

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

Sarah Song and Mik Black

April 16, 2015

## 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 pre-processing. 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](http://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

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

```
> golub.cl
```

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.

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

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.

Comparison: 0-1

2

```
> 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	51.452125	1.796025e-07	4.295669e-06	0.1	0.5860615
KEGG04360	30	35.193509	6.570324e-06	8.426706e-05	0.1	0.5860615
KEGG04010	96	40.777404	1.752705e-06	2.585103e-05	0.1	0.5860615
KEGG04910	53	21.777816	2.537898e-04	2.223683e-03	0.1	0.5860615
KEGG03410	14	40.040059	2.075157e-06	3.010518e-05	0.1	0.5860615
KEGG04650	58	53.920463	1.103493e-07	3.013920e-06	0.1	0.5860615
KEGG04962	14	25.229090	9.194586e-05	9.246336e-04	0.1	0.5860615
KEGG04510	79	52.030331	1.600398e-07	3.934102e-06	0.1	0.5860615
KEGG04270	42	24.825175	1.032369e-04	1.019266e-03	0.1	0.5860615
KEGG04810	84	42.380783	1.220748e-06	1.964189e-05	0.1	0.5860615
KEGG04060	83	57.649662	5.419198e-08	1.918294e-06	0.1	0.5860615
KEGG04062	87	70.607587	5.610503e-09	3.489577e-07	0.1	0.5860615
KEGG03050	22	20.987918	3.229199e-04	2.747778e-03	0.1	0.5860615
KEGG04110	57	46.327670	5.167040e-07	9.728909e-06	0.1	0.5860615
KEGG04971	33	19.064533	5.888905e-04	4.653039e-03	0.1	0.5860615
KEGG04972	33	44.132155	8.292288e-07	1.439584e-05	0.1	0.5860615
KEGG04976	20	19.403696	5.288905e-04	4.254938e-03	0.1	0.5860615
KEGG05110	30	24.810971	1.036597e-04	1.019266e-03	0.1	0.5860615
KEGG04146	18	37.574387	3.694684e-06	4.880032e-05	0.1	0.5860615
KEGG00190	42	14.032358	3.150661e-03	2.212847e-02	0.1	0.5860615
KEGG01100	313	69.895011	6.309175e-09	3.489577e-07	0.1	0.5860615
KEGG05010	65	17.300928	1.040963e-03	7.806808e-03	0.1	0.5860615
KEGG05012	41	11.792235	7.023285e-03	4.638266e-02	0.1	0.5860615
KEGG05016	67	17.363594	1.019776e-03	7.713283e-03	0.1	0.5860615
KEGG04142	52	61.237256	2.806865e-08	1.079975e-06	0.1	0.5860615
KEGG03420	15	15.484007	1.909975e-03	1.374177e-02	0.1	0.5860615
KEGG04141	61	43.161502	1.026357e-06	1.713731e-05	0.1	0.5860615
KEGG03018	13	5.431912	8.549321e-02	4.977465e-01	0.1	0.5860615
KEGG04144	52	37.787256	3.512529e-06	4.709732e-05	0.1	0.5860615
KEGG04020	55	32.084611	1.435281e-05	1.716427e-04	0.1	0.5860615
KEGG04666	43	46.505825	4.975229e-07	9.571398e-06	0.1	0.5860615
KEGG05100	31	33.781340	9.329447e-06	1.162834e-04	0.1	0.5860615
KEGG00350	10	5.163053	9.581005e-02	5.541659e-01	0.1	0.5860615
KEGG04514	62	30.108931	2.402932e-05	2.691749e-04	0.1	0.5860615
KEGG04530	36	31.095936	1.854001e-05	2.158821e-04	0.1	0.5860615
KEGG04670	52	33.856570	9.155178e-06	1.157415e-04	0.1	0.5860615
KEGG05160	55	53.505956	1.196423e-07	3.096946e-06	0.1	0.5860615
KEGG03430	13	22.840756	1.844695e-04	1.665783e-03	0.1	0.5860615
KEGG05200	149	68.034766	8.603400e-09	4.478590e-07	0.1	0.5860615
KEGG05210	37	29.737191	2.652154e-05	2.897570e-04	0.1	0.5860615
KEGG05213	28	26.480816	6.452479e-05	6.639698e-04	0.1	0.5860615
KEGG05416	41	21.371049	2.871904e-04	2.467475e-03	0.1	0.5860615
KEGG04120	29	13.923334	3.273264e-03	2.280854e-02	0.1	0.5860615

KEGG04974	27	6.653981	5.140452e-02	3.053059e-01	0.1	0.5860615
KEGG04210	41	25.794077	7.829317e-05	7.963881e-04	0.1	0.5860615
KEGG05142	55	49.447678	2.694878e-07	6.114975e-06	0.1	0.5860615
KEGG05014	23	31.606850	1.623516e-05	1.915647e-04	0.1	0.5860615
KEGG05130	24	8.239661	2.713860e-02	1.644958e-01	0.1	0.5860615
KEGG05131	31	55.243538	8.545878e-08	2.551955e-06	0.1	0.5860615
KEGG04115	24	37.099129	4.138379e-06	5.385692e-05	0.1	0.5860615
KEGG04916	30	13.646748	3.607485e-03	2.494105e-02	0.1	0.5860615
KEGG05215	46	53.385953	1.224846e-07	3.096946e-06	0.1	0.5860615
KEGG04310	44	41.315269	1.551142e-06	2.366702e-05	0.1	0.5860615
KEGG04350	23	24.300792	1.201271e-04	1.147699e-03	0.1	0.5860615
KEGG03013	38	13.379371	3.965031e-03	2.699127e-02	0.1	0.5860615
KEGG04145	75	146.861808	4.447553e-13	2.010185e-10	0.1	0.5860615
KEGG04520	32	22.207678	2.229148e-04	1.972691e-03	0.1	0.5860615
KEGG05410	31	18.282987	7.562727e-04	5.819702e-03	0.1	0.5860615
KEGG05414	32	10.341813	1.204386e-02	7.779746e-02	0.1	0.5860615
KEGG00010	37	9.063638	1.964873e-02	1.207514e-01	0.1	0.5860615
KEGG04380	71	32.680575	1.232254e-05	1.514565e-04	0.1	0.5860615
KEGG04620	48	49.019006	2.942818e-07	6.510638e-06	0.1	0.5860615
KEGG04630	54	41.009056	1.662694e-06	2.493908e-05	0.1	0.5860615
KEGG05140	56	41.989367	1.332522e-06	2.105749e-05	0.1	0.5860615
KEGG05145	69	47.765451	3.816431e-07	8.041338e-06	0.1	0.5860615
KEGG05212	42	26.528157	6.367518e-05	6.629358e-04	0.1	0.5860615
KEGG04640	63	123.041575	5.115131e-12	1.508883e-09	0.1	0.5860615
KEGG00980	10	66.696592	1.079104e-08	5.305311e-07	0.1	0.5860615
KEGG00983	11	44.129914	8.296347e-07	1.439584e-05	0.1	0.5860615
KEGG00240	30	74.320240	3.081965e-09	2.727394e-07	0.1	0.5860615
KEGG00480	14	89.964548	3.026550e-10	5.356709e-08	0.1	0.5860615
KEGG00590	16	39.391204	2.410915e-06	3.441203e-05	0.1	0.5860615
KEGG00860	15	49.760861	2.527749e-07	5.886681e-06	0.1	0.5860615
KEGG00030	15	13.506746	3.790243e-03	2.600144e-02	0.1	0.5860615
KEGG00230	50	24.125069	1.264215e-04	1.190182e-03	0.1	0.5860615
KEGG00071	18	39.257416	2.487030e-06	3.493499e-05	0.1	0.5860615
KEGG03320	18	55.009039	8.939526e-08	2.551955e-06	0.1	0.5860615
KEGG04920	27	62.446658	2.260875e-08	9.425648e-07	0.1	0.5860615
KEGG05150	32	56.063892	7.306855e-08	2.412054e-06	0.1	0.5860615
KEGG00620	14	24.286911	1.206120e-04	1.147699e-03	0.1	0.5860615
KEGG04930	21	19.258351	5.537710e-04	4.414966e-03	0.1	0.5860615
KEGG04664	36	62.245608	2.343224e-08	9.425648e-07	0.1	0.5860615
KEGG04722	54	56.026347	7.359200e-08	2.412054e-06	0.1	0.5860615
KEGG04912	33	13.323658	4.044143e-03	2.731966e-02	0.1	0.5860615
KEGG00280	19	38.660972	2.858611e-06	3.891902e-05	0.1	0.5860615
KEGG00310	12	28.018168	4.216839e-05	4.496029e-04	0.1	0.5860615
KEGG00380	15	103.491944	5.077894e-11	1.123424e-08	0.1	0.5860615
KEGG00640	14	47.605074	3.946596e-07	8.122213e-06	0.1	0.5860615
KEGG00650	10	4.237641	1.426346e-01	8.196420e-01	0.1	0.5860615
KEGG00020	14	13.152966	4.297080e-03	2.859183e-02	0.1	0.5860615
KEGG04012	38	23.225345	1.645928e-04	1.501617e-03	0.1	0.5860615
KEGG05220	46	38.691417	2.838289e-06	3.891902e-05	0.1	0.5860615
KEGG00564	14	58.921920	4.279369e-08	1.577934e-06	0.1	0.5860615

KEGG05340	26	146.639555	4.543033e-13	2.010185e-10	0.1	0.5860615
KEGG00500	12	28.113816	4.108093e-05	4.433499e-04	0.1	0.5860615
KEGG05120	34	65.157949	1.405379e-08	6.545760e-07	0.1	0.5860615
KEGG05323	45	85.832094	5.425801e-10	6.859398e-08	0.1	0.5860615
KEGG03040	33	19.524418	5.091239e-04	4.133493e-03	0.1	0.5860615
KEGG04660	50	10.494546	1.136995e-02	7.398434e-02	0.1	0.5860615
KEGG00410	12	46.645514	4.830102e-07	9.498695e-06	0.1	0.5860615
KEGG05221	39	41.452923	1.503599e-06	2.334411e-05	0.1	0.5860615
KEGG04340	11	6.073128	6.534459e-02	3.829595e-01	0.1	0.5860615
KEGG05218	31	20.513822	3.737548e-04	3.120334e-03	0.1	0.5860615
KEGG04512	26	24.645916	1.087092e-04	1.057171e-03	0.1	0.5860615
KEGG05146	49	71.444582	4.892881e-09	3.489577e-07	0.1	0.5860615
KEGG05222	46	43.526104	9.470522e-07	1.611724e-05	0.1	0.5860615
KEGG04610	14	72.954692	3.832568e-09	3.083311e-07	0.1	0.5860615
KEGG03030	19	22.769488	1.884236e-04	1.684303e-03	0.1	0.5860615
KEGG04622	20	53.826381	1.123894e-07	3.013920e-06	0.1	0.5860615
KEGG00970	16	23.403392	1.561698e-04	1.439614e-03	0.1	0.5860615
KEGG03015	19	46.791930	4.682779e-07	9.418270e-06	0.1	0.5860615
KEGG04970	36	63.607321	1.841606e-08	8.148673e-07	0.1	0.5860615
KEGG04370	35	31.024253	1.889009e-05	2.171018e-04	0.1	0.5860615
KEGG04662	45	44.427951	7.774477e-07	1.412827e-05	0.1	0.5860615
KEGG00051	16	26.636897	6.176816e-05	6.507370e-04	0.1	0.5860615
KEGG00052	15	19.849740	4.596460e-04	3.801543e-03	0.1	0.5860615
KEGG04114	40	21.595025	2.682605e-04	2.327430e-03	0.1	0.5860615
KEGG04540	35	9.106446	1.932494e-02	1.195920e-01	0.1	0.5860615
KEGG04914	33	18.313146	7.489572e-04	5.813964e-03	0.1	0.5860615
KEGG04070	28	20.870886	3.347487e-04	2.821303e-03	0.1	0.5860615
KEGG04720	34	7.744109	3.305226e-02	1.989775e-01	0.1	0.5860615
KEGG04730	31	78.228678	1.675825e-09	1.647807e-07	0.1	0.5860615
KEGG00561	12	88.191090	3.878922e-10	5.721105e-08	0.1	0.5860615
KEGG00330	20	70.508466	5.702581e-09	3.489577e-07	0.1	0.5860615
KEGG03008	10	14.128713	3.046345e-03	2.156697e-02	0.1	0.5860615
KEGG00520	15	8.466957	2.481070e-02	1.514228e-01	0.1	0.5860615
KEGG04672	24	44.399449	7.822850e-07	1.412827e-05	0.1	0.5860615
KEGG05144	32	81.917469	9.608954e-10	1.062934e-07	0.1	0.5860615
KEGG05310	21	32.129242	1.418916e-05	1.716427e-04	0.1	0.5860615
KEGG05320	25	15.995128	1.606747e-03	1.165488e-02	0.1	0.5860615
KEGG05322	40	70.237398	5.962643e-09	3.489577e-07	0.1	0.5860615
KEGG05330	24	19.655395	4.885585e-04	4.003252e-03	0.1	0.5860615
KEGG04612	39	55.220397	8.583898e-08	2.551955e-06	0.1	0.5860615
KEGG04940	24	9.486543	1.668563e-02	1.054714e-01	0.1	0.5860615
KEGG05332	24	10.138221	1.300856e-02	8.341999e-02	0.1	0.5860615
KEGG05143	19	24.060403	1.288248e-04	1.200041e-03	0.1	0.5860615
KEGG05214	37	16.531583	1.342565e-03	9.900887e-03	0.1	0.5860615
KEGG05219	22	48.867088	3.036379e-07	6.553786e-06	0.1	0.5860615
KEGG05223	31	16.965995	1.162369e-03	8.644046e-03	0.1	0.5860615
KEGG00270	13	9.247003	1.830088e-02	1.148611e-01	0.1	0.5860615
KEGG04966	10	42.692106	1.138933e-06	1.866486e-05	0.1	0.5860615
KEGG04621	20	55.022897	8.915738e-08	2.551955e-06	0.1	0.5860615
KEGG04623	17	17.496447	9.763539e-04	7.448511e-03	0.1	0.5860615

KEGG04330	16	14.667409	2.526630e-03	1.803184e-02	0.1	0.5860615
KEGG04964	10	29.959732	2.499846e-05	2.765308e-04	0.1	0.5860615
KEGG04150	18	11.009560	9.376387e-03	6.146416e-02	0.1	0.5860615
KEGG04973	20	13.182805	4.251686e-03	2.850410e-02	0.1	0.5860615
KEGG05216	19	30.751272	2.028858e-05	2.301852e-04	0.1	0.5860615
KEGG05020	20	16.352442	1.425270e-03	1.042394e-02	0.1	0.5860615
KEGG04742	10	9.165107	1.889037e-02	1.177260e-01	0.1	0.5860615
KEGG00562	15	18.867003	6.271148e-04	4.911213e-03	0.1	0.5860615
KEGG00510	15	7.675775	3.396901e-02	2.031147e-01	0.2	1.0000000
KEGG00250	11	9.616124	1.587530e-02	1.010712e-01	0.2	1.0000000
KEGG04960	19	6.414720	5.672222e-02	3.346433e-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 un-pooled 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
<code>var.equal=F</code>	5.5	6.8
<code>comp=8</code>	6	7.6
<code>dist.method="euclidean"</code>	4.8	6
<code>permu="ByRow"</code>	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.

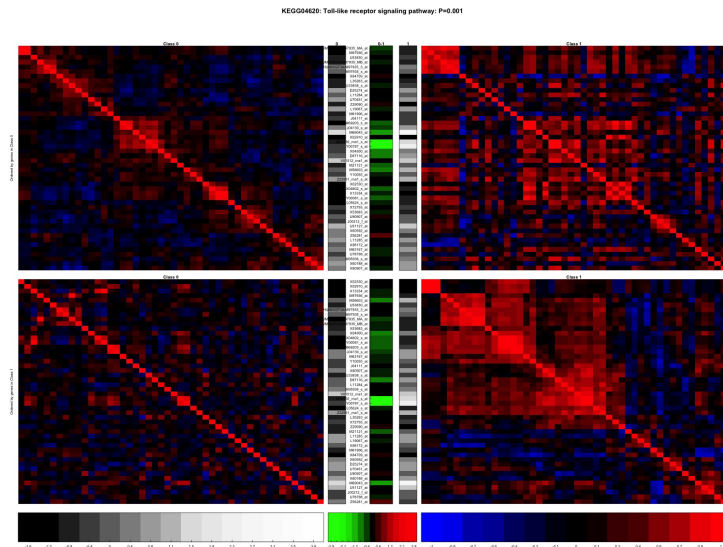


Figure 1: KEGG04620

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])
+   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.

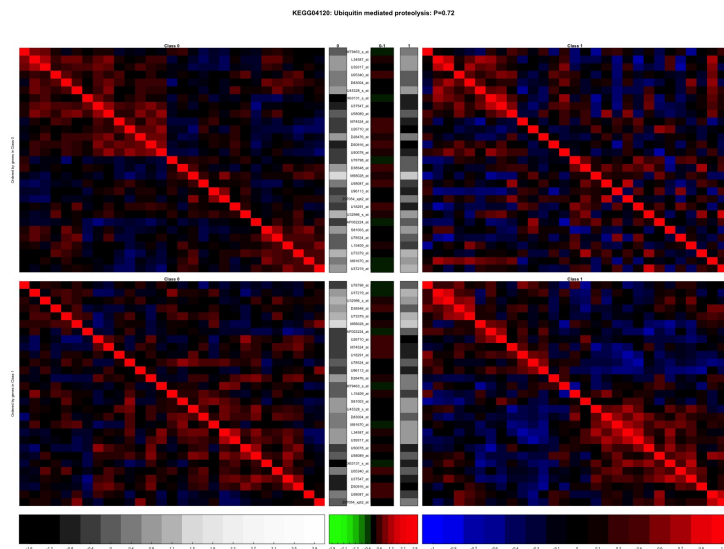


Figure 2: KEGG04120

## 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))
> #### transform data matrix only ####
> newdata <- aveProbe(x=golub, ids=ids)$newx
> #### transform both data and imat ####
> output <- aveProbe(x=golub, imat=imat, ids=ids)
> newdata <- output$newx
> newimat <- output$newimat
> newimat <- newimat[,apply(newimat, 2, sum)>=10]
> dim(newdata)

[1] 2519  38

> dim(newimat)

[1] 2519 151
```

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.