

# Using Categories to Model Genomic Data

R. Gentleman

January 4, 2026

## Introduction

The analysis of genomic data is an important challenge in bioinformatics. The particular aspect of that problem that we will address is the analysis of gene expression data in conjunction with category data. Category data, are data that map from specific entities (genes in this case) to categories, or classes. In some cases the mapping will be a partition, so that each gene maps to one and only one category, but in most cases genes will map to multiple categories. One important aspect of category data is that the categories and the mappings from entities to categories are known *a priori* and are not determined from the experimental data.

There are very many biological examples of category data, and to some extent they may be approached using similar statistical methods. Examples include the mapping from genes to chromosomes (which is nearly a partition), the mappings from genes to pathways, the mappings from genes to GO (The Gene Ontology Consortium, 2000) classifications, or the mappings from genes to protein complexes. The categories themselves may have complex relationships, as is the case with GO and protein complexes, but for now we concentrate on the mappings between genes and categories.

To make some of the concepts concrete and to provide extensive, but related, examples we will make use of a microarray data set (Chiaretti et al., 2004). The data come from a clinical trial in acute lymphoblastic leukemia (ALL). We will focus our attention on the patients with B-cell derived ALL, and in particular on comparing the group with BCR/ABL (9;22 translocation) to those with no observed cytogenetic abnormalities.

Category analysis is similar to the approach taken in Mootha et al. (2003). However, category analysis can be viewed as an extension of that methodology that makes its application both simpler and richer. Among the important concepts is the notion that the search for sets of differentially expressed genes is not always the right approach for analyzing gene expression data. Given that a microarray typically measures the levels of coordinated gene expression averaged over a few thousands of cells it seems that a more holistic approach is sometimes warranted. We are often more interested in categories where the constituent genes show coordinated changes in expression over the experimental conditions than in sets of differentially expressed genes. The methods presented in this paper are one approach to taking a more global view.

```
> library("Biobase")
> library("annotate")
> library("hgu95av2.db")
> library("KEGGREST")
> library("genefilter")
```

```
> library("Category")
> library("ALL")
```

Let's consider the comparison of two different groups, or phenotypes and we assume that there are some number of DNA microarrays that have been obtained for each group. The actual method of comparing the expression levels in the two groups is, in some sense, irrelevant to the subsequent discussion and readers can easily substitute their own favorite methods. We refer readers to von Heydebreck et al. (2004) or Smyth (2004) for a general discussion of some of the issues involved in gene filtering. Here we will use a  $t$ -test.

First we subset the ALL data to the two phenotypes that we would like to compare, those with BCR/ABL and those with no cytogenetic abnormalities, labeled NEG.

```
> ## subset of interest: 37+42 samples
> data(ALL)
> esetA <- ALL[, intersect(grep("^B", as.character(ALL$BT)),
+                           which(as.character(ALL$mol) %in% c("NEG", "BCR/ABL")))]
> esetA@annotation = "hgu95av2"
> esetA$mol.biol = factor(esetA$mol.biol)
> esetASub = nsFilter(esetA, var.cutoff=0.5)$eset
> ##set up some colors
>
> BCRcols = ifelse(esetASub$mol == "BCR/ABL", "goldenrod", "skyblue")
> library("RColorBrewer")
> cols = brewer.pal(10, "RdBu")
>
```

## Category Analysis

We have a set of data where there have been  $G$  measurements on each of  $n$  samples. We use  $\mathbf{E}$  to denote the  $G$  by  $n$  data matrix. We consider the case where a univariate test statistic can be computed for each entity (gene in our case) and denote the resulting  $G$  vector by  $\mathbf{x}$ .

There is a given fixed set of categories,  $\mathcal{C}$ , and a set of entities (genes)  $\mathcal{G}$  from which we can compute the incidence matrix  $\mathbf{A}$ , where  $a[i, j] = 1$  if entity  $j$  is in category  $i$ . The question of interest is the identification of the elements of  $\mathbf{z} = \mathbf{Ax}$  that are unusually large or small.

## Implementation

Consider the two-sample problem. Assume that there are  $n$  microarrays available, and that they have collected data on  $n$  samples under two conditions. Suppose that we have chosen to use a test statistics,  $T$ . There are several different methods of generating values of  $\mathbf{x}_T$  under the null hypothesis that there is no difference between the two conditions. These include permuting the sample labels, carrying out a bootstrap simulation, or using any one of a number of other methods for generating a reference distribution. Once this reference distribution,  $\{\mathbf{x}^b\}_{b=1}^B$  has been computed, it induces a distribution on  $\mathbf{z}$ , where  $\mathbf{z}^b = \mathbf{Ax}^b$ . Hence, for each  $z_i$  we can compute marginal tests of whether that particular  $z_i$  is extreme relative to the joint distribution.

## Parametric Assumptions

Suppose that  $\mathbf{X}$  is multivariate  $N(\mu, \Sigma)$ . The statistics are computed as  $\mathbf{Z} = \mathbf{A}\mathbf{X}$ , and hence  $\mathbf{Z}$  also follows a multivariate Normal distribution with mean  $\mathbf{A}\mu$  and variance  $\mathbf{A}\Sigma\mathbf{A}'$ . But if  $\Sigma$  is unknown then it is not possible to carry out inference on  $\mathbf{Z}$ . For our situation  $\Sigma$  is too large to easily be estimated.

We note that if  $\mathbf{X}$  is made up of two sample  $t$ -statistics with  $n$  reasonably large then the elements of  $\mathbf{X}$  are approximately  $N(0, 1)$  random variables. If the genes were independent  $\mathbf{Z}$  divided by the square root of the row sums of  $\mathbf{A}$  will itself be approximately multivariate Normal with mean zero and  $\Sigma$  will be the  $m$  by  $m$  identity matrix. We use this approximation below.

To carry this out we make use of the `rowttests` function in the `genefilter` package.

```
> ttests = rowttests(esetASub, "mol.biol")
> ##find the probes that we are going to use
> fL = findLargest(featureNames(esetASub), abs(ttests$statistic), "hgu95av2")
> fL2 = probes2Path(fL, "hgu95av2")
> length(fL2)

[1] 2058

> inBoth = fL %in% names(fL2)
> fL = fL[inBoth]
>
```

In the computation to get `fL2` we first found all duplicate mappings to LocusLink identifiers and then selected the one with the largest  $t$ -statistic. Note that we passed the absolute value of the observed  $t$ -statistic in and so obtain the most extreme value. Then we remove all LocusLink identifiers that do not map to any known pathway and find that we have 2058 genes left.

In the next code segment, we first reorder the genes by the `fL2` values and then compute  $t$ -tests, on a per row basis, using `mol.biol` variable in the phenotypic data to define the groups.

```
> eS = esetASub[match(names(fL2), featureNames(esetASub)), ]
> tobs = rowttests(eS, "mol.biol")
```

Next we create the adjacency matrix that maps genes/probes to pathways. We compute `Amat` and rearrange its columns to follow the order of genes.

```
> Amat = t(PWAmat("hgu95av2"))
> AmER = Amat[, names(fL)]
```

We will make the decision that we are not interested in pathways that have fewer than 5 members that are in the experimentally observed genes. In your own analysis, you will need to adapt this choice.

```
> rs = rowSums(AmER)
> AmER2 = AmER[rs>5, ]
> rs2 = rs[rs>5]
> nCats = length(rs2)
```

There are 193 pathways (categories) that will be used for the analysis.

```
> qqnorm(tA)
```



Figure 1: A qq-plot where each point represents a different category (in this case a pathway). There is one pathway with a remarkably low value.

```
> ##compute observed stats
> tA = AmER2 %*% tobs$statistic
> tA = tA/sqrt(rs2)
> names(tA) = row.names(AmER2)
>
```

And now we can examine the resultant qq-plot, which is shown in Figure 1.

In Figure 1 we see that there is one pathway for which the aggregate statistic seems to be unusually large, and negative. While there are a number of pathways that seem to have elevated levels of activity among those with BCR/ABL, the ribosome pathway stands out with a large and negative aggregate statistic. We can find that pathway and examine the data a bit more. Then in the next section we make use of the permutation approach to assess significance.

```
> byTT = names(tA)[tA < -7]
>
```

Figure 2 suggests that the level of activity among genes involved in the Ribosome pathway seems larger in those labeled NEG than the samples from patients with BCR/ABL. Perhaps indicating that ribosomal activity is suppressed in those with BCR/ABL.

We can further investigate the relationship between gene expression and the Ribosome pathway by examining heatmaps for this pathway. First, we examine the heatmap for those genes that were selected by our filtering approach, Figure 3, the colors in the top bar for Figure 3 are gold for BCR/ABL and blue for NEG, while in Figure 4 they are green for males and grey for females.

```

> KEGGmnplot(byTT, eS, group=eS$mol, data="hgu95av2",
+             main=paste(getPathNames(byTT)[[1]], paste("Overall:",
+             round(tA[byTT], 3)), sep="\n"))
>

```



Figure 2: NA point represents a gene in the pathway and the  $x$ -value is determined by the mean expression in the BCR/ABL group while the  $y$ -value is determined by the mean in the NEG group.

```

> tmp1 = KEGG2heatmap(byTT, eS, data="hgu95av2",
+                       main=paste(getPathNames(byTT)[[1]], paste("Overall:",
+ round(tA[byTT], 3)), sep="\n"), col=hmcol,
+                       ColSideColors=spcol)
>

```

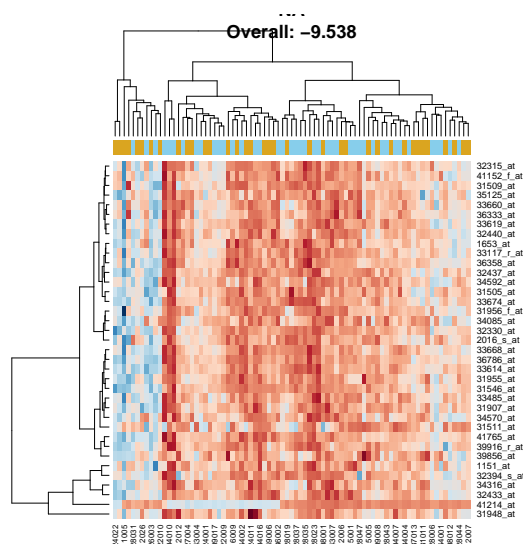


Figure 3: The heatmap for the category with the Ribosome pathway, using all gene.

```

> hmcol <- rev(colorRampPalette(brewer.pal(10, "RdBu"))(256))
> spcol <- ifelse(eS$mol.biol=="BCR/ABL", "goldenrod", "skyblue")

```

Here the patterns of expression do not seem to corroborate the finding. Yes there are differences in the patient samples with respect to the expression of mRNA for ribosomal mRNAs, but the differences are not so striking. So what is going on? Well, one of the genes gives us a hint - if you examine the heatmap carefully you will notice that one probe seems to be expressed in one set of samples and not in another - and the probe is 41214\_at, which corresponds to RPS4Y1, which is a sex-linked ribosomal protein. If we then redraw the heatmap but color the sidebar according to sex, (green for males and slate for females) we see that these probes almost exactly separate the males and females (and suggest that there may be some errors in the meta-data).

So we have found something that appears to be real, at least biological, but which may well be uninteresting. Gender is slightly confounded with the relationship between BCR/ABL and NEG cytogenetics and hence we will sometimes find effects that may be more correctly attributed to sex. To better understand this effect it will be necessary to modify the testing procedure so that sex can be adjusted for. This is reasonably straightforward and involves fitting a per gene linear model, with both a BCR/ABL effect and a sex effect and then using the standardized coefficients for the BCR/ABL effect in the GSEA analysis.

```

> spcol2 <- ifelse(eS$sex=="M", "lightgreen", "slategrey")
> tmp2 = KEGG2heatmap(byTT, eS, data="hgu95av2",
+                       main=paste(getPathNames(byTT)[[1]], paste("Overall:",
+                       round(tA[byTT], 3)), sep="\n"), col=hmcol,
+                       ColSideColors=spcol2)
>

```



Figure 4: A heatmap of the selected probes for mRNA in the Ribosome pathway. The color bar is green for males and grey for females.

## A permutation distribution

If you are uncomfortable with the Normal-theory argument given in the previous section then it is important to assess the significance of the observed test statistics with respect to a reference distribution. To that end, we consider permuting the sample labels (that is which of the two groups, BCR/ABL or NEG, a patient belongs to). In the code below we consider 500 permutations.

```
> v1 = ttperm(exprs(eS), eS$mol.biol, B=NPERM)
> permDm <- matrix(0.0, nrow=length(v1$perms[[1]]$statistic),
+                  ncol=length(v1$perms))
> for (j in 1:ncol(permDm)) {
+   permDm[, j] <- v1$perms[[j]]$statistic
+ }
> permD = AmER2 %*% permDm
> ##no need to do this second step - if we don't do it for tobs
> permD2 = sweep(permD, 1, sqrt(rs2), "/")
> pvals = matrix(NA, nr=nCats, ncol=2)
> dimnames(pvals) = list(row.names(AmER2), c("Lower", "Upper"))
> for(i in 1:nCats) {
+   pvals[i,1] = sum(permD2[i,] < tA[i])/NPERM
+   pvals[i,2] = sum(permD2[i,] > tA[i])/NPERM
+ }
> ord1 = order(pvals[,1])
> lowC = (row.names(pvals)[ord1])[pvals[ord1,1] < 0.05]
> highC = row.names(pvals)[pvals[,2] < 0.05]
> getPathNames(lowC)
```

```
$`path:hsa03450`
[1] NA
```

```
$`path:hsa03010`
[1] NA
```

```
$`path:hsa03440`
[1] NA
```

```
$`path:hsa03320`
[1] NA
```

```
$`path:hsa03030`
[1] NA
```

```
$`path:hsa03008`
[1] NA
```

```
$`path:hsa03430`
```



```
[1] NA
```

```
> lnhC = length(highC)
```

There are 55 pathways that have  $p$ -values less than 0.05 where the mean is higher in the BCR/ABL group. We print the

```
> getPathNames(highC)[1:5]
```

```
$`path:hsa04610`
```

```
[1] NA
```

```
$`path:hsa04142`
```

```
[1] NA
```

```
$`path:hsa04360`
```

```
[1] NA
```

```
$`path:hsa05130`
```

```
[1] NA
```

```
$`path:hsa05131`
```

```
[1] NA
```

```
>
```

Notice that we have used quite a large  $p$ -value, although our adjustment should be for the 193 categories that we are testing, and so it will not be too dramatic. We can visualize the differences in group means, just as we did before. These are shown in Figures 5 through 7.

Unfortunately, as we can see from the visualizations none of these category plots are especially compelling. If we return to the category NA  $p$ -value is 0.988.

## Using other functions

In the examples above we used, perhaps the simplest form of per category statistic, the summation. Extending the model to deal with virtually any other per-category function is quite simple and support for doing this is available with the (very simple) `applyByCategory` function.

```
> med = applyByCategory(tobs$statistic, AmER2, FUN=median)
> wt = applyByCategory(tobs$statistic, AmER2,
+ FUN = function(x) with(wilcox.test(x), c(statistic, p=p.value)))
> head(t(wt[,order(wt[,2])]))
```

	V	p
03010	0	1.455192e-11
04621	500	6.905757e-07

```

> KEGGmnplot(highC[lnhC], eS, group=eS$mol, data="hgu95av2",
+             main=paste(getPathNames(highC[lnhC])[[1]], paste("Overall:",
+             round(tA[highC[lnhC]], 3)), sep="\n"), pch=16, col="blue")
>
>

```

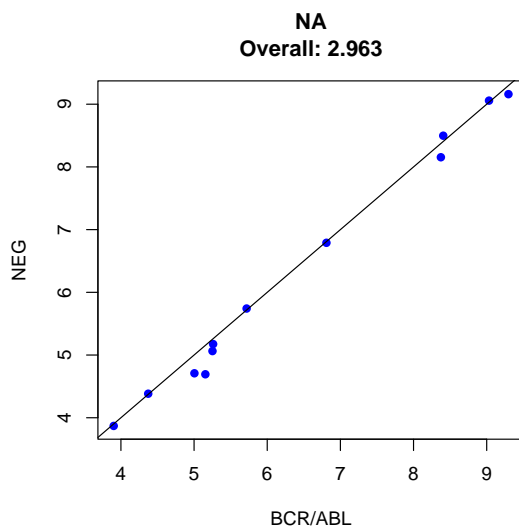


Figure 5: The per category mean plot for pathways that are deemed to be differentially expressed using a permutation approach.

```

> KEGGmnplot(highC[lnc-1], eS, group=eS$mol, data="hgu95av2",
+           main=paste(getPathNames(highC[lnc-1]))[[1]], paste("Overall:",
+           round(tA[highC[lnc-1]], 3)), sep="\n")
>
>

```



Figure 6: The per category mnplot for pathways that are deemed to be differentially expressed using a permutation approach.

```

> KEGGmnplot(highC[lnhC-2], eS, group=eS$mol, data="hgu95av2",
+             main=paste(getPathNames(highC[lnhC-2])[1], paste("Overall:",
+             round(tA[highC[lnhC-2]], 3)), sep="\n"))
>
>

```

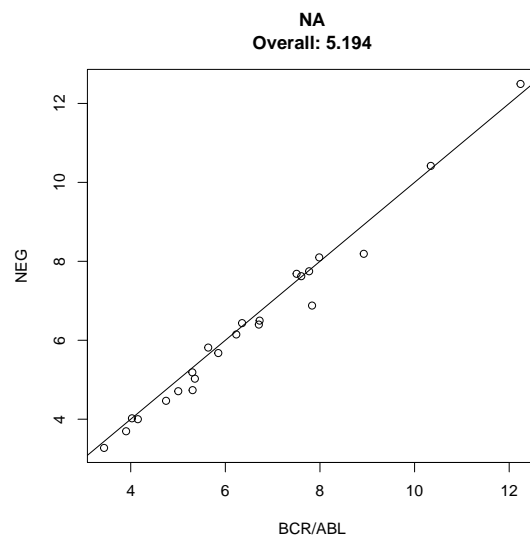


Figure 7: The per category mean plot pathway that are deemed to be differentially expressed using a permutation approach.

```

> tmp2 = KEGG2heatmap(highC[lnhC-2], eS, data="hgu95av2",
+                       main=paste(getPathNames(highC[lnhC-2]))[[1]], paste("Overall:",
+ round(tA[highC[lnhC-2]], 3)), sep="\n"), col=hmcol,
+                       ColSideColors=spcol)
>

```

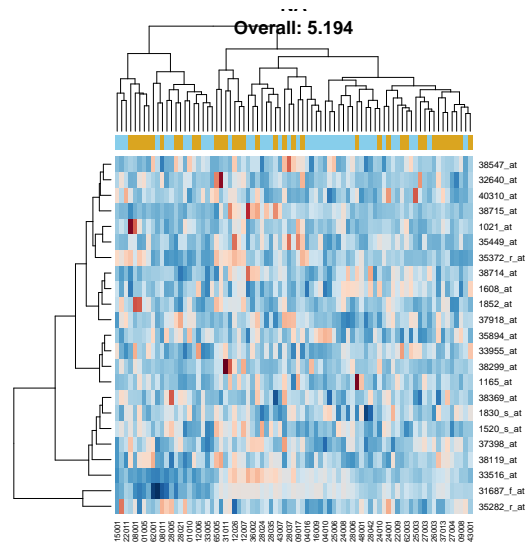


Figure 8:

03030 18 1.884997e-06  
05323 917 1.396277e-05  
05146 913 1.781248e-05  
05416 583 2.788709e-05

>

## References

- S. Chiaretti, X Li, R Gentleman, A Vitale, M. Vignetti, F. Mandelli, J. Ritz, , and R. Foa. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103:2771–2778, 2004.
- V. K. Mootha, C. M. Lindgren, et al. PGC-1 $\alpha$ -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34:267–273, 2003.
- Gordon K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):article 3, 2004.
- The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
- A. von Heydebreck, W. Huber, and R. Gentleman. Differential expression with the bioconductor project. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. John Wiley and Sons, 2004.