Data Processing Effects on the Interpretation of Microarray Gene Expression Experiments

Katrin Fundel[a], Robert Küffner[a], Thomas Aigner[b], and Ralf Zimmer[a]
[a]Institut für Informatik, Ludwig-Maximilians-Universität München,
Amalienstrasse 17, 80333 München, Germany
[b]Lehrstuhl für Pathologie, Universität Erlangen-Nürnberg,
Krankenhausstr. 8-10, 91054 Erlangen, Germany
{fundel,kueffner,zimmer}@bio.ifi.lmu.de, thomas.aigner@patho.imed.uni-erlangen.de

Abstract:
Motivation: Microarray gene expression data is collected at an increasing pace and numerous methods and tools exist for analyzing this kind of data. The aim of this study is to evaluate the effect of the basic statistical processing steps of microarray data on the final outcome for gene expression analysis; these effects are most problematic for one-channel cDNA measurements, but also affect other types of microarrays, especially when dealing with grouped samples. It is crucial to determine an appropriate combination of individual processing steps for a given dataset in order to improve the validity and reliability of expression data analysis.

Results: We analyzed a large gene expression data set obtained from a one-channel cDNA microarray experiment conducted on 83 human samples that have been classified into four Osteoarthritis related groups. We compared different normalization methods regarding the effect on the identification of differentially expressed genes. Furthermore, we compared different methods for combining spot p-values into gene p-values, and propose Stouffer's method for this purpose. We developed several quality and robustness measures which allow to estimate the amount of errors made in the statistical data preparation.

Conclusion: The apparently straight forward steps of gene expression data analysis, i.e. normalization and identification of differentially expressed genes, can be accomplished by numerous different methods. We analyzed multiple combinations of a number of methods to demonstrate the possible effects and therefore the importance of the single decisions taken during data processing. An overview of these effects is essential for the biological interpretation of gene expression measurements. We give guidelines and tools for evaluating methods for normalization, spot combination and detection of differentially regulated genes.

1 Introduction

Today, numerous methods and tools exist for analyzing gene expression data. New normalization techniques are presented, e.g. [Edw03, FC04, WBHW03, ZLS05], so are methods for detecting differentially expressed genes, e.g. [CNGGC04, CC03, CHQ$^+$05, TTC01, YDFQ05]. Tools aim at analyzing microarray data in a largely automated way, e.g. [CKP$^+$04, HZZL02, HVAS$^+$04, KWSPF03, PGM04], many of them even integrate

gene expression data with further information obtained from e.g. ontologies, pathway databases or text mining.
Yet, comparisons between different normalization methods were focussed mainly on Affymetrix and two-channel cDNA microarrays [BIAS03, PYK+03], and do not consider sample groups. Generally, existing literature offers little guidance on how to decide which method to use, how to compare different methods and their outcomes, especially for one channel cDNA data, and how to check the correspondance of possible outcomes to a biologists expectation and downstream interpretation.
The aim of the study presented here is to demonstrate that the 'higher-level' outcome, i.e. a list of differentially regulated genes, of any microarray experiment is closely related to the 'low-level' details of data processing, that individual microarray data processing steps can not be considered as independent and that it is crucial to be careful in every decision taken during microarray data processing in order to obtain reliable results. More precisely, our goal is to investigate the importance of cDNA microarray data normalization and processing for the identification of differentially expressed genes. Therefore we apply different normalization techniques and evaluate the differences in the final result.
We also compare different methods for combining spot p-values to gene p-values, that is another neglected problem.
Finally, the large number of samples allows us to perform a stability analysis on the significantly regulated genes. Recently, it has been shown [MKH05] that in numerous published large studies on differential gene expression differentially expressed genes are highly unstable for subsets of the analyzed samples. Thus, we propose a procedure which estimates the errors and quantifies their amount via a robustness analysis, because a gold standard is not available.

We present a study conducted on one-channel cDNA microarray data analysis. The analyzed dataset represents 83 samples of human joint cartilage classified into four disease-related groups of osteoarthritis (OA), for reviews on osteoarthritis see [ABZZ02, AD03, ABSZ04]. Given the difficulty of obtaining human joint samples this represents a large data set. On the other hand, groups of about 20 samples allow for statistical robustness and quality analysis. The data was collected to identify differentially regulated genes which are of potential interest for understanding disease mechanisms, diagnosis and medical therapy. The full dataset used for this study and its biological interpretation is going to be published in the near future. The study presented here is not intended to focus on the intrinsic content of the underlying data.

2 Dataset

The data analyzed in the present study was obtained from a custom designed cDNA microarray. The microarrays were produced and measured by GPC-Biotech AG (Martinsried, Germany). A part of the spotted cDNA had been preselected for OA-relevant genes. Scanning of the radiolabeled arrays was done by phosphorimaging. Each microarray contains 7467 spots, 5517 spots represent 3648 genes, there are 1 to 74 spots per gene on the array, and 1062 genes are represented by more than one spot. The fact that the number of spots

per gene varies complicates data analysis as the information obtained from several spots needs to be combined for obtaining information for a gene, which is needed for biological interpretation; yet this applies to most cDNA and oligonucleotide microarrays and therefore represents a common problem.
Primary data analysis, i.e. local background correction and removal of outlier spots, was also done by GPC-Biotech with proprietary software. The data set is described in the following as: $X = X_{ks} = \{x_{ks}|k = 1...83, s = 1...7467\}$, where s: spots, k: samples.
83 samples of human cartilage were analyzed. The samples were classified based on histological criteria into four groups: normal (*n*), early degenerative cartilage (*e*), peripheral (*p*) and central (*c*) Osteoarthritis. Furthermore the class 'late OA' (*l*) was defined as the combined set of peripheral and central OA, this represents all samples of patients severely affected by Osteoarthritis. It is known that *n* and *e* as well as *p* and *c* are very similar from a physiological point of view, whereas *n* is very different from *p* and *c* and, consequently, also from *l*.
One of the main goals of the experiment was to identify differentially regulated genes for the group pairs *ne*, *np*, *nc*, *ep*, *ec*, *pc*, *nl*, *el*. The expression value distribution (figure 1) shows most data concentrated in a very small range (75% of the values are $<$0.13, 99% are $<$5.91) and some values are significantly larger (overall maximum at 539.9). Due to the technique (cDNA spots of different sequences, radioactive detection), the expression values for different spots representing the same gene can vary significantly.

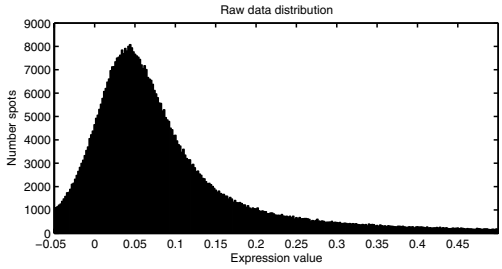

Figure 1: Distribution of raw data of the analyzed dataset (before outlier removal, i.e. 83*7467 spots).

Outlier detection
Three outlier detection methods relying on different principles were chosen for obtaining a meaningful result: cluster analysis, PCA and analysis of expression value distributions. These analyses were done with the expression values, and with data transformed to z-scores over spots, both before and after applying the normalizations described below. The z-score transformation normalizes the values for each spot over all samples, i.e. every spot mean is 0 with standard deviation 1. Thus, the contribution of every spot to an overall analysis is the same, and consequently this transformed data can yield complementary results to results obtained directly from the raw data.
In cluster analysis, we noticed samples that were clustered with significant distance to all other samples. For PCA analysis, the first two main components were plotted against

each other; this yields dense groups of samples belonging to the same sample class and the groups of different classes partially overlapping with each other. Some samples were clearly separated from the dense group of all other samples belonging to the same sample class in this PCA plot; this was interpreted as an indication for being outliers. For analysis of expression value distributions boxplots and histogramms were inspected visually. We classified a sample as outlier if it was conspicuous in at least two of the three types of analysis.
We found 5 of the 83 samples to be outliers. The outlier analysis before and after normalization yielded the same set of outlier samples, this confirms that the classification of a sample as outlier does not depend on the type of normalization applied to the data.
The outlier samples were removed for further analysis. The remaining 78 samples are classified as follows: 18 normal (*n*), 20 early degenerative cartilage (*e*), 21 peripheral OA (*p*), and 19 central OA (*c*). The raw data excluding outlier samples represents the starting point for all further analysis we present here.

3 Data Processing

3.1 Normalization

Centralization [ZAZL01] is a normalization method that estimates for each pair of arrays the quotient of the constants of proportionality and subsequently computes an optimally consistent scaling for the samples based on the matrix of pairwise quotients. Centralization needs two parameters describing the range of reliable measurements to be used. For estimating these parameters, we analyzed the distribution of expression levels. We investigated the effect of different pairs of parameters (3 lower limits, 4 upper limits) and the effect of iterative application of centralization (up to 10 iterations). We required the centralized data to show little variation of the 25%, 50%, and 75% percentiles after applying centralization once, and also after several iterations of centralization. The range of expression values 0.03-1 produced reliable yet conservative results, i.e. the smallest variance after one and up to 10 iterations, and therefore was applied for final centralization.

Percentile Normalization is a method that adjusts a certain percentile to the same level for all samples by applying a multiplicative factor to each sample. We used the 50% (eq. median) and 75% percentiles, which are typically used.

MAD Scale Normalization adjusts the median of all samples to a common expression level and the median absolute deviation (MAD), which is a very robust measure of scale for the variability of distributions, to a common level. Typically, the median and MAD are fixed to 0 and 1, respectively; we applied a variance in that we transformed the median and MAD back to the original scales. For each sample k and spot s the original value x_{ks} was transformed into the normalized value x''_{ks} according to the following equation:

$$x'_{ks} = \frac{x_{ks} - \mathrm{median(x_k)}}{\mathrm{MAD(x_k)}}$$

$$x''_{ks} = x'_{ks} * \mathrm{MAD(X)} + \mathrm{median(X)}$$

$$\mathrm{MAD}(\mathrm{x_k}) = \mathrm{median}(|\mathrm{x_k} - \mathrm{median}(\mathrm{x_{k\cdot}})|)$$

where: MAD: median absolute deviation; k: sample; s: spot measured in sample k; X: entire dataset.

Flooring
The raw expression intensities contain negative values due to background correction of the original data performed by GPC-Biotech. Negative data are not appropriate for computing fold changes and p-values. Also expression values very close to 0 are not appropriate because they lead to inappropriate high fold changes; therefore it is important to estimate a reliable floor value. The general lower bound of measured intensity accurracy was estimated by an analysis of p-values versus individual spot expression values: p-values for all group comparisons were calculated by the two-sided Wilcoxon ranksum test and plotted against the underlying spot expression values (data not shown). This analysis showed that p-values smaller than 10^{-3} were based almost exclusively on expression values above 0.01. Therefore, the background level was estimated to be at 0.01 and expression values $<$0.01 were set to 0.01 for all further analysis.

3.2 Differential Expression

p-values
Differently expressed genes were detected based on the following procedure: First, the two-sided Wilcoxon ranksum test was applied for calculation of p-values for spots. Next, these spot p-values are combined to obtain overall **gene p-values**. This two-step procedure is necessary because of the high variability in expression values measured for different spots representing the same gene. We applied three different methods for combining spot p-values into gene p-values:

(1) Fisher's inverse chi-square method [Fis32]. This method uses the fact that given a uniform distribution U, $-2*log(U)$ has a chi-square distribution with two degrees of freedom, and the sum of two independent chi-square variables is again chi-square distributed (with four degrees of freedom). Consequently, the combined p-value $p_{chi}(g)$ for a gene g can be computed as:

$$p_{chi}(g) = 1 - \chi^2_{2d}(\sum_s -2 * log(p_s))$$

where p_s are the p-values for spots s representing gene g (in our case obtained from the two-sided Wilcoxon ranksum test), d is the number of spots s representing gene g, and $\chi^2_d(x)$ is the cumulative distribution function of the chi-square distribution with d degrees of freedom.

(2) A variant of Fisher's inverse chi-square method that also considers the directions associated to individual spot p-values:

$$p_{dirchi}(g) = \min(1 - \chi^2_{2d}(\sum_s -2 * log(p_s^{dir})))$$

where p_s^{dir} are the onesided spot p-values (Wilcoxon ranksum test) for all spots s representing gene g; these onesided spot p-values are determined for both regulation directions; the overall combined gene p-value then equals to the smaller of the two combined p-values, each of them corresponding to one test direction.

(3) Stouffer's method [Ros84]. This method transforms p-values to z-scores assuming a normal distribution($p_s \rightarrow Z_s$), which is a straightforward calculation as the onesided p-value $p_s^{onesided}$ corresponds to the area under the normal cumulative distribution function between $-\infty$ and $-|Z_s|$.

$$p_s^{onesided} = \frac{1}{\sqrt{2\pi}} \int_{-\inf}^{-|Z_s|} e^{\frac{-t^2}{2}} dt$$

Each Z_s gets the sign corresponding to the regulation direction of the corresponding spot, the z-scores of spots representing one gene are summed, and the sum is scaled in order to account for the number of combined spots:

$$Z_{overall} = \sum_s Z_s / \sqrt{k}$$

where k is the number of tests, i.e. the number of spots to be combined. Finally the z-scores are transformed back to p-values ($Z_{overall} \rightarrow p_{overall}$) by the integration of the area under the curve as described above.

q-values

The gene p-values are converted into q-values by use of the R-library 'qvalue'[ST03]. The q-value quantifies the false discovery rate, i.e. a q-value of 0.01 indicates that when selecting significant genes as the subset of all genes having a q-value $\leq$0.01, 1% of the selected genes have to be expected to be false positives. The q-value computation implies the estimation of π_0, i.e. the number of non-regulated genes. This is done by analyzing the distribution of p-values; the uniform distribution underlying the given p-value distribution is estimated and the area under this uniform distribution estimates π_0. The number of regulated genes can thus be estimated by $1 - \pi_0$. Different methods for estimating π_0 are implemented in the applied R-library, we used the bootstrap method as this is recommended as a robust method by the authors.

Fold change

Given two sample groups $C_1, C_2 \in \{n, e, p, c, l\}, C_1 \neq C_2$ the overall fold-change for a gene g was estimated as follows: A spot s for the gene g is taken into account if at least one expression value in the groups under investigation is above the floor value (0.01); for each spot we compute fold-changes ($\mathrm{sfc}_{\mathrm{S_g}}^{\mathrm{C_1,C_2}}$) for all pairs of samples derived from the two groups to be compared. The median of these spot fold-changes is used as overall estimate for the gene-fold change ($\mathrm{fc(g)}^{\mathrm{C_1,C_2}}$).

$$S'_g := \{s\ spot | s \text{ represents gene g}\}$$

$$s \in S_g := \left\{ s \in S'_g | \exists k \in \{C_1 \cup C_2\} : \mathrm{x_{ks}} > 0.01 \right\}$$

$$\mathrm{sfc}_{\mathrm{S_g}}^{\mathrm{C_1,C_2}} := \{\log_2(\mathrm{x_{is}}/\mathrm{expr_{js}}) | \mathrm{i} \in \mathrm{C_1} \wedge \mathrm{j} \in \mathrm{C_2}, \mathrm{s} \in \mathrm{S_g}\}$$

$$\mathrm{fc(g)}^{\mathrm{C_1,C_2}} = 2^{\mathrm{median}(\mathrm{sfc}_{\mathrm{S_g}}^{\mathrm{C_1,C_2}})}$$

where: $\mathrm{x_{ks}}$ is the expression value of spot s in sample k.

We apply three different methods for combining spot p-values into gene p-values, and one method for computing fold changes. The independent determination of gene-fold change and directed gene p-value makes it possible that the gene p-value is given for the opposite direction to the fold change-direction. We analyzed the data for this and found that this effect only occurs for few fold changes that are very close to 1; therefore this does not imply problems for further biological interpretation of data.

4 Results and Discussion

4.1 Normalization

Normalization can significantly change the original data. Typically, the effect of normalization is evaluated by visual inspection of boxplots. A boxplot shows the 25% percentile and 75% percentile of a dataset as lower and upper boundary of a box and the median as horizontal line within the box, it shows whiskers of a length that is typically proportional to the interquartile range, and all data points lying outside these whiskers are displayed individually as outliers. For between slide normalizations, the individual samples are listed on the x-axis, and the expression values are plotted as a dataset on the y-axis (Figure 2, left panel).
Boxplots of the type of data described above typically show small boxes and whiskers but numerous outlier values due to the data distribution. While boxplots are easy to generate and interprete, we suggest in addition to these a different type of plot for evaluating the effect of normalization, especially for experiments dealing with samples belonging to different classes.

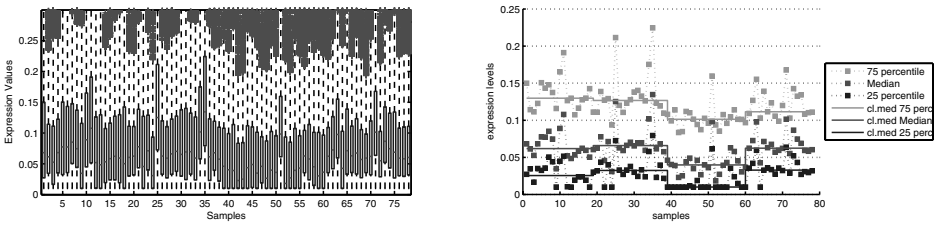

Figure 2: Boxplot (left panel) and group-level plot (right panel) for the same data. The group-level plot shows the 25%, 50% and 75% percentile for each sample (as does the boxplot) and additionally shows the median over these values for each sample group representing different disease stages (cl.med: class median).

This **group-level plot** also shows the 25%, 50% and 75% percentiles for the individual samples as does the boxplot. Data displayed as outlier in boxplots is ignored as it is not in the focus of normalization. Most importantly the plot additionally shows the group-levels of the plotted percentiles, i.e. the median of the corresponding percentile over all samples belonging to the same group. This group-level allows to identify group-specific variations within data, which may not be inherent to the biological samples under investigation. For the investigated samples, analysis of total mRNA content showed no group specific variations on the mRNA level; variations must be due to experimental setup or any other undesired effect. Figure 2 shows a boxplot and group-level plot for our dataset. The group-level plot clearly shows the different levels of expression data for the different sample groups, in the boxplot this is much less evident.

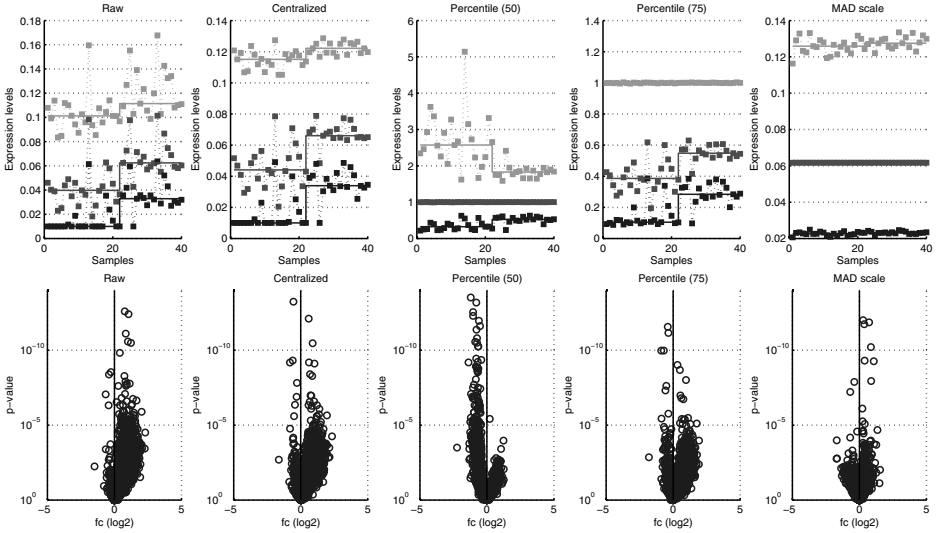

Figure 3: Effects of normalization: group-level plots for p(1-21) and c(22-40) samples and volcano plots for group comparison p versus c for raw data and four different normalizations (for details see section 4.1).

Differing levels of expression data between groups of samples evidently affect the calculation of p-values and fold changes. This results in an artificially high number of differentially regulated genes. Another effect may be that more genes seem to be regulated in one direction than in the other even though this might not be the case after adequate normalization and might also not be expected from prior biological knowledge. Figure 3 contrasts normalized data with raw data for the comparison pc, it shows group-level plots and the resulting p-values and fold changes. The raw and centralized data yield assymetric fold-change distributions, more genes appear upregulated than downregulated from p to c due to the differences in group level. The 50% percentile normalization produces more down-regulated than upregulated genes. For the analyzed data, we expect that approximately the same number of genes are up- and down-regulated. Only the 75% percentile normalization and the MAD scale normalization yield approximately symmetric distributions.

Far more normalization techniques than the ones analyzed here exist (e.g. [HvHS$^+$02, YDL$^+$02, SS03, Edw03, BGOT04]); most of the newer normalization techniques are non-linear as this is assumed to perform generally better than linear techniques. Some of them focus on two-channel or Affymetrix-type data and can therefore not easily be applied to one-channel cDNA data, others can directly or after slight adaptation be applied to this kind of data. The study presented here concentrated on some normalization techniques and shows that different normalization methods yield different results and therefore we propose analyses that should be performed after any normalization to test its appropriateness.

4.2 p-value combination

Within cDNA microarrays, the number of spots per gene typically varies, and the experimental technique of spotting different clones for a given gene results in high variability of the measured data for different spots representing the same gene. This makes it necessary to combine spot p-values to gene p-values to simplify biological interpretation. In principle, different cDNAs representing the same gene do not need to be identical, they can be splicing variants, or they can cover distinct regions within the gene sequence; therefore, it could be interesting to integrate sequence information about the spotted cDNAs. Here, we do not take sequence information into account, we make use of the gene annotation as provided by GPC-Biotech and consider each spot representing the same gene as replicate, irrespective of the specific cDNA being spotted.
To our knowledge, there is no comprehensive analysis about how to best combine spot p-values into gene p-values available, therefore we compared three methods for this task. Figure 4 shows examples of gene p-values obtained from the three investigated methods. We required the total p-value resulting from significant p-values with the same direction of regulation to be at least as significant as the most significant p-value; a total p-value resulting from p-values with inconsistent direction should generally be of lower significance than the most significant underlying spot p-value; and if the individual p-values are approximately equal but the fold changes point in opposite direction, the gene p-value should tend towards 1.
Stouffer's method has so far been, to our knowledge, predominately been used in social sciences, we are not aware of its application for gene expression data. This method assumes a normal distribution of the underlying spot p-values, which might not be the case for the analyzed data. However, we found Stouffer's method to result in the most plausible results, especially in cases where multiple spots were investigated for a gene and individual spots showed fold-changes in opposite direction. This combination method was therefore applied in all other analyzes based on gene p-values shown here if not indicated otherwise.
The most predominantly used method for combining p-values of the analyzed ones is Fisher's inverse chi-square method. This method has the drawback that it does not consider the signs of changes, i.e. the combination of two spots with a significant p-value and opposite regulation direction results in the same gene p-value as two spots of the same direction, and the resulting gene p-value is of higher significance than the p-value of the

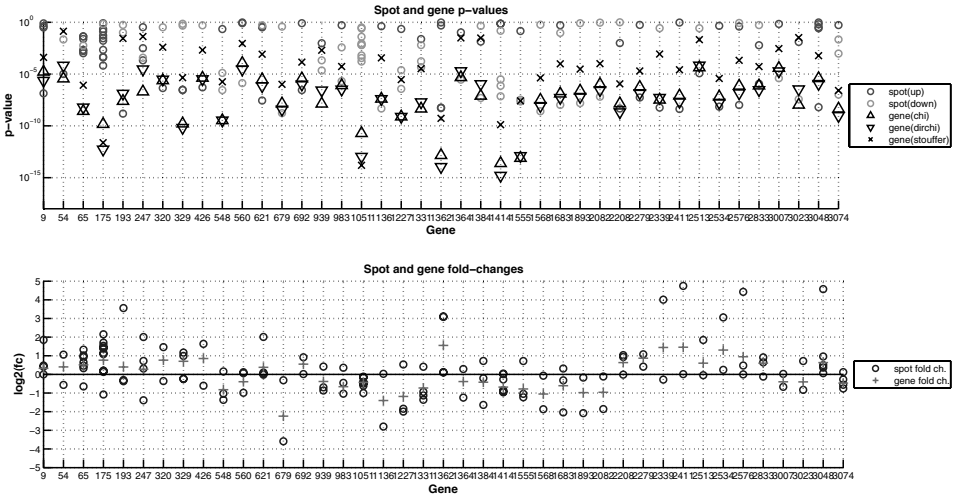

Figure 4: Different methods for combining spot p-values to gene p-values; upper figure: spot and gene p-values; lower figure: spot and gene fold changes. For details see section 3.2 and 4.2.

corresponding spots.
The presented variant of Fisher's inverse chi-square method partially eliminates this effect. Generally, the method renders p-values that are slightly more significant than the original Fisher's inverse chi-square method. If, however, two spots are regulated in opposite directions and both have significant p-values the resulting gene p-value is clearly less significant than the respective value of the original Fisher's inverse chi-square method. Therefore, the variant corresponds better to our requirements than the original method.
Compared to Stouffer's method, we clearly favor the latter because it reflects p-values of opposite direction in a more prominent decrease in significance of the resulting overall p-value than does the chi-square variant. This is obvious given the calculation methods: In Stouffer's method, two spots of opposite directions and approximately equally significant p-values nearly cancel each other out due to the summing of z-scores; in the chi-square variant, one value cannot cancel out another, the more significant one has highest influence, and the less significant one has a minor, yet still increasing effect on the overall significance.

4.3 Number of regulated genes

Cluster analysis of the analyzed expression data showed a good separation between the class pairs *ne* and *pc*, whereas the groups *n* and *e* as well as *p* and *c* were not separated from each other (results not shown here). Interestingly, also in terms of clinical staging, *n* and *e* and *p* and *c* resemble each other, whereas these two group pairs are clearly distinct.

The result of the cluster analysis lead to the expectation, that more genes are significantly regulated in the comparisons *np*, *nc*, and *nl* than in the comparisons *ne* and *pc*.

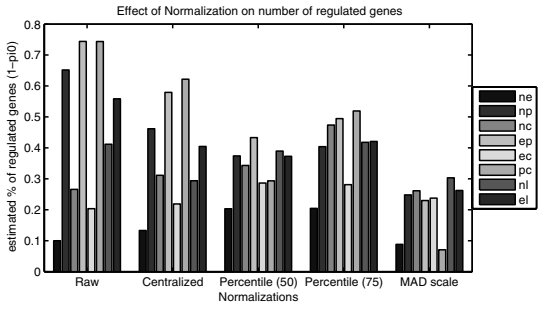

Figure 5: Effect of normalization on the number of significantly regulated genes.

Figure 5 shows the estimated number of significantly regulated genes for different group comparisons and different normalizations. Once again, the figure shows the very important effect of the type of normalization on the outcome of differentially expressed genes. The effect is most pronounced in the comparison *pc*; depending on the method of normalization between 7% and 74% of the genes are regulated. Overall, only the percentile normalization to the median and the MAD scale normalization yield results which support the expectation to find less genes to be regulated in the comparisons *ne* and *pc* compared to *np* and *nc*. The MAD scale normalization yields the smallest number of regulated genes in all comparisons.

4.4 Robustness analysis

The large number of samples allows us to assess the robustness of the differentially regulated genes between two sample groups; we performed a leave-one-out and a subset sampling approach that both show the varying robustness of differentially regulated genes for different group comparisons.

Leave-one-out Analysis
We performed a leave-one-out analysis for estimating the robustness of p-value calculation, i.e. we disregarded one sample at a time and calculated p-values based on the remaining samples. The resulting lists of p-values were compared to the list derived from the full dataset. This analysis was conducted on the MAD scale normalized dataset and with the Stouffer method for combining p-values.
For estimating the robustness, the p-values obtained from the full dataset were considered as standard of truth, we applied a series of cutoff-p-values (between 10^{-7} and 10^{-1}) and determined the fraction of significantly regulated genes from the full dataset that are also significantly regulated to the given cutoff p-value in the leave-one-out datasets. We se-

lected 'robust' differentially expressed genes according to two criteria:
exact: The fraction of genes that are significant in all leave-one-out datasets
relaxed: The fraction of genes that are significant with a p-value of $\leq$2* the cutoff p-value in all leave-one-out datasets.

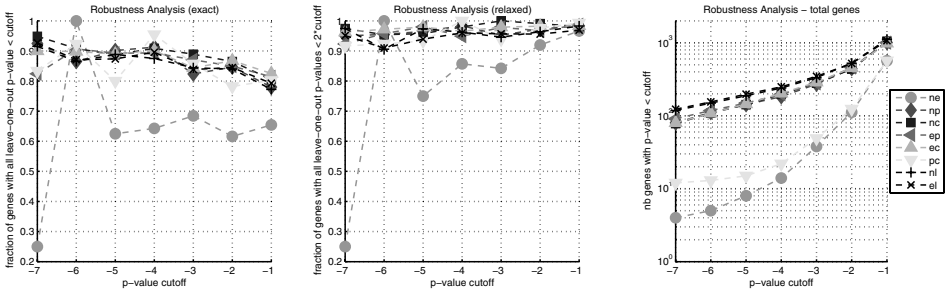

Figure 6: Robustness Analysis of p-value calculation: leave-one-out analysis. Fraction of genes significant at a certain p-value level in the overall p-value calculation that are also significant in the leave-one-out p-value calculation according to two different measures for evaluation *exact* (left panel) and *relaxed* (middle). The right panel shows the total number of genes significant at a certain p-value-level for each group comparison. For details see section 4.4.

The results of this analysis (figure 6) show that the p-values are generally very robust. Considering a cutoff p-value of 10^{-3} the agreement of most group comparisons covers $> 82\%$ of all genes in the strict analysis and $> 93\%$ in the relaxed analysis. The comparison of p-values between normal and early degenerative cartilage shows the least robustness; one reason for this is the small number of significantly regulated genes in this comparison (only 8 genes have a p-value $\leq 10^{-5}$, 38 genes have a p-value $\leq 10^{-3}$), and this also reflects the relatively high similarity of normal and early degenerative cartilage samples.
Overall, this confirms, that the applied methods for normalization and p-value combination yields robust p-values and, thus, the genes selected on the basis of these p-values or the corresponding q-values can be assumed to be appropriate for further biological investigation. An error of about 10% of the significantly differentially regulated genes has to be expected.

Subset Sampling
For estimating the robustness of the most significantly regulated genes for a given group comparison we additionally performed a subset sampling analysis. For each group pair, we generated 50 random subsets of the samples (m=10...18 samples used for each of the groups to be compared) and calculated p-values based on these subsets. Next, we analyzed the top p-value genes; we used the t top genes obtained from the entire sample set as standard of truth and determined the fraction of these top candidates, that are also among the t top candidates of at least $s\%$ of the subset p-value sets. For t we used 50, 75, 100; for s we used 100, 80, 50.
The result for the MAD scale normalized dataset, with the Stouffer method for combining p-values, and for t=50 (t=75 and t=100 yielded very similar results) is shown in figure 7.

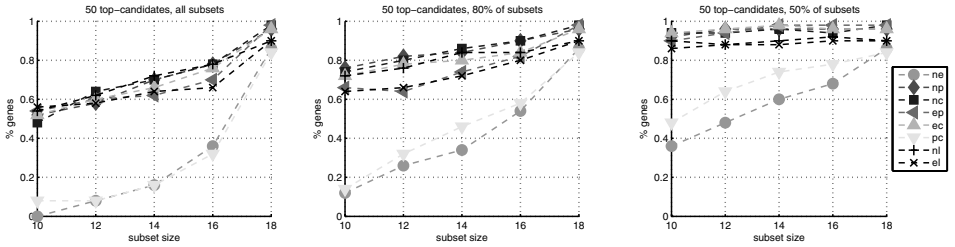

Figure 7: Robustness Analysis of p-value calculation: Subset sampling. Fraction of the 50 top p-value candidates in the overall p-value calculation that are also among the 50 top candidates in at least $s\%$ of the subset-based p-values. Left: all subset p-values ($s = 100$); middle: $s = 80$; right: $s = 50$. For details see section 4.4.

The figure shows that the p-values are of varying stability. Generally, the fraction of stable genes raises when the number of samples in the subset (m) increases. The genes for the group comparisons *ne* and *pc* are significantly less robust than the other comparisons. The other group comparisons show higher stability; for a subset sample size of 10, about 50% of the top-candidates are present in all subset p-value top-candidates; about 90% of the top-candidates are present in half of the subset top-candidates. For these group comparisons, the fraction of stable genes also rises with increasing subset size, but this increase is rather modest compared to *ne* and *pc*. In any case, the analysis yields an overview of the error to be expected within the respective group comparison and the involved differentially regulated genes.

5 Conclusions

The study presented here shows that microarray data normalization and processing has an important effect on the final outcome especially for the identification of differentially expressed genes. It presents the group-level plot as an helpful means for visual inspection of normalization effects on data from classified samples. Furthermore, we compared different methods for combining spot p-values into gene p-values, an important task when dealing with data that bares large inter-spot expression value differences, but neglected so far in our opinion. We found Stouffer's method to work best, which has not been described before for this task.

Finally, we believe that this study shows on exemplary data that it is of vital importance to check every individual step of gene expression data analysis for its appropriateness. Certainly, gene expression data analysis has to fit statistical requirements, but it also needs to account for experimental and biological background knowledge. For most individual processing steps numerous alternatives exist and therefore it is important to test different possibilities and analyze the effects of the decision with appropriate tools. The use of global robustness and quality measures for analyzing individual outcomes can help in estimating the reliability of final microarray study results.

6 Acknowledgement

The authors wish to thank Dr. Eckart Bartnik and Dr. Joachim Saas for helpful discussions. This work is partially funded by projects BEX (Sanofi-Aventis, Frankfurt) and BOA (German ministry for research and education, grant 01GG9824).

References

[ABSZ04] T. Aigner, E. Bartnik, F. Sohler, and R. Zimmer. Functional genomics of osteoarthritis: on the way to evaluate disease hypotheses. *Clin Orthop Relat Res*, 1(427 Suppl):S138–43, 2004.

[ABZZ02] T. Aigner, E. Bartnik, A. Zien, and R. Zimmer. Functional genomics of osteoarthritis. *Pharmacogenomics*, 3(5):635–50, 2002.

[AD03] T. Aigner and J. Dudhia. Genomics of osteoarthritis. *Curr Opin Rheumatol*, 15(5):634–40, 2003.

[BGOT04] K. V. Ballman, D. E. Grill, A. L. Oberg, and T. M. Therneau. Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics*, 20(16):2778–86, 2004.

[BIAS03] B.M. Bolstad, R.A Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[CC03] X. Cui and G. A. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4(4):210, 2003.

[CHQ+05] X. Cui, J. T. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6(1):59–75, 2005.

[CKP+04] H. J. Chung, M. Kim, C. H. Park, J. Kim, and J. H. Kim. ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids Res*, 32(Web Server issue):W460–4, 2004.

[CNGGC04] J. Comander, S. Natarajan, Jr. Gimbrone, M. A., and G. Garcia-Cardena. Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation. *BMC Genomics*, 5(1):17, 2004.

[Edw03] D. Edwards. Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*, 19(7):825–33, 2003.

[FC04] M. Futschik and T. Crompton. Model selection and efficiency testing for normalization of cDNA microarray data. *Genome Biol*, 5(8):R60, 2004.

[Fis32] R. Fisher. *Statistical methods for research workers.* Oliver and Boyd, London, 4th edition edition, 1932.

[HVAS+04] J. Herrero, J. M. Vaquerizas, F. Al-Shahrour, L. Conde, A. Mateos, J. S. Diaz-Uriarte, and J. Dopazo. New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res*, 32(Web Server issue):W485–91, 2004.

[HvHS+02] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002.

[HZZL02] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18 Suppl 1:S145–54, 2002.

[KWSPF03] S. Knudsen, C. Workman, T. Sicheritz-Ponten, and C. Friis. GenePublisher: Automated analysis of DNA microarray data. *Nucleic Acids Res*, 31(13):3471–6, 2003.

[MKH05] S. Michiels, S. Koscielny, and C. Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365(9458):488–92, 2005.

[PGM04] R. Pandey, R. K. Guru, and D. W. Mount. Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics*, 20(13):2156–8, 2004.

[PYK+03] T. Park, S. G. Yi, S. H. Kang, S. Lee, Y. S. Lee, and R. Simon. Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, 4(1):33, 2003.

[Ros84] R. Rosenthal. *Meta-analytic procedures for social sciences*. Beverly Hills, CA: Sage Publications., 1984.

[SS03] G. K. Smyth and T. Speed. Normalization of cDNA microarray data. *Methods*, 31(4):265–73, 2003.

[ST03] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–5, 2003.

[TTC01] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–21, 2001.

[WBHW03] D. L. Wilson, M. J. Buckley, C. A. Helliwell, and I. W. Wilson. New normalization methods for cDNA microarray data. *Bioinformatics*, 19(11):1325–32, 2003.

[YDFQ05] X. Yan, M. Deng, W. K. Fung, and M. Qian. Detecting differentially expressed genes by relative entropy. *J Theor Biol*, 234(3):395–402, 2005.

[YDL+02] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, 2002.

[ZAZL01] A. Zien, T. Aigner, R. Zimmer, and T. Lengauer. Centralization: a new method for the normalization of gene expression data. *Bioinformatics*, 17 Suppl 1:S323–31, 2001.

[ZLS05] Y. Zhao, M. C. Li, and R. Simon. An adaptive method for cDNA microarray normalization. *BMC Bioinformatics*, 6(1):28, 2005.