Section: Computational Methods for High Throughput Genetic Analysis - Expression profiling

Article: Differential Expression with the Bioconductor Project

Anja von Heydebreck[1], Wolfgang Huber[2], Robert Gentleman[3]

[1] Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, 14195 Berlin, Germany
Current address: Department of Bio– and Chemoinformatics, Merck KGaA, 64271 Darmstadt, Germany
Phone +49 6151 72 3235, Fax +49 6151 72 3329, anja.von.heydebreck@merck.de

[2] German Cancer Research Center, Department of Molecular Genome Analysis, 69120 Heidelberg, Germany
Phone +49 6221 42 4709, Fax +49 6221 4252 4709, w.huber@dkfz.de

[3] Department of Biostatistical Science, Dana-Farber Cancer Institute, Harvard School of Public Health, 44 Binney Str., Boston MA 02115-6084
Phone +1 617 632 5250, Fax +1 617 632 2444, rgentlem@jimmy.harvard.edu

Abstract

A basic, yet challenging task in the analysis of microarray gene expression data is the identification of changes in gene expression that are associated with particular biological conditions. We discuss different approaches to this task and illustrate how they can be applied using software from the Bioconductor Project. A central problem is the high dimensionality of gene expression space, which prohibits a comprehensive statistical analysis without focusing on particular aspects of the joint distribution of the genes' expression levels. Possible strategies are to do univariate gene–by–gene analysis, and to perform data–driven nonspecific filtering of genes before the actual statistical analysis. However, more focused strategies that make use of biologically relevant knowledge are more likely to increase our understanding of the data.

Keywords: differential gene expression, microarrays, multiple testing, statistical software, biological metadata

1 Introduction

The measurement of transcriptional activity in living cells is of fundamental importance in many fields of research from basic biology to the study of complex diseases such as cancer. DNA microarrays provide an instrument for measuring the mRNA abundance of tens of thousands of genes. Currently, the measurements are based on mRNA from samples of hundreds to millions of cells, thus expression estimates provide an ensemble average of a possibly heterogeneous population. Comparisons within gene across sample are supported by the current state of the technology, whereas comparisons within samples and between genes are not. In this article we focus on detecting probes that reflect differences in gene expression associated with specific phenotypes of samples. The Bioconductor Project (http://www.bioconductor.org) provides a variety of tools for this purpose. We will illustrate general principles of the analysis, rather than discussing detailed features of individual software packages; for this, we refer the reader to the Bioconductor website and the *vignettes* provided with each of the packages.

Gene expression is a well coordinated system and hence measurements on different genes are in general not independent. Given more complete knowledge of the specific interactions and transcriptional controls it is plausible, and likely, that more precise comparisons between samples can be made by considering the joint distribution of specific sets of genes. However, the high dimension of gene expression space prohibits a comprehensive exploration while the fact that our understanding of biological systems is only in its infancy means that we do not know which relationships are important and should be studied. Practitioners have adopted two different strategies to deal with this situation. One ignores the dependencies and treats each gene as a separate and independent experiment while the other strategy attempts to use relevant biological knowledge to reduce the set of genes to a manageable number.

The first strategy ignores the dependencies between genes and analyses the data gene–by–gene, for example, through statistically testing for association between each gene's expression levels and the phenotypic data. This approach has been popular, largely because it is relatively straightforward and a standard repertoire of methods can be applied. However, the approach has a number of drawbacks: most important is the fact that a large number of hypothesis tests will be carried out. This involves a problematic trade-off between specificity and sensitivity. p–value correction methods provide one mechanism for redress, by focusing on specificity, but they are not a panacea, as the price will usually be a large loss in sensitivity.

The second approach is to reduce the number of hypotheses to a more manageable number that directly address the questions of biological interest. Ideas such as non-specific filtering also fall under this rubric. But the main tool here is the use of prior biological knowledge to focus on a small number of genes and of their associations with each other. Specific examples include examining particular pathways or specific transcription factors (looking for downstream correlates) that have been associated with a disease or condition of interest.

We note that it does not have to be an either-or world. It is entirely possible, and acceptable, to first test specific hypotheses of interest and to then adopt a more exploratory approach as a basis for the next round of experimentation and hypothesis testing.

Carrying out these explorations requires the use of software analysis tools. There are very many different projects and commercial enterprises that provide software solutions. Many of the gene–expression analysis tools present themselves as a graphical user interface to a predefined set of operations on the data. Such systems are easy to use and provided that they incorporate the methodology that the user wants they perform satisfactorily. In contrast the Bioconductor project provides an extensible and flexible set of tools that are associated with a full-fledged programming language which has a large collection of numerical and statistical libraries. While more difficult to use the software is also much more powerful and can accommodate virtually any chosen analysis.

Bioconductor is an international open source and open development software project for the analysis and comprehension of genomic data. Its main engine is *R* [Gentleman and Ihaka, 1996], a language and environment for statistical computing and graphics (http://www.r-project.org). The software runs on all major computing platforms, Unix/Linux, Windows, and Mac OS X. Through its flexible data handling capabilities and well-documented application programming interface (API), it is easy to link it to other applications, such as databases, web servers, numerical or visualization software. *R* and Bioconductor support the concept of *compendiums* [Gentleman and Temple Lang, 2004] which are interactive documents that bundle primary data, processing methods (computational code), derived data and statistical output with the textual documentation and conclusions.

2 Gene selection

Fundamental to the task of analysing gene expression data is the need to select those genes whose patterns of expression are related to a specific phenotype of interest. This problem is hampered by the issues mentioned above, in that we would prefer to model the joint distribution of gene expression but the biological knowledge is lacking and instead genes will be treated as independent experiments. In this section we describe some of the analysis strategies that are in widespread use. Since our own bias is to reduce the set of genes under consideration to those that we consider to be expressed in a substantial subset of the population we will first prefilter the genes using a non-specific approach (that is, we do not use phenotypic information to aid in our decision). In addition to a short description of non-specific filtering we consider several different testing paradigms.

One popular approach is to simply conduct a standard statistical test, such as the t–test, for each gene separately. As [Kendziorski et al., 2003] indicate, analyses that treat genes as separate, independent experiments tend to be less efficient than those that adopt a Bayesian approach in order to take advantage of the shared information between genes. [Baldi and Long, 2001, Tusher et al., 2001, Lönnstedt and Speed, 2002, Smyth, 2004] proposed moderated versions of the t–statistic where the gene–specific variance estimator in the denominator is augmented by a constant that is obtained from the data of all genes. This is especially useful when few replicated samples are available and typically gene–specific variance estimates are highly variable. The moderated t–statistics may be seen as interpolations between a fold–change criterion and the usual t–statistic. Except for the variant of [Tusher et al., 2001], they coincide with the latter in the case of many replicates. Irrespective of the number of replicates, the use of a fold–change criterion may be beneficial, as this helps to screen out genes whose effect sizes are small in absolute terms, even though they may be statistically significant.

To demonstrate some of the differences between these general approaches we consider expression data from 79 samples from patients with acute lymphoblastic leukemia (ALL) that were investigated using HGU95AV2 Affymetrix GeneChip arrays [Chiaretti et al., 2004]. The data were normalized using quantile normalization and expression estimates were computed using RMA [Irizarry et al., 2003]. There are a lot of different choices for the preprocessing. This can have much impact on the results of subsequent analyses [Huber et al., 2002]. The appropriate choice of method depends on the technology and experimental design, and this is still a subject of active research. Here, we side-step these issues and instead focus on the expression values themselves.

Of particular interest is the comparison of samples with the BCR/ABL fusion gene resulting from a chromosomal translocation (9;22) with samples that are cytogenetically normal. There are 37 BCR/ABL samples and 42 normal samples (labeled NEG).

We will demonstrate some aspects of differential gene expression analysis with this example data set. Code chunks using functionality from *R* and Bioconductor illustrate the calculations.

2.1 Nonspecific filtering

This technique has as its premise the removal of genes that are deemed to be not expressed according to some specific criterion that is under the control of the user. One may also want to eliminate from consideration genes that do not show sufficient variation in expression across all samples, as they tend to provide little discriminatory power. Many of the genes represented by the 12625 probe sets on the array are not expressed in B-cell lymphocytes (either in their normal condition or in any of the disease states being considered), which are the cells that were measured in this experiment, and hence the probes for those genes can, and should, be removed from the analysis. In the example below we require estimated intensities to be above 100 fluorescence units in at least 25% of the samples, and the interquartile range (IQR) across the samples on the log base 2 scale to be at least 0.5. We start with an object, `eset`, that contains the expression data and the phenotypic information for the samples.

```
> library(genefilter)
> f1 <- pOverA(0.25, log2(100))
> f2 <- function(x) (IQR(x) > 0.5)
> ff <- filterfun(f1, f2)
> selected <- genefilter(eset, ff)
> sum(selected)
```

```
[1] 2391
```

```
> esetSub <- eset[selected, ]
```

The following analysis will be based on the expression data of the 2391 selected probe sets.

2.2 Differential expression – one gene at a time

We now turn our attention to the process of selecting interesting genes by using a statistical test and treating each gene as an independent experiment. Our setting is the simple case where we compare two phenotypes, the BCR/ABL–positive and the cytogenetically normal samples. Many of the principles of the analysis presented below are also valid in the case of a more general response variable, or in a multi–factorial experiment. We concentrate on the differences between the *mean* expression levels in the two groups. In other analyses different properties of the distributions of the expression levels could also be interesting. For instance, [Pepe et al., 2003] look at the ability of any given gene to serve as a marker for one of the groups, meaning that high (or low) expression of the gene is mainly seen in this group.

Using the *multtest* package, we perform a permutation test for equality of the mean expression levels of each of the 2391 selected genes.

```
> cl <- as.numeric(esetSub$mol == "BCR/ABL")
> resT <- mt.maxT(exprs(esetSub), classlabel = cl, B = 1e+05)
> ord <- order(resT$index)
> rawp <- resT$rawp[ord]
> names(rawp) <- geneNames(esetSub)
```

The histogram of p-values (Figure 1) suggests that a substantial fraction of the genes are differentially expressed between the two groups.

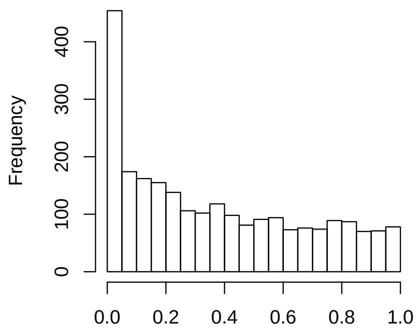

Figure 1: Histogram of p–values for the gene-by-gene comparison between BCR/ABL positive and negative leukemias.

In order to control the family-wise error rate (FWER), that is, the probability of at least one false positive in the set of significant genes, we use the permutation-based maxT-procedure [Westfall and Young, 1993]. We obtain 18 genes with an adjusted p-value below 0.05. Compare this number to the size of the leftmost bar in the histogram — clearly we are missing out on a large number of differentially expressed genes.

Looking at the description of the probe sets with smallest adjusted p–values, we see that the 3 most significant ones represent the ABL1 gene, which, in the form of the BCR/ABL fusion gene, defines one of the phenotypes we are studying.

```
1636_g_at  39730_at   1635_at  40202_at  37027_at
   "ABL1"    "ABL1"    "ABL1"   "BTEB1"   "AHNAK"
```

As illustrated above, the FWER is a very stringent criterion, and in some microarray studies, only few genes may be significant in this sense, even if many more are truly differentially expressed. A more flexible criterion is provided by the false discovery rate (FDR), that is, the expected proportion of false positives among the genes that are called significant. We use the procedure of [Benjamini and Hochberg, 1995] as implemented in *multtest* to control the FDR at a level of 0.05, which leaves us with 109 significant genes (note however that this procedure makes certain assumptions on the dependence structure between genes):

```
> res <- mt.rawp2adjp(rawp, proc = "BH")
> sum(res$adjp[, "BH"] < 0.05)

[1] 109
```

2.3 Multiple probe sets per gene

The annotation package *hgu95av2* provides information about the genes represented on the array, including LocusLink identifiers (http://www.ncbi.nlm.nih.gov/LocusLink), Unigene cluster identifiers, gene names, chromosomal location, Gene Ontology classification, and pathway associations. While the term *gene* has many aspects and can mean different things to different people, we operationalize it by identifying it with entries in the LocusLink database. One problem that does arise is that some genes are represented by multiple probe sets on the chip. The multiplicities for the HGU95AV2 chip are shown in the following table.

Multiplicity	1	2	3	4	5	6	7	8	9
No. LocusLink IDs	6719	1579	505	121	29	18	10	9	1

This leads to a number of complications, as we discuss in the following. Of the 2272 LocusLink IDs that have more than one probe set identified with them, we found that in 510 cases our nonspecific filtering step of Section 2.1 selected some, but not all corresponding probe sets.

In Section 2.2, we found that the three top-scoring probe sets all represented the ABL1 gene. However, there are 5 more probe sets on the chip that also represent the ABL1 gene, none of which passed our filtering step from Section 2.1. The permutation p–values of all eight probe sets are:

```
   1635_at  1636_g_at   39730_at  1656_s_at   32974_at 32975_g_at  2041_i_at
   0.00001    0.00001    0.00001    0.05800    0.23000    0.53000    0.59000
 2040_s_at
   0.76000
```

Some caution must be used in interpreting this particular example. The BCR/ABL–positive samples have an mRNA that is the fusion of the BCR gene and the ABL1 gene, and this fused gene is generally highly expressed. The heterogeneity in the p–values listed might be due to the specific location of the probes in the ABL1 gene and whether or not they are able to detect the fused gene. This question could be further investigated using the *hgu95av2probe* package from Bioconductor, which contains information about the probe sequences and their location.

Figure 2 shows a comparison of the t–statistics calculated in Section 2.2 between 206 *pairs* of probe sets that represent the same gene. While in many

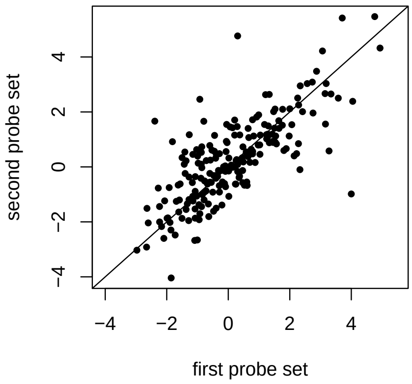

Figure 2: Comparison between the t–statistics for pairs of probe sets that represent the same gene.

cases they are approximately concordant, there are a number of cases where a *significant* p–value would be obtained for one of the probe sets but not for the other.

The consistency of the behavior of different probe sets that are intended to represent the same gene is seldom systematically reported. Here, we found some striking discrepancies. How can these be interpreted? In some cases, different probe sets may represent alternative transcripts of the same gene, which could be expressed differently. However, the magnitude of the problem suggests that there may also be errors in the mapping of the probes to LocusLink records. A better understanding of the set of all possible transcripts and their mapping to the genome, but also further and better quality control procedures are needed.

2.4 The relation between prefiltering and multiple testing

As explained above, the aim of non–specific filtering is to remove genes that, e. g. due to their low overall intensity or variability, are unlikely to carry information about the phenotypes under investigation. The researcher will be interested in keeping the number of tests/genes as low as possible while keeping the interesting genes in the selected subset.

If the truly differentially expressed genes are overrepresented among those selected in the filtering step, the FDR associated with a certain threshold of the test statistic will be lowered due to the filtering. This appears plausible for two commonly used global filtering criteria: *Intensity–based filtering* aims to remove genes that are not expressed at all in the samples studied, and these cannot be differentially expressed. Concerning the *variability across samples*, a higher overall variance of the differentially expressed genes may be expected, because their between–class variance adds to their within–class variance.

To investigate these presumed effects, we compare the scores for intensity and variability that we used in the beginning for gene selection with the absolute values of the t–statistic, which we now compute for all 12625 genes.

```
> IQRs <- esApply(eset, 1, IQR)
> intensityscore <- esApply(eset, 1, function(x) quantile(x, 0.75))
> abs.t <- abs(mt.teststat(exprs(eset), classlabel = cl))
```

The result is shown in Fig. 3. Gene selection by the interquartile range (IQR) seems to lead to a higher concentration of differentially expressed genes, whereas for the intensity–based criterion, the effect is less pronounced.

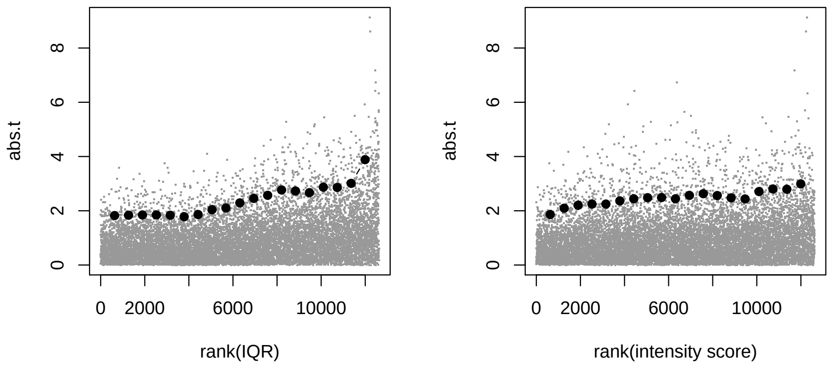

Figure 3: Plots of the absolute values of the t–statistic (y–axis) against the ranks of the values of the two filtering criteria: left, interquartile range (IQR), right, an overall intensity score. The larger dark dots indicate the 95%–quantiles of the absolute value of the t–statistic computed for moving windows along the x–axis.

2.5 Moderated t–statistics

Many microarray experiments involve only few replicated samples, which makes it difficult to estimate the gene-specific variances that are used e.g. in the t-test. In the *limma* package, an Empirical Bayes approach is implemented that employs a global variance estimator s_0^2 computed on the basis of all genes' variances. The resulting test statistic is a moderated t–statistic, where instead of the single–gene estimated variances s_g^2, a weighted average of s_g^2 and s_0^2 is used. Under certain parametric assumptions, this test statistic can be shown to follow a t-distribution under the null hypothesis with the degrees of freedom determined by the data [Smyth, 2004].

With 79 samples at hand, there is no big difference between the ordinary and the moderated t–statistic. Here we use the ALL data to illustrate the behavior of the different approaches for small sample sizes: We repeatedly draw random small sets of arrays from each of the two groups and apply different statistics for differential expression. The results are compared to those of the analysis of the whole data set. As an approximation, we declare the 109 genes with a FDR below 0.05, based on the whole set of samples, as truly differentially expressed genes.

```
> is.diff <- res$adjp[order(res$index), "BH"] < 0.05
```

Now we consider a random subset of 4 arrays per group and compute ordinary as well as moderated t–statistics for them.

```
> groupsize <- 4
> design <- cbind(c(1, 1, 1, 1, 1, 1, 1, 1), c(0, 0, 0, 0, 1, 1,
+     1, 1))
> g1 <- sample(which(esetSub$mol == "NEG"), groupsize)
> g2 <- sample(which(esetSub$mol == "BCR/ABL"), groupsize)
> subset <- c(g1, g2)
> fit <- lm.series(exprs(esetSub)[, subset], design)
> eb <- ebayes(fit)
> tsub <- mt.teststat(exprs(esetSub)[, subset], classlabel = cl[subset],
+     test = "t.equalvar")
> rawpsub <- 2 * (1 - pt(abs(tsub), df = 2 * groupsize - 2))
```

For a comparison, we define true positives as those genes that are among the 100 most significant ones in the tests based on the small data set, and which were also selected in the analysis of the full data set.

```
> TPeb <- sum(is.diff[order(eb$p.value[, 2])[1:100]])
> TPtt <- sum(is.diff[order(rawpsub)[1:100]])
```

Repeating this procedure for 50 random subsets of arrays, we then compare the numbers of true positives between the two methods. The results are shown in Figure 4.

It appears that for small sample sizes, the moderated t–statistic, which may be seen as an interpolation between the t–statistic and a fold–change criterion, has a higher power than the simple t–test.

3 Asking specific questions — using metadata

We now turn our attention to analyses that are based on asking more specific and detailed questions that relate directly to the known biology of the system or disease under study. We examine three different biological aspects. First, we ask whether differentially expressed genes are enriched on specific chromosomes.

3.1 Chromosomal location

In the following, we look at all genes with an unadjusted $p < 0.01$ (taking the median p–value in the case of several probe sets per gene), and conduct a Fisher test for each chromosome to see whether there are disproportionally many differentially expressed genes on the given chromosome. The information on the chromosomal location of genes is taken from the annotation package *hgu95av2*.

```
> ll <- getLL(geneNames(esetSub), "hgu95av2")
> chr <- getCHR(geneNames(esetSub), "hgu95av2")
```

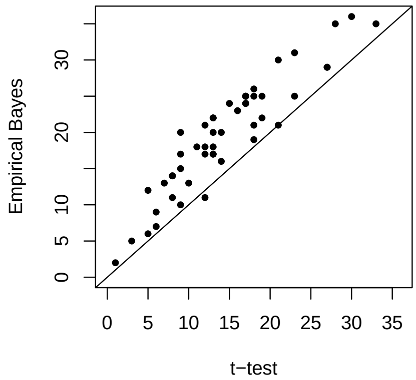

Figure 4: Number of true positives among the top 100 genes selected by the t–test (x–axis) and a test based on a moderated t–statistic, as implemented in the *limma* package. (y–axis).

```
> names(chr) <- names(rawp) <- NULL
> chromosomes <- unique(chr[!is.na(chr)])
> ll.pval <- exp(tapply(log(rawp), ll, median))
> ll.chr <- tapply(chr, ll, unique)
> ll.diff <- ll.pval < 0.1

> p.chr <- sapply(chromosomes, function(x) {
+     fisher.test(factor(ll.chr == x), as.factor(ll.diff))$p.value
+ })

> signif(sort(p.chr), 2)

     7     17      X     15      8     21      Y      3      6     22      4
0.0062 0.1100 0.1800 0.1800 0.2000 0.2900 0.3300 0.3600 0.4400 0.4800 0.5600
    12      5      9      1     10      2     16     18     20     11     19
0.5900 0.6100 0.6700 0.6900 0.6900 0.7000 0.8200 0.8300 0.8700 0.9100 0.9200
    14     13
1.0000 1.0000
```

We identify chromosome 7 as being of potential interest. Out of 87 genes that mapped there, 35 were differentially expressed according to our criterion. We further explore this by visualizing gene expression on Chromosome 7. In Figure 5 we present the output of the `plotChr` function from the *geneplotter* package. In this plot gene expression values for each sample are separated according to the strand of the chromosome that the gene is located on. Then a lowess smoother is applied (across genes, within sample) and the resultant smooth estimate is plotted. The top part of the plot represents the positive strand while the bottom part of the plot represents genes on the negative strand. High expression values are near the outside of the plotting region while low expression values correspond to the center (i.e. the expression values are mirrored). One can readily detect regions of heterogeneity between samples (such as the left end of both strands and again around `38299_at` on the negative strand).

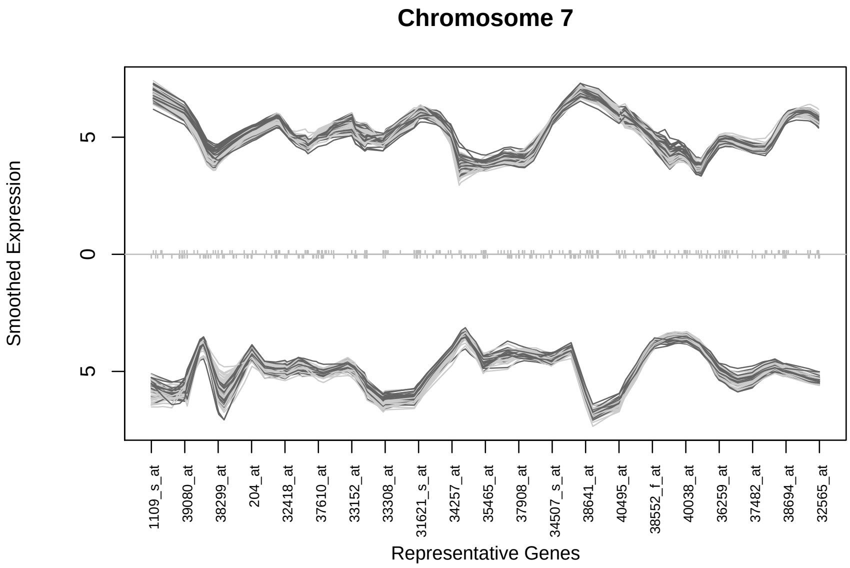

Figure 5: Smoothed expression for chromosome 7, both strands are plotted separately.

3.2 Using Gene Ontology data

A second source of valuable biological data that is easily accessible through Bioconductor software is data from the Gene Ontology (GO). It is known that many of the effects due the BCR/ABL translocation are mediated by tyrosine kinase activity. It will therefore be of interest to examine genes that are known to have tyrosine kinase activity. We examine the set of GO terms and identify the term, `GO:0004713` from the *molecular function* portion of the GO hierarchy as referring to `protein-tyrosine kinase activity`. We can then obtain all Affymetrix probes that are annotated at that node, either directly or by inheritance, using the following command.

```
> tykin <- unique(lookUp("GO:0004713", "hgu95av2", "GO2ALLPROBES"))
```

We see that 234 probe sets are annotated at this particular term. Of these only 41 were selected by the non-specific filtering step of Section 2.1. We focus our attention on these probes and repeat the permutation t-test analysis of Section 2.2.

In the analysis of the GO–filtered data, 7 probe sets have FWER–adjusted p–values less than 0.1. They are printed below, together with the adjusted p-values from the first analysis that involved 2391 genes.

```
[1] "GO analysis"

   1635_at  1636_g_at   39730_at 40480_s_at  2039_s_at   36643_at  2057_g_at
   0.00001    0.00001    0.00001    0.00002    0.00030    0.02383    0.08282

[1] "All Genes"

   1635_at  1636_g_at   39730_at 40480_s_at  2039_s_at   36643_at  2057_g_at
   0.00001    0.00001    0.00001    0.00095    0.01407    0.46938    0.82884
```

Due to the reduced number of tests in the analysis focused on tyrosin kinases, we are left with more significant genes after correcting for multiple testing. For instance, the probe set `36643_at`, which corresponds to the gene DDR1, was not significant in the unfocused analysis, but would be if instead the investigation was oriented towards studying tyrosine kinases.

3.3 Using pathways

In a closely related disease, chronic myeloid leukemia, there is evidence of a BCR/ABL-induced loss of adhesion to fibronectin and the marrow stroma. This observation suggests that there may be substantial differences between the BCR/ABL samples and the NEG samples with respect to the expression of genes involved in the integrin-mediated cell adhesion pathway. A version of this pathway was obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) as Pathway 04510. The relevant genes were identified and mapped to the corresponding Affymetrix identifiers using the *hgu95av2* and *KEGG* packages from the Bioconductor Project.

There were 110 probes that correspond to 67 LocusLink identifiers. Now we can ask whether any of these are differentially expressed between the two groups. Users can either apply our non-specific filter first, or not, depending on their particular point of view.

We did not apply the non-specific filtering of Section 2.1 and simply applied the `mt.maxT` function to obtain permutation adjusted p-values for differences between the BCR/ABL and the NEG group. This analysis identified four probe sets that had significant p-values. Three corresponded to FYN and one to CAV1. To compare with the other analyses reported above we determined that there were four probe sets included on the HGU95AV2 chip for FYN and of these three were selected by the non-specific filtering in Section 2.1 and the same three were selected here. In Section 2.2, we found two of the three FYN probes had significant p-values, after adjustment. On the other hand CAV1 is represented by a single probe set and it was not selected in Section 2.1 because it was expressed at levels above 100 in only nine samples.

A scatterplot matrix of the gene expression data for the four selected probes is presented in Figure 6. This plot provides corroboration for some of some of the points made previously about probe set fidelity. For two of the FYN probe sets the agreement is remarkably good. But their correlation with the third set of FYN measurements is poorer. We can also see that CAV1 expression is related to the BCR/ABL phenotype, but does not seem to be related to FYN expression.

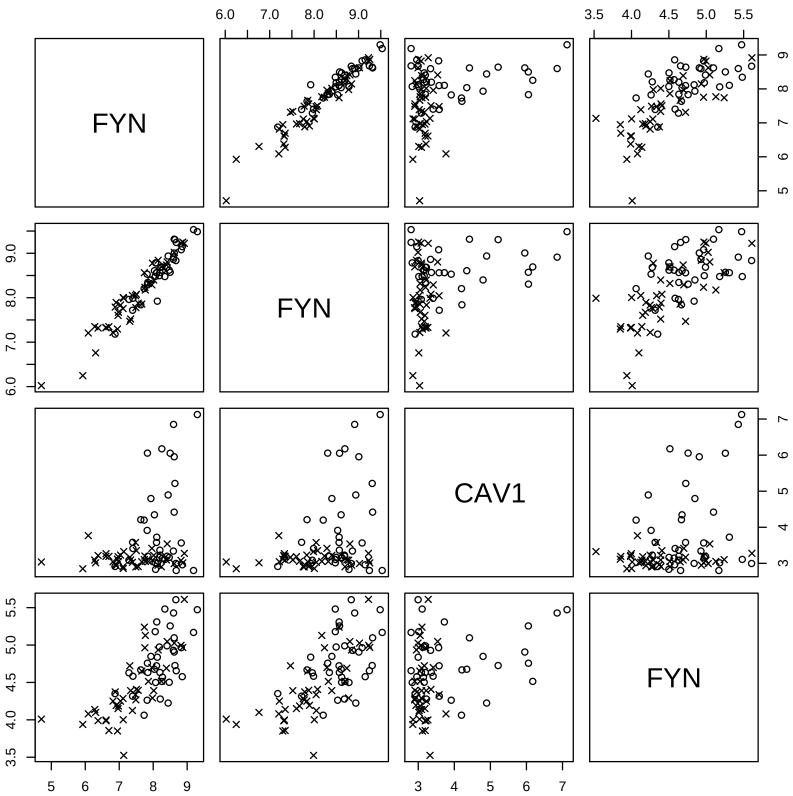

Figure 6: Pairwise scatterplots of the expression values for the four probe sets selected. Crosses indicate BCR/ABL negative samples, circles denote BCR/ABL positive samples.

4 Discussion

Understanding the complex dependencies between the transcriptional activities of genes remains an important, difficult, and essentially unsolved problem. Current analysis of differential gene expression mostly works on a gene–by–gene basis, where the necessary concepts are within reach and established statistical tests can be used. Statisticians have directed a lot of attention to methods for p–value correction. This has led to widespread adoption of these methods and particularly of the false discovery rate (FDR). They rely on the assumption that those tests with the most extreme p–values are most likely to have arisen from hypotheses that are truly false. The decision boundary is shifted sufficiently to ensure that some required proportion of those hypotheses deemed false are in fact truly false. But such an approach cannot be viewed as a solution to the real problem of identifying all truly false hypotheses. We note that in adjusting the rejection boundary one has also taken an unknown number of truly false hypotheses and deemed them to be true. Furthermore, there is no guarantee that the biologically most relevant genes are the ones with the most extreme p-values. Genome–wide expression profiles may be dominated by secondary changes in gene expression, such that the primary effectors may be buried within long lists of statistically significant genes. There is no simple fix to this problem apart from testing fewer and more directed hypotheses and where possible this should be done.

In the examples of Section 3 we demonstrated a variety of analyses incorporating biological information that are currently available and easily carried out. We showed that such analyses have the capability of extracting more information from the data than an omnibus approach can. Increases in our understanding of the relevant biology coupled with gains in statistical knowledge will likely lead to even more promising analyses.

Acknowledgements

We are grateful to Drs. J. Ritz and S. Chiaretti of the DFCI for graciously providing their data. And also, to the developers of R and those of the Bioconductor software packages that have made it possible to so easily consider a wide variety of approaches to analysing these data.

References

[Baldi and Long, 2001] Baldi, P. and Long, A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t–test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300.

[Chiaretti et al., 2004] Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood* 2004 (in press)

[Gentleman and Ihaka, 1996] Gentleman, R. and Ihaka, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314. http://www.r-project.org.

[Gentleman and Temple Lang, 2004] R. Gentleman and D. Temple Lang. Statistical analyses and reproducible research. *Unpublished Manuscript*, 2004.

[Huber et al., 2002] Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. *Bioinformatics*, 18 Suppl. 1, S96–S104.

[Irizarry et al., 2003] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264.

[Kendziorski et al., 2003] Kendziorski, C., Newton, M., Lan, H., and Gould, M. (2003). On parametric Empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22(24):3899–3914.

[Lönnstedt and Speed, 2002] Lönnstedt, I. and Speed, T.P. (2002). Replicated microarray data. *Statistica Sinica*, 12:31–46.

[Pepe et al., 2003] Pepe, M., Longton, G., Anderson, G., and Schummer, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics*, 59:133–142.

[Smyth, 2004] Smyth, G. (2004). Linear models and Empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Article 3.

[Tusher et al., 2001] Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98:5116–5121.

[Westfall and Young, 1993] Westfall, P. and Young, S. (1993). *Resampling-based multiple testing: examples and methods for p-value adjustment*. John Wiley and Sons.