for the analysis of groups of genes in microarray experiments. It utilizes inter-gene correlation information to detect significant alterations in the activities of gene sets. Incorporating additional (usually functional) information into the data analysis process allows gene interactions to be investigated in a statistical framework. One of the reasons that gene set analysis is becoming important is that it is suitable for detecting small coordinated changes in expression of groups of genes which are functionally related, which may not be considered significant in a single gene analysis. This vignette gives a tutorial-style introduction to the functions in the pcot2 package. These functions are used for testing and visualizing changes in expression activity for groups of genes.

2 Example: ALL/AML data

In this example the ALL/AML leukemia data set of Golub *et al.*(1999) is used to illustrate the functionality of the pcot2 package. This data set contains 38 bone marrow samples obtained from adult leukemia patients, 11 relating to acute myeloid leukemia (AML, class 1) and 27 relating to acute lymphoblastic leukemia (ALL, class 0). Gene expression levels were measured using Affymetrix high density oligonucleotide arrays containing 6817 human genes, of which 3051 genes were considered suitable for analysis by Golub et al.(1999) after preprocessing. This data set is available as part of the multtest package and gene sets are defined as KEGG pathways using the hu6800.db annotation package. Both packages can be downloaded from www.bioconductor.org.

```
> library(pcot2)
> library(multtest)
> library(hu6800.db)
> set.seed(1234567)
```

3 The pcot2 function

The pcot2 function implements the PCOT2 testing method, which is a two-stage permutation-based approach for testing changes in activity in pre-specified gene sets. The function requires at least three inputs: gene expression data, sample class labels, and a gene category indicator matrix. The gene expression data should be in the form of a matrix with no missing values. Data preprocessing (e.g. normalization) must therefore take place before running the PCOT2 analysis.

```
> data(golub)
> rownames(golub) <- golub.gnames[, 3]
> colnames(golub) <- golub.cl
```

The class labels represent two distinct experimental conditions (e.g., AML and ALL).

```
> golub.cl
```

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

The gene category indicator matrix is designed to indicate presence or absence of genes in the pre-defined gene categories (e.g., gene pathways). The indicator matrix contains rows representing gene identifiers for genes present in the expression data, and columns representing pre-defined group names. The values 1 or 0 indicate the presence or absence of a gene in a particular group.

In this example, the `hu6800.db` annotation package is used to define the KEGG (http://www.genome.jp/kegg/pathway.html) pathways for all of 3051 genes in the data. The `getImat` function is used to generate an indicator matrix which includes 65 KEGG pathways containing at least 10 of the total 3051 genes.

```
> KEGG.list <- as.list(hu6800PATH)
> imat <- getImat(golub, KEGG.list, ms = 10)
> colnames(imat) <- paste("KEGG", colnames(imat), sep = "")
> dim(imat)
```

```
[1] 3051  142
```

Permutations are used to produce p-values based on the null distribution of the T^2 statistic. By default `pcot2` will automatically run 1000 permutations. In order to minimize the time taken to build this vignette, only 10 permutations have been performed.

```
> results <- pcot2(golub, golub.cl, imat, iter = 10)
```

```
Comparison: 0-1
```

The output from the `pcot2` function can contain information on either all pathways or just significantly differentially expressed pathways, based on the value of α used in the function, where α determines the significance threshold for the permutation p-values. For each KEGG pathway, the number of genes in the pathway is listed, along with Hotelling's T^2 statistic. These are followed by parametric p-values for the test statistic, both raw and adjusted. The last two columns provide raw and adjusted permutation-based p-values. The default adjustment method is the false discovery rate controlling method of Benjamini and Yekutieli (2001).

```
> results$res.sig

[1] Num         T2          P.nor       P.adj       P.permu     P.permu.adj
<0 rows> (or 0-length row.names)

> results$res.all

          Num        T2        P.nor        P.adj P.permu P.permu.adj
KEGG04080  55 53.233934 1.261880e-07 3.100248e-06     0.1   0.5823641
KEGG04360  30 35.193509 6.570324e-06 8.331504e-05     0.1   0.5823641
KEGG04010  97 40.757244 1.760778e-06 2.563535e-05     0.1   0.5823641
KEGG04910  54 21.839940 2.490615e-04 2.083086e-03     0.1   0.5823641
KEGG03410  14 40.040059 2.075157e-06 2.966310e-05     0.1   0.5823641
KEGG04650  59 55.645654 7.912906e-08 2.488424e-06     0.1   0.5823641
KEGG05322  45 70.923249 5.327594e-09 3.150664e-07     0.1   0.5823641
KEGG04962  14 25.229090 9.194586e-05 8.815494e-04     0.1   0.5823641
KEGG04510  79 52.030331 1.600398e-07 3.812785e-06     0.1   0.5823641
KEGG04270  43 23.290321 1.614646e-04 1.426315e-03     0.1   0.5823641
KEGG04810  83 40.998056 1.666856e-06 2.472580e-05     0.1   0.5823641
KEGG04520  33 21.495315 2.765195e-04 2.288392e-03     0.1   0.5823641
KEGG04670  53 34.386677 8.020671e-06 1.000918e-04     0.1   0.5823641
KEGG04060  83 57.649662 5.419198e-08 1.775220e-06     0.1   0.5823641
KEGG04062  87 70.607587 5.610503e-09 3.150664e-07     0.1   0.5823641
KEGG03050  23 26.894107 5.749360e-05 5.870257e-04     0.1   0.5823641
KEGG04110  57 46.327670 5.167040e-07 1.015571e-05     0.1   0.5823641
KEGG03320  18 55.009039 8.939526e-08 2.508401e-06     0.1   0.5823641
KEGG04971  34 19.068265 5.881924e-04 4.578533e-03     0.1   0.5823641
KEGG05110  30 24.810971 1.036597e-04 9.701946e-04     0.1   0.5823641
KEGG04146  20 32.548726 1.274396e-05 1.541415e-04     0.1   0.5823641
KEGG00190  43 14.212036 2.959080e-03 2.022960e-02     0.1   0.5823641
KEGG01100 307 65.043122 1.433592e-08 6.261543e-07     0.1   0.5823641
KEGG05010  67 16.031326 1.587297e-03 1.134472e-02     0.1   0.5823641
KEGG05012  43 10.731022 1.040403e-02 6.704555e-02     0.1   0.5823641
KEGG05016  70 31.545201 1.649643e-05 1.907258e-04     0.1   0.5823641
KEGG04142  53 61.157011 2.847713e-08 1.017658e-06     0.1   0.5823641
KEGG03420  15 15.484007 1.909975e-03 1.352798e-02     0.1   0.5823641
KEGG04141  63 41.998038 1.329932e-06 2.133839e-05     0.1   0.5823641
KEGG04144  52 36.689760 4.565751e-06 5.982591e-05     0.1   0.5823641
KEGG04020  56 32.309882 1.354665e-05 1.613676e-04     0.1   0.5823641
KEGG04666  42 45.439744 6.246788e-07 1.197847e-05     0.1   0.5823641
KEGG05100  30 32.603669 1.256653e-05 1.541415e-04     0.1   0.5823641
KEGG00350  10  5.163053 9.581005e-02 5.458337e-01     0.1   0.5823641
KEGG04514  61 29.696185 2.681273e-05 2.848641e-04     0.1   0.5823641
KEGG04530  36 31.095936 1.854001e-05 2.082285e-04     0.1   0.5823641
KEGG03430  13 22.840756 1.844695e-04 1.593718e-03     0.1   0.5823641
KEGG05200 149 68.145490 8.444726e-09 4.426114e-07     0.1   0.5823641
KEGG05210  36 29.839909 2.580669e-05 2.779315e-04     0.1   0.5823641
KEGG05213  28 26.480816 6.452479e-05 6.341105e-04     0.1   0.5823641
KEGG05416  40 20.558833 3.685836e-04 2.987395e-03     0.1   0.5823641
KEGG04120  29 12.630167 5.181351e-03 3.394611e-02     0.1   0.5823641
KEGG04210  41 25.794077 7.829317e-05 7.599188e-04     0.1   0.5823641
```

KEGG05142	55	49.447678	2.694878e-07	6.053400e-06	0.1	0.5823641
KEGG05014	23	31.606850	1.623516e-05	1.905066e-04	0.1	0.5823641
KEGG05130	23	7.705335	3.356919e-02	1.984347e-01	0.1	0.5823641
KEGG05131	30	55.207896	8.604510e-08	2.505479e-06	0.1	0.5823641
KEGG04115	24	37.099129	4.138379e-06	5.514506e-05	0.1	0.5823641
KEGG04916	31	13.686536	3.557264e-03	2.410940e-02	0.1	0.5823641
KEGG05215	47	53.971118	1.092671e-07	2.850310e-06	0.1	0.5823641
KEGG04310	44	41.315269	1.551142e-06	2.391166e-05	0.1	0.5823641
KEGG04350	23	24.300792	1.201271e-04	1.089933e-03	0.1	0.5823641
KEGG04145	74	143.647979	6.066259e-13	2.384620e-10	0.1	0.5823641
KEGG05410	31	18.282987	7.562727e-04	5.717069e-03	0.1	0.5823641
KEGG05414	32	10.341813	1.204386e-02	7.636115e-02	0.1	0.5823641
KEGG00010	37	9.063638	1.964873e-02	1.188282e-01	0.1	0.5823641
KEGG04620	48	49.019006	2.942818e-07	6.426718e-06	0.1	0.5823641
KEGG04630	54	41.009056	1.662694e-06	2.472580e-05	0.1	0.5823641
KEGG05140	55	42.007433	1.327132e-06	2.133839e-05	0.1	0.5823641
KEGG05212	42	26.528157	6.367518e-05	6.336821e-04	0.1	0.5823641
KEGG04640	61	123.722346	4.746648e-12	1.243925e-09	0.1	0.5823641
KEGG00980	10	66.696592	1.079104e-08	5.302389e-07	0.1	0.5823641
KEGG00983	12	44.930783	6.971132e-07	1.304915e-05	0.1	0.5823641
KEGG00240	30	74.320240	3.081965e-09	2.692239e-07	0.1	0.5823641
KEGG00480	14	89.964548	3.026550e-10	4.758896e-08	0.1	0.5823641
KEGG00590	17	41.335666	1.543998e-06	2.391166e-05	0.1	0.5823641
KEGG00860	15	49.760861	2.527749e-07	5.844985e-06	0.1	0.5823641
KEGG00030	15	13.506746	3.790243e-03	2.525302e-02	0.1	0.5823641
KEGG00230	51	25.179015	9.327181e-05	8.834880e-04	0.1	0.5823641
KEGG00071	18	39.257416	2.487030e-06	3.491574e-05	0.1	0.5823641
KEGG04920	27	62.446658	2.260875e-08	8.772488e-07	0.1	0.5823641
KEGG00620	14	24.286911	1.206120e-04	1.089933e-03	0.1	0.5823641
KEGG04930	21	19.258351	5.537710e-04	4.353700e-03	0.1	0.5823641
KEGG04664	36	62.245608	2.343224e-08	8.772488e-07	0.1	0.5823641
KEGG04722	55	55.236204	8.557909e-08	2.505479e-06	0.1	0.5823641
KEGG04912	34	13.567744	3.709441e-03	2.492591e-02	0.1	0.5823641
KEGG00280	19	38.660972	2.858611e-06	3.874855e-05	0.1	0.5823641
KEGG00310	12	28.018168	4.216839e-05	4.362162e-04	0.1	0.5823641
KEGG00380	15	103.491944	5.077894e-11	9.980492e-09	0.1	0.5823641
KEGG00640	14	47.605074	3.946596e-07	8.165211e-06	0.1	0.5823641
KEGG00650	10	4.237641	1.426346e-01	8.067487e-01	0.1	0.5823641
KEGG00020	14	13.152966	4.297080e-03	2.838931e-02	0.1	0.5823641
KEGG04012	38	23.225345	1.645928e-04	1.437794e-03	0.1	0.5823641
KEGG05220	47	39.126096	2.564215e-06	3.536779e-05	0.1	0.5823641
KEGG00564	14	58.921920	4.279369e-08	1.462784e-06	0.1	0.5823641
KEGG05340	25	148.792814	3.701484e-13	2.384620e-10	0.1	0.5823641
KEGG00500	12	28.113816	4.108093e-05	4.306331e-04	0.1	0.5823641
KEGG05120	34	65.157949	1.405379e-08	6.261543e-07	0.1	0.5823641
KEGG03040	40	17.132641	1.100194e-03	8.083768e-03	0.1	0.5823641
KEGG04660	50	10.494546	1.136995e-02	7.267445e-02	0.1	0.5823641
KEGG00410	12	46.645514	4.830102e-07	9.736885e-06	0.1	0.5823641
KEGG05221	39	35.710984	5.788211e-06	7.460069e-05	0.1	0.5823641
KEGG04340	11	6.073128	6.534459e-02	3.749880e-01	0.1	0.5823641

KEGG05218	31	20.513822	3.737548e-04	2.998397e-03	0.1	0.5823641
KEGG04512	26	24.645916	1.087092e-04	1.005486e-03	0.1	0.5823641
KEGG05146	49	71.444582	4.892881e-09	3.150664e-07	0.1	0.5823641
KEGG05222	46	43.526104	9.470522e-07	1.654588e-05	0.1	0.5823641
KEGG04610	13	71.766981	4.642947e-09	3.150664e-07	0.1	0.5823641
KEGG03030	19	22.769488	1.884236e-04	1.610185e-03	0.1	0.5823641
KEGG04622	20	53.826381	1.123894e-07	2.850310e-06	0.1	0.5823641
KEGG00970	16	23.403392	1.561698e-04	1.395220e-03	0.1	0.5823641
KEGG04970	37	63.587530	1.848021e-08	7.646835e-07	0.1	0.5823641
KEGG04370	35	31.024253	1.889009e-05	2.091722e-04	0.1	0.5823641
KEGG04662	45	44.427951	7.774477e-07	1.421448e-05	0.1	0.5823641
KEGG00051	16	26.636897	6.176816e-05	6.225846e-04	0.1	0.5823641
KEGG00052	15	19.849740	4.596460e-04	3.650200e-03	0.1	0.5823641
KEGG04114	41	21.934126	2.420699e-04	2.046379e-03	0.1	0.5823641
KEGG04540	35	9.106446	1.932494e-02	1.177760e-01	0.1	0.5823641
KEGG04914	33	18.313146	7.489572e-04	5.716736e-03	0.1	0.5823641
KEGG04070	29	20.991795	3.225358e-04	2.641405e-03	0.1	0.5823641
KEGG04720	35	7.609501	3.488383e-02	2.031509e-01	0.1	0.5823641
KEGG04730	31	78.228678	1.675825e-09	1.656859e-07	0.1	0.5823641
KEGG00561	12	88.191090	3.878922e-10	5.082625e-08	0.1	0.5823641
KEGG00330	21	71.283630	5.022943e-09	3.150664e-07	0.1	0.5823641
KEGG00520	15	8.466957	2.481070e-02	1.489005e-01	0.1	0.5823641
KEGG04672	23	43.067585	1.047902e-06	1.790981e-05	0.1	0.5823641
KEGG05144	32	78.189327	1.685960e-09	1.656859e-07	0.1	0.5823641
KEGG05310	20	31.461775	1.685710e-05	1.920711e-04	0.1	0.5823641
KEGG05320	24	14.784894	2.426270e-03	1.694922e-02	0.1	0.5823641
KEGG05330	23	18.244124	7.658108e-04	5.734037e-03	0.1	0.5823641
KEGG04612	39	43.632799	9.250903e-07	1.652950e-05	0.1	0.5823641
KEGG04940	23	9.603295	1.595365e-02	9.876079e-02	0.1	0.5823641
KEGG05332	23	10.257857	1.243218e-02	7.819260e-02	0.1	0.5823641
KEGG05214	38	16.596301	1.313947e-03	9.477189e-03	0.1	0.5823641
KEGG05219	22	48.867088	3.036379e-07	6.451825e-06	0.1	0.5823641
KEGG05223	31	16.965995	1.162369e-03	8.461521e-03	0.1	0.5823641
KEGG04621	22	54.830169	9.252661e-08	2.508401e-06	0.1	0.5823641
KEGG04966	10	42.692106	1.138933e-06	1.905148e-05	0.1	0.5823641
KEGG04623	17	17.496447	9.763539e-04	7.241520e-03	0.1	0.5823641
KEGG04330	16	14.667409	2.526630e-03	1.742469e-02	0.1	0.5823641
KEGG04150	18	11.009560	9.376387e-03	6.092261e-02	0.1	0.5823641
KEGG05216	19	30.751272	2.028858e-05	2.215377e-04	0.1	0.5823641
KEGG05020	21	14.773131	2.436126e-03	1.694922e-02	0.1	0.5823641
KEGG04742	10	9.165107	1.889037e-02	1.160269e-01	0.1	0.5823641
KEGG00562	15	18.867003	6.271148e-04	4.833650e-03	0.1	0.5823641
KEGG00510	15	7.675775	3.396901e-02	1.992996e-01	0.2	1.0000000
KEGG00270	12	8.220476	2.734539e-02	1.628690e-01	0.2	1.0000000
KEGG00250	11	9.616124	1.587530e-02	9.876079e-02	0.2	1.0000000
KEGG04960	19	6.414720	5.672222e-02	3.279009e-01	0.3	1.0000000
KEGG04260	29	2.355025	3.299165e-01	1.000000e+00	0.4	1.0000000
KEGG05412	26	3.301194	2.153740e-01	1.000000e+00	0.5	1.0000000
KEGG05211	31	2.628229	2.913801e-01	1.000000e+00	0.5	1.0000000

In the `pcot2` function, the T^2 statistic can be calculated in two ways, using either a pooled estimate of correlation for the two classes (default) or an unpooled estimate. And users can set *var.equal=F* if the correlation structure is assumed to differ across the two classes.

In the first step of the PCOT2 analysis, the dimensionality of the gene expression data is reduced via principal coordinates. The default dimensionality in the `pcot2` function is set as *ncomp=2*. In the second step of the PCOT2 analysis, the distances between the transformed groups are calculated via euclidean distances by default. Other distances (e.g., correlation or Spearman distances) can also be used by defining *dist.method* in the function. A permutation p-value for each category is calculated by re-arranging the sample labels. The permutations can also be performed by permuting rows (genes), using `permu='ByRow'`.

Table 1 lists computation times (in minutes) required to run 1000 permutations of the `pcot2` function on the AML/ALL data under various parameter configurations. The two machines used were a 3.2GHz Pentium 4 with 1Gb RAM running Microsoft Windows XP and R 2.1.0 (PC), and a 1.70GHz Pentium M with 256Mb of RAM running Fedora Core 3 and R 2.2.0 (Unix).

Table 1: *Computation times (minutes, 1000 permutations)*

Changes	PC machine	UNIX machine
default setting	5.6	6.8
var.equal=F	5.5	6.8
comp=8	6	7.6
dist.method="euclidean"	4.8	6
permu="ByRow"	5.6	6.8

4 The `corplot` and `corplot2` functions

The `corplot` and `corplot2` functions enable visualization of both correlation and gene expression information for a particular gene category, in particular the groups identified as being differentially expressed. The plot produced by the `corplot` function displays the pooled correlation calculated from the two classes, while the `corplot2` function produces a plot based on unpooled correlation. Gene names can be added to the plot using *add.name=T* (default). The font size can be changed by setting the *font.size* argument. The *main* option specifies the title of the plot.

```
> sel <- c("04620", "04120")
> pvalue <- c(0.001, 0.72)
> library(KEGG.db)
> pname <- unlist(mget(sel, env = KEGGPATHID2NAME))
> main <- paste("KEGG", sel, ": ", pname, ": ", "P=", pvalue, sep = "")
> for (i in 1:length(sel)) {
+     fname <- paste("corplot2-KEGG", sel[i], ".jpg", sep = "")
+     jpeg(fname, width = 1600, height = 1200, quality = 100)
+     selgene <- rownames(imat)[imat[, match(paste("KEGG", sel,
+         sep = "")[i], colnames(imat))] == 1]
+     corplot2(golub, selgene, golub.cl, main = main[i])
```

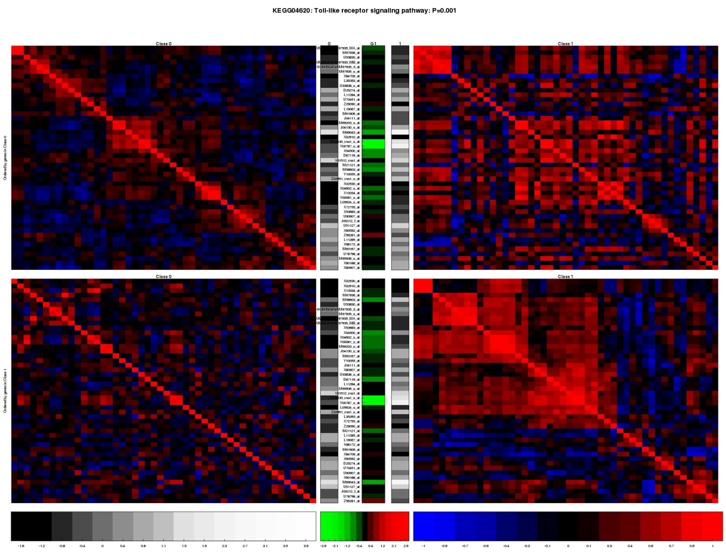

Figure 1: KEGG04620

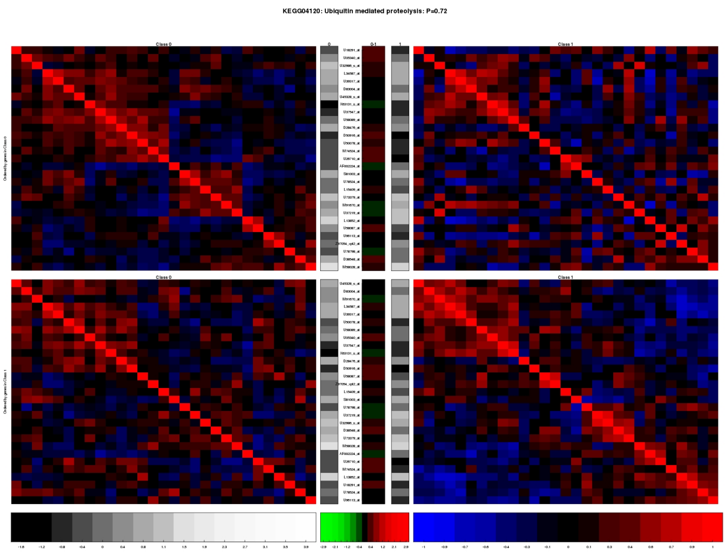

Figure 2: KEGG04120

```
+       dev.off()
+ }
```

The argument *inputP* allows users to input the *p*-values of individual genes calculated using other approaches, such as the limma package (Smyth *et al.*, 2004), allowing the results from both per-gene and per-pathway analysis to be printed on a single plot. To allow users to identify genes from in correlation image plots, the argument *gene.locator=T* allows the selection of interesting (e.g., highly correlated and differential expressed between two classes) genes by clicking beginning and end points on the main diagonal of the image plots. This prints the identifiers for the selected genes. Further details of this functionality are provided in the HowToUseGeneLocator.pdf document. The usage of `corplot2` is similar to that for the `corplot` function.

5 The `aveProbes` function

In Affymetrix gene expression data, a unique gene can often link to multiple probe sets, with such genes then having a greater influence on the pathway analysis (particularly if the gene is differentially expressed). In order to solve this problem, the `aveProbe` function is provided to change the multiple probe data to the unique gene data by taking the median of the probe values. This function can be used to transform both expression data and the indicator matrix by providing a vector of unique gene identifiers.

```
> pathlist <- as.list(hu6800PATH)
> pathlist <- pathlist[match(rownames(golub), names(pathlist))]
> ids <- unlist(mget(names(pathlist), env = hu6800SYMBOL))
> newdata <- aveProbe(x = golub, ids = ids)$newx
> output <- aveProbe(x = golub, imat = imat, ids = ids)
> newdata <- output$newx
> newimat <- output$newimat
> newimat <- newimat[, apply(newimat, 2, sum) >= 10]
> dim(newdata)

[1] 2557   38

> dim(newimat)

[1] 2557  136
```

After the multiple probe data set has been changed to the unique gene symbol data, further analysis such as testing and visualizing pathways can be done on the new data set.

References

[1] Benjamini,B.Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165-1188.

[2] Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

[3] Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, **286**, 531-537.

[4] Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, No.1, Article 3.