4 Selecting normalization genes for small diagnostic microarrays

In the last two chapters I showed that a diagnostic chip can be derived in the early onset of a clinical trial. I also pointed out how to derive a diagnostic signature using gene selection for SVM classification. Though, in order to make a diagnostic microarray chip work in practice it has to be normalized. This is needed to bring chips on a comparable scale and account for experimental artifacts. In this chapter, I show that standard microarray normalization methods do not work for diagnostic microarrays. I propose two alternative normalization strategies and evaluate them on simulated and real datasets.

With the concept of diagnostic microarrays new problems arise. A first important step in microarray analysis is normalization. The overall intensity of microarrays can vary. This can reflect global differential gene expression, but it is more likely due to experimental artifacts. Consequently, array-to-array normalization is crucial for microarray analysis (Yang *et al.*, 2002; Kroll and Wölfl, 2002; Smyth and Speed, 2003).

Standard normalization protocols rely on the assumption that the majority of genes on the microarray are not differentially expressed between samples (Yang *et al.*, 2002). For whole genome microarrays this is likely to be true, but on a diagnostic microarray the genes are selected to be differentially expressed between disease entities. Consequently, for diagnostic microarrays a fundamental assumption of microarray normalization does not hold. This has negative effects on the quality of gene expression measurements. Assume that a diagnostic signature consists of 10 genes, all of them higher expressed in disease type A than in type B. Since there are also scale differences due to experimental artifacts, the microarrays need to be normalized. Normalizing them to constant average expression also eliminates the biological differences between A and B. The dilemma is that global differences can be either artifacts or the manifestation of molecular difference between the disease types. Thus, diagnostic microarrays need to be designed in a way that allows for the discrimination of the two different effects.

One way to address the problem is to include additional genes on the microarray that are exclusively used for normalization. Typically, one uses housekeeping genes, which are thought to be expressed at a constant level. However, it has been found that housekeeping genes are occasionally regulated, too (Foss *et al.*, 1998; Schmittgen and Zakrajsek, 2000; Neuvians *et al.*, 2005). One solution is to identify non-regulated housekeeping genes from the set of all housekeeping genes for a given study (Pfaffl *et al.*, 2004). We suggest a data driven approach to select normalization genes from the pool of all genes on the microarray.

Not only the diagnostic signature should be derived from the analysis of a whole genome microarray study but this data is also used for finding normalization genes.

Here, we address the problem of selecting normalization genes from the expression data itself. We compare two simple strategies in the context of simulation experiments as well as in real world applications. The first strategy aims to find control genes that are not influenced by the disease type and can therefore be used for normalization. The second strategy aims to find genes that complement the discriminatory genes on the diagnostic microarray in a way such that normalization on all genes together is not any more influenced by the diseases type. We call this novel concept *balanced signatures*.

The chapter is organized as follows: First we demonstrate the problems occurring when standard normalization protocols are used for small diagnostic chips. In section 4.1 we discuss alternative strategies for normalization gene selection and the concept of balanced signatures. In section 4.5 we compare the novel methods to a standard normalization in the setting of a controlled simulation experiment and in section 4.6 on a dataset from a clinical study on leukemia and on a dataset from a clinical study on lung cancer. We close with a summary and a discussion of our findings.

4.1 Problems of standard normalization methods for diagnostic chips

Standard microarray normalization protocols can not directly be applied to diagnostic microarrays. Ignoring the special character of normalization on diagnostic microarrays leads to a loss of the biological signal. To illustrate this normalization effect on real data, we used a publicly available dataset on acute lymphocytic leukemia (ALL) in children (Yeoh *et al.*, 2002), which we described in detail in chapter 2. We applied a standard normalization protocol where we preprocessed the data using background correction followed by probeset summarization and finally normalization on the summary values. Background correction was done using perfect match (PM) probes only, ignoring mismatch (MM) probes. Probeset summary was done using an additive model fitted by a median polish procedure. Finally, the data was quantile normalized. We used the *RMA* package (Irizarry *et al.*, 2003b) with default parameters to perform all three steps. Note that the probeset summarization step takes logarithms of the data and hence transforms expression levels to an additive scale. Here, fold changes of molecule abundance correspond to differences in the normalized data.

We designed a virtual diagnostic microarray for discriminating between patients displaying a TEL-AML translocation (group A) and those displaying either a BCR-ABL or a E2A-PBX1 translocation (group B). To this end, we chose the 10 genes with the highest expression differences

$$\Delta_i = \sum_{j \in J_A} \frac{x_{ij}}{|J_A|} - \sum_{j \in J_B} \frac{x_{ij}}{|J_B|}$$

where J_A, J_B are the set of samples in group A and B, respectively. x_{ij} is the normalized gene expression intensity of gene i in sample j. Then we went back to the non-normalized raw data of only these 10 genes and discarded all other expression data. Using only the remaining raw data of the 10 genes we repeated the same normalization steps that were used for the large Affymetrix microarray. This yielded a virtual diagnostic microarray of 10 genes. Since normalization was not done on an array-by-array, nor on a gene-by-gene basis, but borrowed information across both genes and microarrays the results of the two normalizations were different although the underlying raw data was identical.

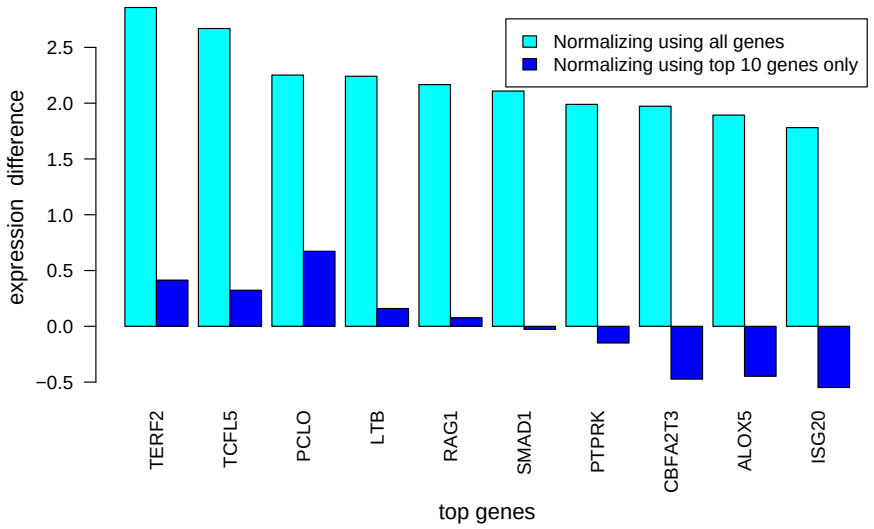

Figure 4.1: The global signal normalization effect resulting from standard normalization protocols: Shown are changes of expression difference, when switching from a whole genome microarray to a small diagnostic microarray chip. The top genes are those genes with the maximal expression difference between TEL-AML versus BCR-ABL and E2A-PBX1. Note, that expression differences on log scale reflect fold changes.

When switching from the whole genome microarray to the small diagnostic array the expression differences between the two cytogenetically different groups of patients vanished almost completely. Normalization of the small array has destroyed the original signal that is needed for diagnosis (Fig. 4.1). We refer to this effect as the *global signal normalization effect*. Not only did the expression differences vanish, but the average correlation between the genes also changed from 0.73 to -0.1.

Algorithms

We showed that standard normalization applied to diagnostic microarrays can substantially skew results and is a problem for diagnosis. In the following section we propose two different strategies to circumvent these problems. The first strategy aims at finding genes that can be used solely for normalization. Several methods for finding normalization genes are suggested and compared. The second strategy aims at finding genes that can be used for normalization and additionally also for classification.

4.2 Selection of normalization genes

We have argued that a microarray carrying only differentially expressed genes can hardly be used to distinguish biological effects from experimental artifacts. To overcome the problem we suggest to include additional normalization genes on a diagnostic microarray that are then used to adjust for experimental artifacts but leave the biological signal intact. Like the signature genes, the normalization genes can be selected based on the data from a genomewide expression study. While signature genes should correlate with the disease labels of patients, the normalization genes should not.

For the signature genes it is most important that the correlation of expression levels to the disease labels does not only hold for the training data on which the genes were found but generalizes to new samples. In the same way the desired properties of normalization genes also need to generalize to new data. Hence, criteria for normalization need to be chosen such that they enable both, a good normalization of diagnostic microarrays and at the same time generalize well to new samples. Note that these two requirements do not implicate each other.

Let p_s be the number of genes that form the diagnostic signature. In experimental settings p_s was in the range of 5-50 genes (Li and Yang, 2002; Bø and Jonassen, 2002; Li, 2005). Let p_n be the number of additional genes used on the microarray for array-to-array normalization. The total number of genes on the diagnostic microarray is thus $p_d = p_s + p_n$. Both the signature genes and the normalization genes are selected based on genomewide microarray data measured with whole genome microarrays holding $p_l \gg p_d$ genes. In this context x_{ij} denotes the expression of gene i in patient j. As we aim at diagnostic differentiation into groups we can assume without loss of generality that the samples fall into two different disease entities represented by class labels A and B. If there should be more classes, it is always possible to construct a binary classification tree where the first group is compared to all others. Then the second group is compared to the rest excluding the first group and so on.

The open question is how to select normalization genes. We propose two novel, completely data driven methods for normalization. The first method selects genes solely used for normalization according to criteria listed below. The second method aims at balancing the signature and is described in section 4.3.

1. **Low variance genes:**
   Calculate the empirical variance σ_i^2 of all p_l genes and choose the p_n genes with the smallest variance in the data. Use only these genes for array-to-array normalization. In our preprocessing protocol the background correction and probeset summarization remain unchanged but only these p_n genes are used for the final normalization step.
   In this approach, we aim for the genes with the most constant expression in both disease populations. Population variances are not known and we select the genes

due to their variances on the expression data of the genomewide study. This idea is similar to the use of housekeeping genes, whose expression is assumed to hardly vary between patients. Observed differences in measurements are hence most likely due to experimental artifacts. However, we do not select housekeeping genes based on a priori knowledge, but from the data at hand.

2. **Small coefficient of variation:**
   Calculate the empirical variance σ_i^2 and the empirical mean μ_i of all p_l genes and choose the p_n genes with the smallest coefficient of variation $\frac{\sigma_i}{\mu_i}$ in the data. Use only these p_n genes for array-to-array normalization.
   In this approach, we aim for the genes with low variance that additionally have high intensity. The idea is to exclude low variance genes within the background noise.

3. **Small differences of average expression:**
   Calculate the differences $\Delta_i = \sum_{j \in J_A} x_{ij}/|J_A| - \sum_{j \in J_B} x_{ij}/|J_B|$ between the two groups for all p_l genes and choose the p_n genes with the smallest absolute Δ_i. Use only these genes for array-to-array normalization.
   In this approach we allow the genes to vary between patients but this variability should not correlate with the disease type. Note that the genes are typically not constant and therefore not housekeeping genes. Still they allow for normalization if the property of small expression differences generalizes well to the diagnostic microarray.

As a control we use randomly sampled genes for normalization. Here of course we have no problem with generalization. One might expect, that the above methods are more effective, but this needs to be proven.

For the evaluation of the real datasets we included the normalization results obtained when using standard housekeeping genes. For this, we used the following 3' variants of the housekeeping probesets supplied on Affymetrix GeneChips: beta-actin, GAPDH, ISGF3, 18S rRNA, transferrin receptor and 28S rRNA.

4.3 Selection of a balanced signatures

This approach does not use different genes for normalization and diagnosis, but tries to find a set of genes, which serves both tasks at the same time. Starting from a non balanced set of signature genes, choose p_n genes from all p_l genes such that the variation of the average gene expression per microarray is minimized

$$\sum_{j \in J} (x_{.j} - x_{..})^2 \to min \quad \Longrightarrow \quad \sum_{j \in J} \left(\sum_{i \in I_d} \frac{x_{ij}}{|I_d|} - \sum_{i \in I_d} \sum_{j \in J} \frac{x_{ij}}{|I_d| * |J|} \right)^2 \to min \quad \Longrightarrow$$

$$\sum_{j \in J} \left(\sum_{i \in I_d} \left(x_{ij} - \sum_{j \in J} \frac{x_{ij}}{|J|} \right) \right)^2 \rightarrow min$$

where $x_{.j}$ denotes the average expression of genes on the diagnostic array j, J is the set of all samples, I_d is the set of all genes on the diagnostic microarray, and $x_{..}$ the average gene expression over all diagnostic microarrays. This is done using a greedy forward selection, which is summarized in figure (4.2). In contrast to the methods above, the normalization is now done using both signature and normalization genes. The strategy here is not to find genes that are not affected by expression difference between the two disease groups, but genes that compensate this effect. For example, if the signature genes are all up-regulated in group A, the goal is to compensate for this effect by choosing genes which are down regulated. This method does not distinguish between the discriminating genes and the genes for normalization any more. The normalization genes are now themselves differentially expressed and can hence be included into the signature.

Greedy forward selection:

Let: $J = J_A \cup J_B$, be all samples in group A and B, $|J|$ is the number of all samples
I_l, be the set of all genes on the whole genome microarray
I_s, be the set of given genes of the diagnostic signature
$I_n = \{\}$, be the initially empty set of normalization genes

for $k = 1..p_n$ (for each normalization gene)
$I_d = I_s \cup I_n$
for $g \in I_l \setminus I_d$ (for each gene not yet used on the diagnostic microarray)
calculate $v_g = \sum_{j \in J} \left(\sum_{i \in I_d \cup g} \left(x_{ij} - \frac{\sum_{j \in J} x_{ij}}{|J|} \right) \right)^2$
$I_n = I_n \cup \text{argmin}_g v_g$

Figure 4.2: Pseudo code for greedy forward selection of balancing genes

In the absence of experimental artifacts the summed up expression levels for each sample should be constant. In this way, these genes allow us to distinguish between differential expression and experimental artifacts. Similar to the first two methods, there is again a generalization problem. We balance the signature on the training set. Its normalization performance for the diagnostic microarray however depends on how well the balance between up- and down-regulated genes generalizes to new data.

4.4 Normalization of small diagnostic microarrays

Normalization of small diagnostic microarrays was done by subtracting the sample wise mean of the normalization genes from all genes. Let x_{ij} be the expression of gene i in patient j. Let I_n be the set of normalization genes, and $p_n = |I_n|$ the number of normalization genes. For all normalization genes the sample wise mean ν_j was calculated: $\nu_j = \sum_{i \in I_n} \frac{x_{ij}}{p_n}$. Normalization was then done by subtracting ν_j from all genes resulting

in normalized data y_{ij}: $y_{ij} = x_{ij} - \nu_j$. For the balanced signature I_n included all genes and therefore $\nu_j = x_{.j}$

4.5 Results on simulated data

The two normalization methods for diagnostic microarrays described in the previous section need to be evaluated with respect to their power in compensating the global signal normalization effect and producing diagnostic arrays that distinguish well between two disease entities. Before we evaluate our methods on real data in the next section we make use of the more transparent setting of a simulation study, in which the population differences, the biological variability among individuals and the experimental variability are modeled independently of each other.

Simulated data was generated according to a multivariate normal distribution, including strong correlation of genes, a large spectrum of expression intensities and non constant expression differences between the two groups A and B.

In total we simulated expression values for 3000 genes on 50 microarrays representing two groups A and B of 25 microarrays each. We first generated the covariance matrix Σ by randomly drawing from an inverse Wishart distribution with 3150 degrees of freedom and a 3000x3000 identity matrix as a scale matrix. Then, we generated a vector of 3000 population means for each group μ_i^A, μ_i^B by independently drawing from a $N(0,1)$ normal distribution. The actual expression data was generated by drawing from a multivariate normal distribution with covariance matrix Σ and means μ_A for the first 25 microarrays and μ_B for the next 25 microarrays. Finally, this data was perturbed by multiplying with a random scaling factor and adding a random offset both drawn from a N(0,0.3) normal distribution. The generation of the data was done twice. Once for a training set and once for a test set.

In this simulation with three successive randomization generating μ^A, μ^B and Σ corresponds to the population properties of the genes. Drawing from a multivariate $N(\mu^{A,B}, \Sigma)$ distribution accounts for biological variability among individuals, while the perturbing the data accounts for global experimental artifacts. The differences Δ_i display the typical continuous spectrum known from real expression data (figure 4.3).

As we have stressed before, the expression patterns of the normalization genes need to generalize from the training set to new data in the same way as the signature patterns do. From the theoretical considerations of the previous section it becomes clear that small variance genes have the potential to compensate for the global signal normalization effect. But the genes need to have small variances not only on the training data but also and more importantly on the data that is generated using the diagnostic array. In general, the variance will be higher than it is on the training data. The same problem occurs for genes with small average expression differences and balanced signatures. To

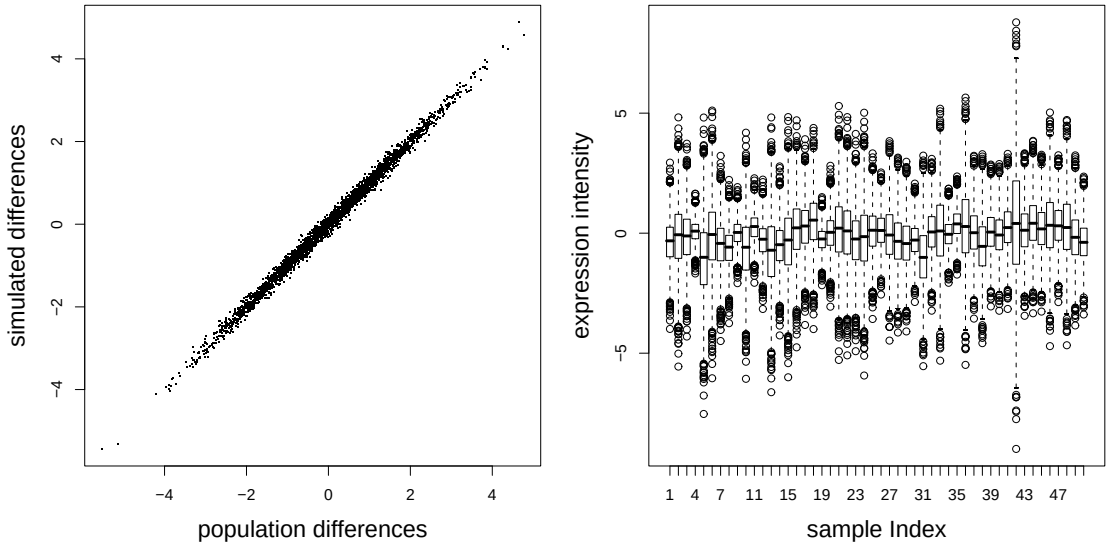

Figure 4.3: The left plot shows the genewise population differences contrasted with the mean differences in simulated data. Population differences $\mu_i^A - \mu_i^B$ were set for each gene by randomly drawing from N(0,1). Simulated differences stem from drawing data from a multivariate distribution with these given population means. The right plot shows boxplots of all 3000 genes for all 50 samples of the simulated data.

this end, we simulated a training and a test set with 50 samples. Both sets have the same underlying gene means and covariance structure. To avoid overfitting, only the training data was used to select the normalization genes and only the test set was used to evaluate the normalization strategies. The diagnostic signature consists of $p_s = 10$ genes with the largest difference of population means. It is unbalanced. For the purpose of normalization $p_n = 10$ additional genes were picked according to the suggested methods.

Using the standard normalization protocol destroys the signal completely, while using random normalization genes already recovers the signal partially (left plot in figure 4.4). However, both versions, data based selection of normalization genes and balanced signatures, recover population differences more accurately and perform similarly to each other (right plot in figure 4.4).

We repeated the data simulation 30 times and recorded for each simulation the distance between the real underlying expression differences of the signature genes and the expression differences obtained by the various normalization methods. This sum of squared error plot shows that balanced signatures perform slightly better than the other methods (figure 4.5).

4.6 Results on a leukemia study

We now proceed from a simulation study to applications on real datasets. Of course, in real datasets we do not know how many genes are deregulated and how many are nec-

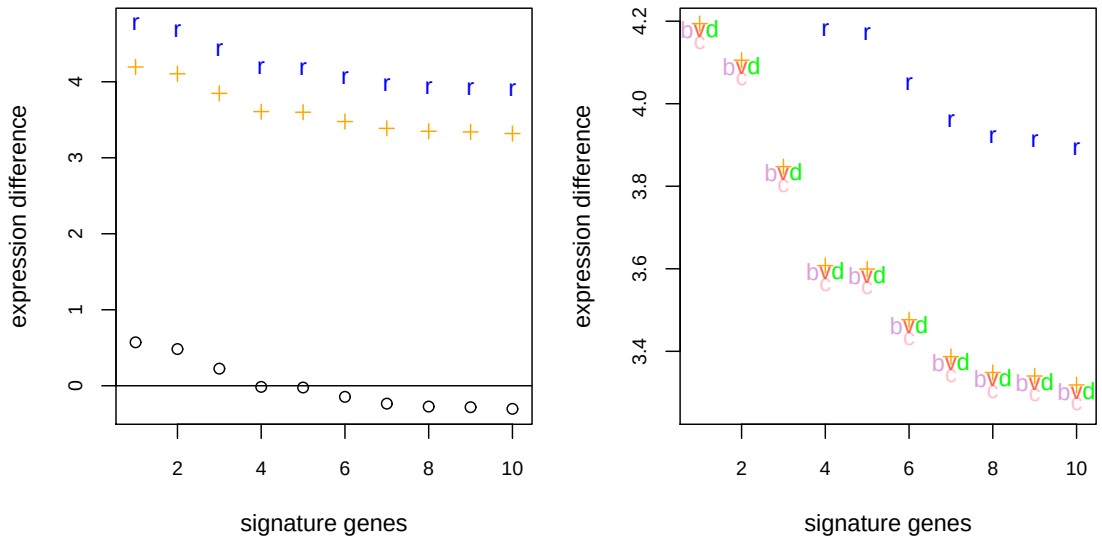

Figure 4.4: Effects of different normalization methods for diagnostic microarrays evaluated on simulated data. "+" depicts expression differences in the test data of the signature genes after normalization with all 3000 genes. This, we would like to recover with normalization methods for diagnostic microarrays, too."o" corresponds to using the standard protocol on the diagnostic microarray. Here, all the signal is lost. "r" corresponds to a normalization of the diagnostic microarray with 10 random genes. It already recovers the signal partially. The right plot is a closeup of the left plot, showing additionally the performance of the proposed normalization schemes. "+" and "r" are the same as in the left plot. Additionally, normalization using lowest variance "v", smallest difference "d", smallest coefficient of variation "c" and balanced signatures "b" are shown. For better visibility the symbols "b" and "d" are slightly moved to the side so that they do not overlap.

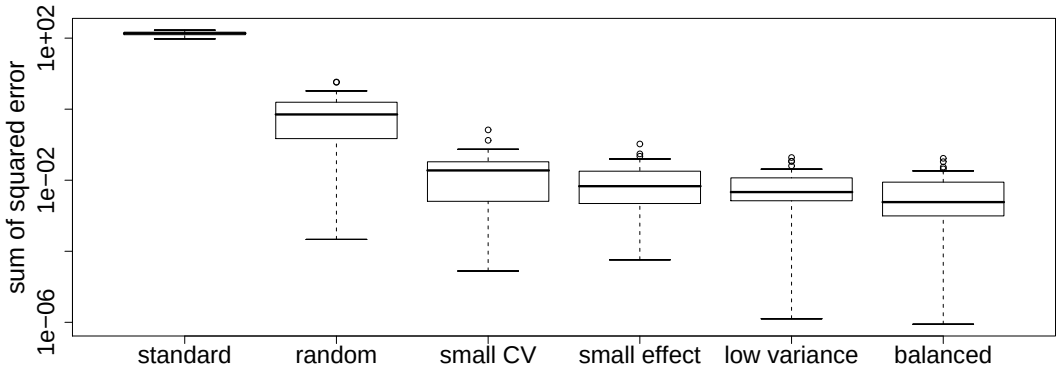

Figure 4.5: Sum of squared errors to the real underlying expression differences of the proposed normalization methods and the standard protocol averaged over 30 runs of the simulated data. "small CV" depicts the normalization method using smallest coefficient of variations and "small effect" depicts the normalization method using small differences of average expression.

essary for achieving optimal classification accuracy. Therefore, we ran the *MCRestimate* package (Ruschhaupt *et al.*, 2004), that uses a nested cross validation loop to avoid biased estimators of classification performance. Our own results analyzing various datasets with *MCRestimate* showed that most datasets can be classified optimally with a 2-50 genes and only very few need more than 50. This is in concordance with findings from other authors (Li and Yang, 2002; Bø and Jonassen, 2002; Li, 2005). When applying it to the leukemia study (Yeoh *et al.*, 2002), described in section 4.1, we found that in this case $p_s = 5$ genes reached the optimal classification accuracy of 99%. Thus, we selected $p_s = 5$ signature genes with the highest absolute equal variance t-score. In addition, $p_n = 5$ normalization genes were determined according to the criteria from the previous sections. For simplicity, the number p_n of additional genes for normalization was set to p_s. In preliminary studies this provided good results but further research on determining the optimal p_s and p_n simultaneously is needed.

The second dataset we analyzed was a study on 86 primary lung adenocarcinoma and 10 normal lung tissues (Beer *et al.*, 2002). Here, we aimed for a classification of normal versus carcinoma. *MCRestimate* achieved 100% accuracy using 3 genes. Thus, we selected $p_s = p_n = 3$ for this dataset.

We randomly split the whole datasets equally into a training and test set. For the training set we applied the gold standard normalization using all genes of the whole genome microarray. Then, we proceeded in the same way as described in the previous sections. Both, signature and normalization genes were derived using only the training data. For each sample in the test set a virtual diagnostic microarray was constructed using only the raw data of the signature and the normalization genes. This virtual diagnostic microarray was normalized using the methods described in section 4.2 and 4.3, resulting in seven different test datasets: standard protocol, Affymetrix housekeeping genes, random normalization genes, low variances, small coefficient of variation, small differences and balanced signatures. On the such normalized test set we evaluated the normalization methods with respect to the diagnostic performance of a support vector machine using cross validation. For this, we used the SVM from the package *e1071* in R (Ihaka and Gentleman, 1996) with a linear kernel and default parameters. The dataset was randomly split in equally sized training and test sets. This was repeated 100 times and the evaluation steps were rerun for every data partitioning (figure 4.6).

The standard protocol reduces the classification accuracy substantially, while both normalization gene selection and balanced signatures yield satisfying results. Affymetrix housekeeping genes for normalization work well on the leukemia dataset, but fail on the lung dataset. Balanced signatures provide the best results in both datasets.

For the leukemia dataset classification accuracy was significantly better for all our methods as compared to the standard protocol ($p < 10^{-15}$). "Balanced normalization" outperformed all other normalizations ($p < 10^{-8}$), too. Standard normalization was also clearly inferior in the lung dataset ($p < 10^{-14}$). When further testing "balanced normalization" against other normalizations p-values were below 0.001 for all but "small effect normal-

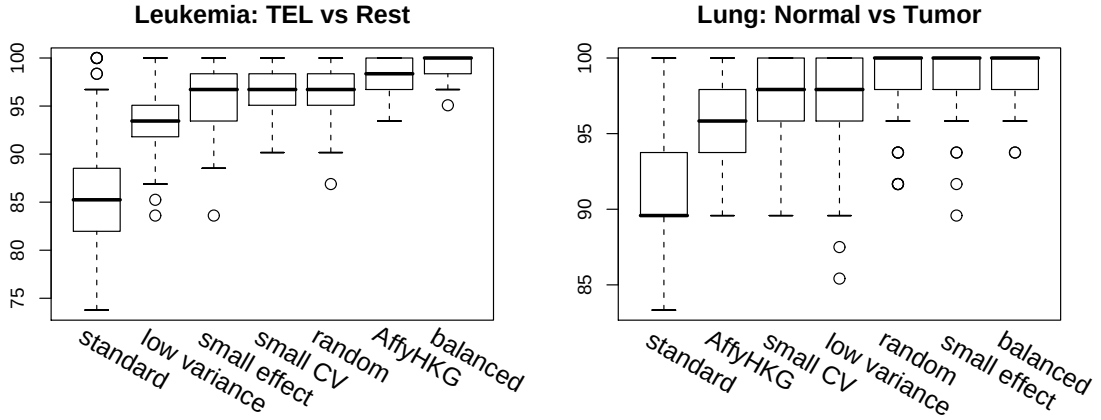

Figure 4.6: Cross validation results of predictive performance of the same diagnostic signature used with different normalization strategies for diagnostic microarrays. The left plot shows classification accuracies for distinguishing TEL-AML1 from other groups in leukemia ($p_s = p_n = 5$). The right plot shows classification accuracies for distinguishing normal from adenocarcinomas in lung ($p_s = p_n = 3$).

ization" and "random normalization", where significance was not reached ($p = 0.13$ and $p = 0.15$ respectively).

4.7 Discussion

In this chapter I addressed the problem of normalizing diagnostic microarrays. I showed that using a standard normalization protocol from large microarrays has fatal effects. They are most pronounced when the diagnostic signature is unbalanced, containing more up- than down-regulated genes or vice versa. However, in most microarray datasets there are more significantly up- than down-regulated genes or vice versa, emphasizing the need for new normalization strategies. Here, I introduced two strategies to overcome this problem: data driven normalization gene selection and balanced signatures. Both gave better results for diagnostic microarrays than the standard normalization protocol. Using Affymetrix housekeeping genes performs well in the analyzed leukemia dataset but does not work for the lung dataset, indicating that the genes are actively regulated in these tissues.

As standard normalization protocol I have chosen the RMA procedure. Of course it is not the only protocol in use. However, the global signal normalization effect is generic and not restricted to this protocol. Any normalization which assumes unchanged expression for the majority of genes on the microarray is expected to suffer from the same problem. An advantage of both methods is that the normalization genes can be selected with no additional experimental cost and little computational effort.

Hua et al. stressed that optimal feature size depends strongly on the classifier and feature-label distribution and that a choice of optimal feature size can greatly improve accuracy of the classification (Hua *et al.*, 2005). Hence, for assessing how many genes should be used

for a diagnostic microarray I used a nested cross validation for SVMs (Ruschhaupt *et al.*, 2004). By this, I determined the number of genes making up the diagnostic signature (p_s) and set it to the number of genes needed for achieving the optimal classification accuracy.

In conclusion, balanced signatures perform well with respect to recovering the real underlying signal as well as for classification. This was verified on a simulated test dataset as well as on two real microarray datasets.