4 Exemplary data sets

4.1 Outline

Throughout the thesis we use a set of six microarray comparisons to illustrate the reviewed methods as well as the novel approaches. In Section 4.2 we introduce the underlying microarray studies and show the effects of preprocessing. In Section 4.3 we further explore the data by applying the (positive) false discovery rate adjustments as introduced in Section 3.2.

4.2 Six microarray comparisons on cancer

We introduce six data sets derived in four microarray studies exploring various kinds of cancer or different clinically relevant outcomes of cancer patients. Each data set is a comparison of two clinical classes. The data sets will serve for illustrating the methods discussed in this thesis. All experiments were carried out on Affymetrix GeneChip® HGU95Av2 arrays coding for 12625 genes. For details on this array technology we refer to Sections 2.2 and 2.3.

ALL 1 The study of Yeoh *et al.* (2002) consists of 327 patients in total. The patients suffered from pediatric acute lymphoblastic leukemia (ALL), a frequent blood-cancer type. The samples divided into cytogenetically distinct subgroups that most often were characterized by a certain chromosomal aberration. We compared the 27 samples with fusion protein E2A-PBX1 to the 18 normal samples. The latter patients suffered from ALL but did not show any of the tested chromosomal aberrations.

ALL 2 From the study on ALL above we chose a second comparison. A subset of 43 out of 327 patients had T-lineage ALL and was followed up for long-

term risk of relapse. For 37 of these patients an event was observed: 26 patients remained in complete remission while 11 patients suffered from blood-cancer relapse. We compared these two groups as one expects different gene activity for relapse than for non-relapse patients.

Breast-cancer 1 The second microarray study comprised a total of 89 patients with breast cancer (Huang *et al.*, 2003). The samples were characterized by their progression status (remission or relapse) and the risk of relapse depending on whether the lymph nodes were affected or not. Here we compared the 18 samples with high risk of relapse to the 19 low-risk samples.

Breast-cancer 2 The fourth comparison consists of the 18 patients with relapse against the 34 patients remaining relapse-free in the breast-cancer study above.

Lung-cancer Bhattacharjee *et al.* (2001) conducted a study on classification of 186 lung-carcinoma patients that divided into several histologically defined subtypes. Besides, a set of 17 samples from normal lung tissues was included. We compared the 21 patients suffering from squamous cell lung carcinomas to the normal samples.

Prostate-cancer The last comparison was taken from a study on prostate-cancer patients by Singh *et al.* (2002). Similar to the lung-cancer study above, the data set included prostate samples from disease-free patients. We compared the 52 patients with prostate cancer to the 50 normal samples.

We applied the preprocessing as described in Section 2.3 separately to the four microarray studies. In Figure 4.1 we illustrate the normalization effect. Shown are kernel density estimates of expression values per sample. On the left-hand side we display the original values without normalization. Only the last summarization step was applied to yield comparable scales. The densities on the right-hand side are based on the fully preprocessed data. Within each study, the values are distributed consistently on the same scale. The Breast-cancer, Lung-cancer and Prostate-cancer samples show wide intensity distributions, the ALL samples are closely distributed around a mean value of 8.

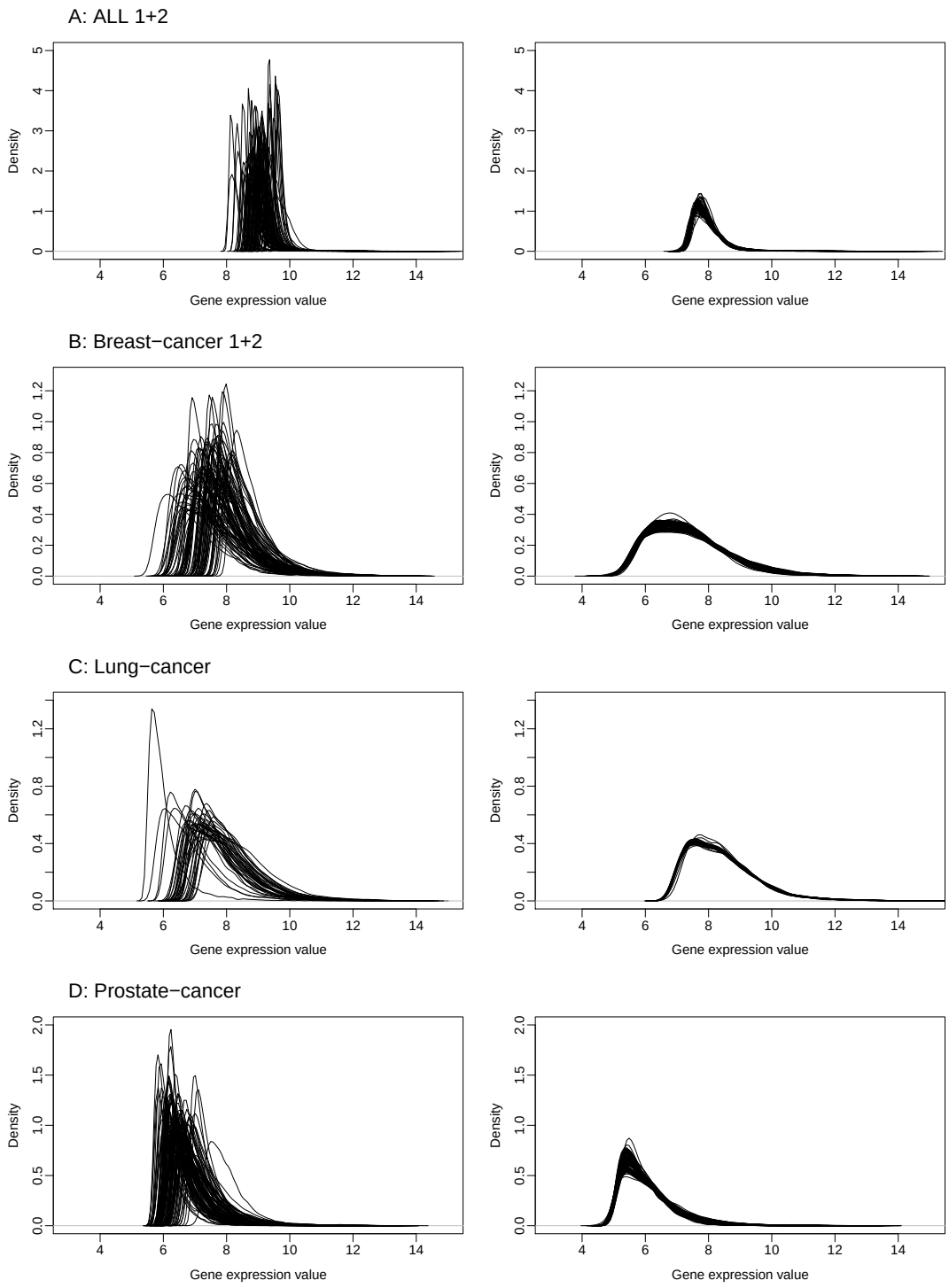

Figure 4.1: Distribution of preprocessed expression values in the exemplary data sets. For each sample, we show kernel density estimates based on data without normalization (left-hand side) and with normalization (right-hand side). Each curve represents the distribution of expression values taken from one microarray. After preprocessing, the values are distributed consistently within each study.

4.3 Exploring differential expression

After preprocessing, the ALL and Breast-cancer samples were divided into the four comparisons introduced above (ALL 1, ALL 2, Breast-cancer 1 and Breast-cancer 2). According to the methods described in Section 2.4, we derived scores and p-values for the six comparisons. To this end, we computed z-scores as defined in Equation (2.10). For each comparison, the fudge factor was set to the median of the respective pooled standard deviations per genes. The resulting z-scores were transformed into pooled empirical p-values based on $B = 1000$ class-label permutations prior to applying Equations (2.13) and (2.14). A first impression on the amount of differential expression is given in Figure 4.2. Shown are the resulting p-value distributions for each comparison. Note that the histograms are plotted on the same scale for enhanced comparison. A strong over-representation of small p-values relates to a high amount of induced genes. The plots exhibit substantial differences between data sets. The six comparisons were chosen to illustrate diverse outcomes of the statistical analysis of gene expression data: the Lung-cancer and Prostate-cancer data sets both compare cancer patients to disease-free patients. In both cases we observe strong over-representation of small p-values and thus many possibly induced genes. We conclude that cancer substantially changes the gene activity in affected cells when compared to the normal state of the same tissue.

The situation changes when comparing diseased patients only as was done in the ALL and Breast-cancer studies. The second comparisons of both studies evaluated differential expression of relapses. The Breast-cancer 2 set exhibits substantial differences in expression whereas the p-value distribution of the ALL 2 set is close to uniform. The p-value distribution of the first Breast-cancer comparison appears to be even more uniform. We conclude that the lymph-node status reflecting risk of relapse does not change the gene expression substantially. Finally, we observe a moderate over-representation of small p-values for the ALL 1 data set, where two cancer subtypes were compared to each other. From this observation we conclude that different cytogenetical states induce different genes.

We proceed with the significance analysis by applying the two p-value filter strategies introduced in Section 3.2, that is the false discovery rate of Benjamini and Hochberg (1995) and the positive false discovery rate of Storey (2003). To this

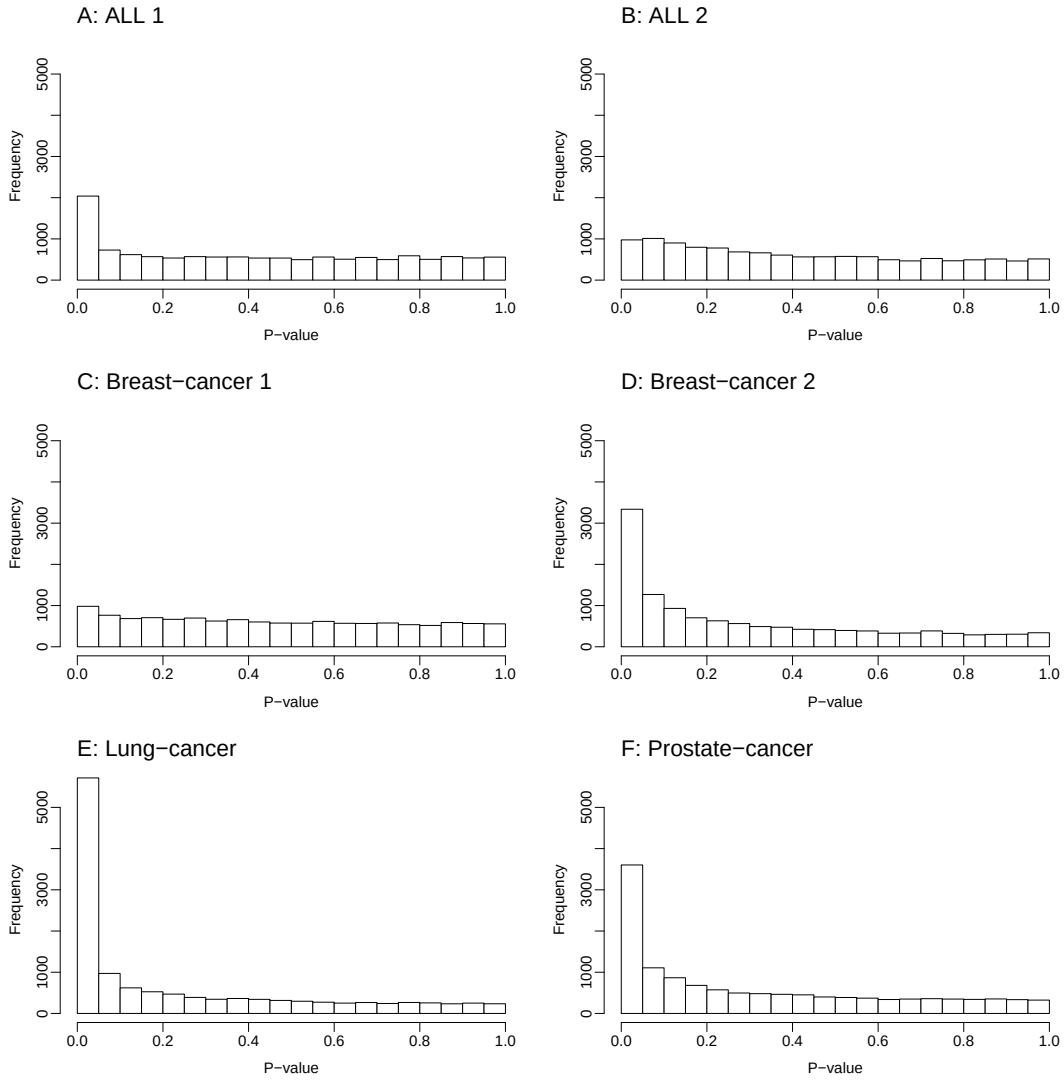

Figure 4.2: Assessing the amount of induced genes. Shown are histograms of empirical p-values for the different comparisons. A high over-representation of small p-values might lead to a longer list of induced genes.

Table 4.1: Application of false discovery rate filters. Shown are the numbers of genes with FDR-adjusted p-values or q-values below the desired level. The last column contains the estimated value of the percentage of non-induced genes π_0.

Comparison	Size of gene list with FDR ≤ 0.05	Size of gene list with pFDR ≤ 0.05	$\widehat{\pi_0}$
ALL 1	876	928	0.8706
ALL 2	0	0	0.7678
Breast-cancer 1	0	3	0.8817
Breast-cancer 2	749	1857	0.4851
Lung-cancer	4615	6048	0.3738
Prostate-cancer	1572	2364	0.5190

end, we computed FDR-adjusted p-values as defined in Equation (3.5) using package *multtest* by K.S. Pollard, Y. Ge and S. Dudoit. Q-values as defined in Equation (3.12) together with an estimate of the percentage of non-induced genes π_0 were derived using package *qvalue* by A. Dabney and J.D. Storey. In both cases, the p-value filters were set such that the corresponding false discovery rates did not exceed 5%. The numbers of genes with p-values passing the filters are shown in Table 4.1 along with Storey's estimate of π_0. The results reflect our first conclusions drawn from the p-value histograms in Figure 4.2. The ALL 2 and the Breast-cancer 1 comparisons merely lead to no or only a few genes identified as differentially expressed. For the remaining four comparisons, up to 6000 genes are in the lists of induced genes. The lists are larger when the p-value filter was set with respect to the positive false discovery rate. This effect is triggered by the estimated percentage $\widehat{\pi_0}$. If $\widehat{\pi_0} = 1$, the two sizes are equal. Smaller values of $\widehat{\pi_0}$ result in large differences between the FDR- and the pFDR-lists. However, the value of $\widehat{\pi_0}$ alone does not correlate with the list size. For example, we observe nearly equal estimates for ALL 1 and Breast-cancer 1, yet the sizes differ substantially. Recall from Figure 3.2 that the estimation of the percentage of non-induced genes relates to the horizontal separation of the complete p-value histogram. Going back to the histograms shown in Panels A and C of Figure 4.2 we observe similar heights of the uniform histogram parts. Only the p-value abundance of the first bar in Panel A is spread over the entire histogram in Panel C, which might explain the slightly higher π_0 estimate for Breast-cancer 1.

Besides the raw numbers for one choice of positive false discovery rate, we might explore how the list sizes increase with increasing thresholds. To this end, we show estimated positive false discovery rates with associated list sizes in Figure 4.3. Col-

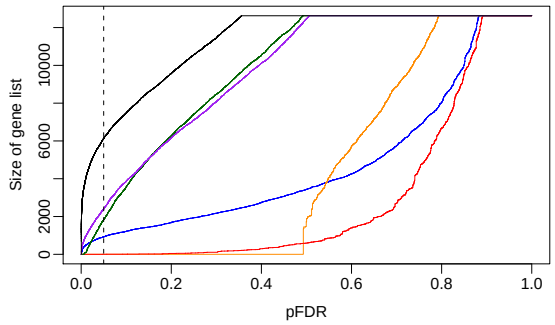

Figure 4.3: Application of the positive false discovery rate. For each comparison, q-values were computed from the set of empirical p-values. Shown are the resulting positive false discovery rate thresholds and the corresponding sizes of lists including genes with lower or equal q-values. The vertical line marks the exemplary choice of pFDR ≤ 0.05, see Table 4.1. Colors correspond to comparisons: ALL 1 (blue), ALL 2 (orange), Breast-cancer 1 (red), Breast-cancer 2 (green), Lung-cancer (black) and Prostate-cancer (purple).

ors correspond to the different comparisons. The rate of size increase reflects the over-representation of small p-values in Figure 4.2: even for small choices of the positive false discovery rate, many genes are identified as differentially expressed in the Lung-cancer comparison. The curves of Breast-cancer 2 and Prostate-cancer increase slower and are almost identical. On the other hand, no genes will pass the p-value filter in the ALL 2 comparison without choosing the positive false discovery rate to be at least 0.5.

We do not apply one of the local false discovery rate estimation methods introduced in Sections 3.4 and 3.5 here but will introduce a novel estimator in the following chapter. The results will be shown there.

4.4 Comparison of pooled and gene-wise p-values

In the end of Chapter 2 we discussed differences between two possible ways to compute p-values from permutation scores. In this section, we will investigate the difference between pooled and gene-wise p-values in real data. To this end, we computed gene-wise p-values as defined in Equation (2.15) for the exemplary ALL 1 comparison. We used a fixed set of $B = 10000$ permutations plus the observed one and computed the p-value of the ith gene by comparing the observed z-score s_{i0} to the set of permutation scores $(s_{ib})_{b=0,\ldots,B}$. Since we used the z-score,

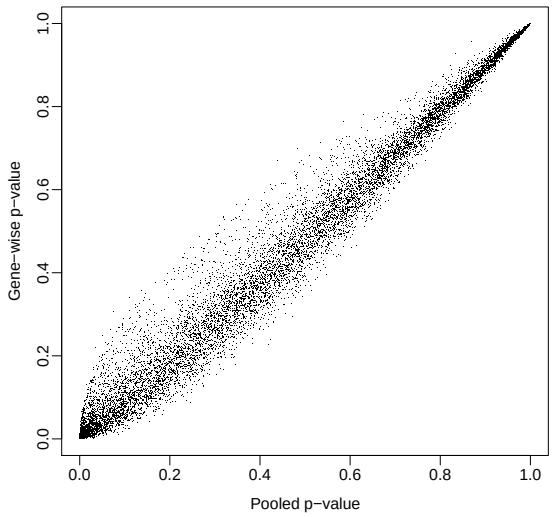

Figure 4.4: Pooled versus gene-wise p-values for the ALL 1 data set. The two computation methods lead to different p-values and thus to a different ranking of genes.

which does not scale properly for different variances, gene-wise p-values differ substantially from pooled p-values, see Figure 4.4.

We further analyzed a set of 200 genes with largest differences between gene-wise and pooled p-values. The results are shown in Figure 4.5. The first two panels on the left-hand side display the two main sets of genes with either gene-wise score distributions being narrower than the pooled score distribution (first row) or gene-wise score distributions being wider than the pooled score distribution (second row). The dashed line in the left-hand side plots denotes the pooled score distribution that was used to compute pooled p-values. In the first case, the pooled p-values are always larger than the gene-wise ones. In the second case, the gene-wise p-values are larger. Both cases might be explained by examining the underlying expression values. Indeed, the genes within each set showed similar behavior. On the left-hand side we display the gene expression values per patient for two typical members of the two gene sets. In the first case, the expression values are almost constant. Thus, even with regularized z-score, the small variances lead to narrow score distributions and the gene-wise p-value of the observed score is smaller than the p-value computed from the pooled null distribution. For genes of the second set, we commonly observed larger variances often accompanied by an outlying expression value. The outlier causes high absolute scores and thus score distributions with heavy tails. In the last two rows of Figure 4.5 we show two ex-

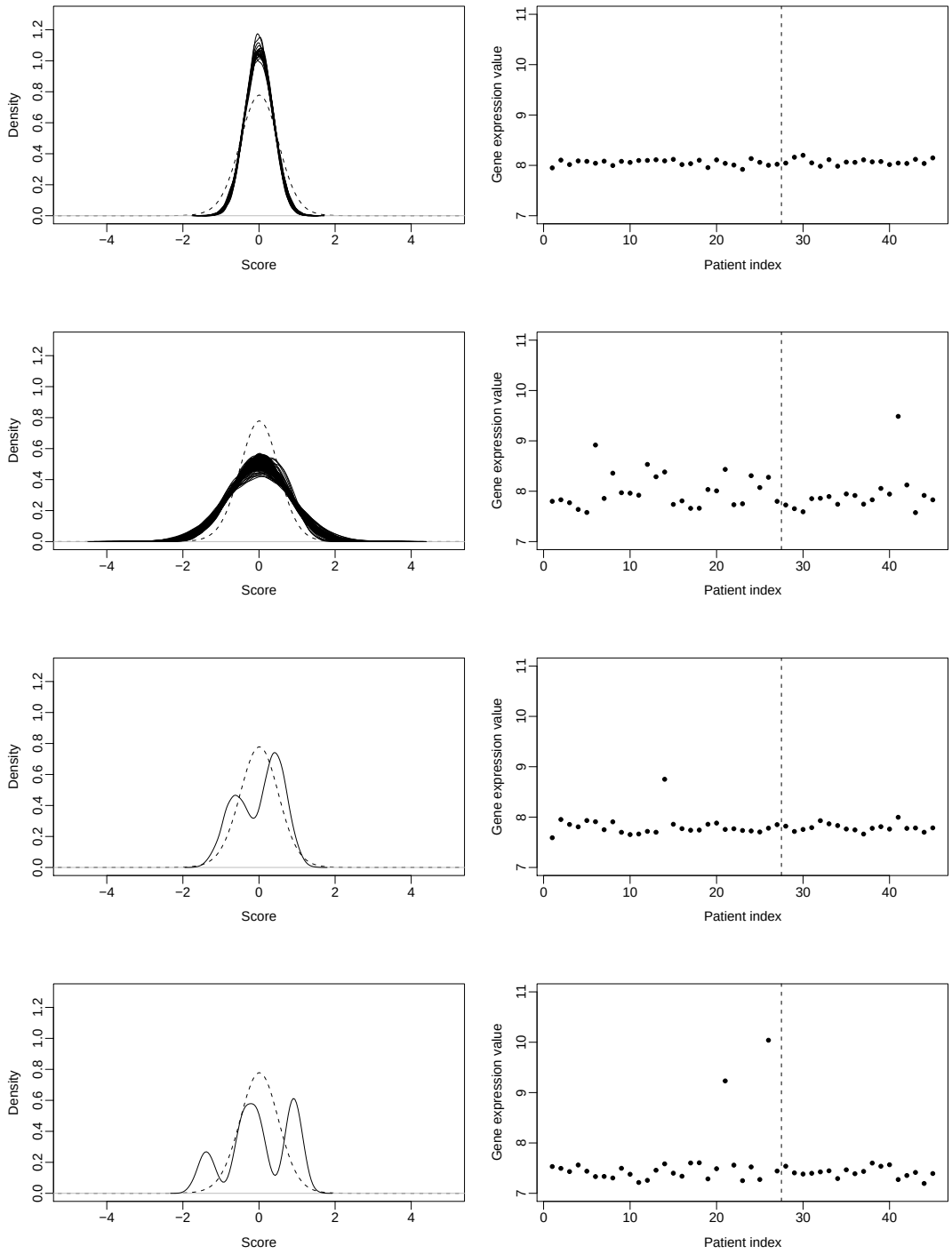

Figure 4.5: Close examination of the first 200 genes with largest differences between gene-wise and pooled p-values in the ALL 1 comparison. Left-hand side: kernel density estimates of gene-wise score distributions. Dashed curve denotes the pooled score distribution. On the right-hand side we display respective gene expression values. For the first two plots, only one exemplary gene was chosen. The dashed line denotes the boundary between E2A-PBX1 patients (left) and normal patients (right).

treme cases out of the set of 200 genes. Here, for two genes with small variances we observed one or even two outlying expression values. The influence of the outliers is strong enough to cause multimodal score distributions. In case of two outliers, we observe three modes: if both values were shuffled into the same patient group, we get either high positive or high negative scores. If the values were assigned to different patient groups, the effects cancel out and the scores contribute to the central bump.

We further conducted a simulation to investigate influences on the false discovery rate estimates. To generate expression data with different variances per gene, we first computed gene-wise variances of the ALL 1 data set. Expression values were mean centered beforehand (per gene). From the set of 12625 variances, 2500 were randomly chosen. Data were generated for 2500 genes and 20 samples by drawing from a normal distribution with mean zero and respective variances per gene. We induced the first 500 genes by adding a value of 2 to the first ten samples. Z-scores were computed for 10000 permutations. Based on these we transformed the observed scores to gene-wise p-values. Pooled p-values were computed based on 1000 permutations. From both sets of p-values we estimated the positive false discovery rate and compared the results to the true values known in simulation. The data-generating process was repeated 100 times, including the sampling of variances. In the top row of Figure 4.6 we display the results of 15 simulation runs. In both cases, we observed accurate q-value estimates for the first 500 gene ranks, which agrees well with the simulation model. For higher ranks, the methods showed different behavior. On average, gene-wise p-values lead to unbiased q-values estimates, which includes the risk of under-estimation of the true positive false discovery rate. However, the differences between estimated and true values are smaller than for the pooled case. Here the estimates are conservative in the beginning but tend to severe under-estimation of the positive false discovery rate for higher ranks. When focusing on q-value ranks 1 to 500, we observed that pooling leads to larger differences than gene-wise computation in some of the 100 simulation runs, see second row of Figure 4.6.

The results above suggest to use gene-wise computation of p-values. Our observation in the simulation setting agrees with the discussion following the article of Ge *et al.* (2003). The authors argue that gene-wise p-values are needed to prove certain features of FDR-adjustment procedures like strong control of the true false discovery rate. With gene-wise p-values, the estimates are unbiased on average,

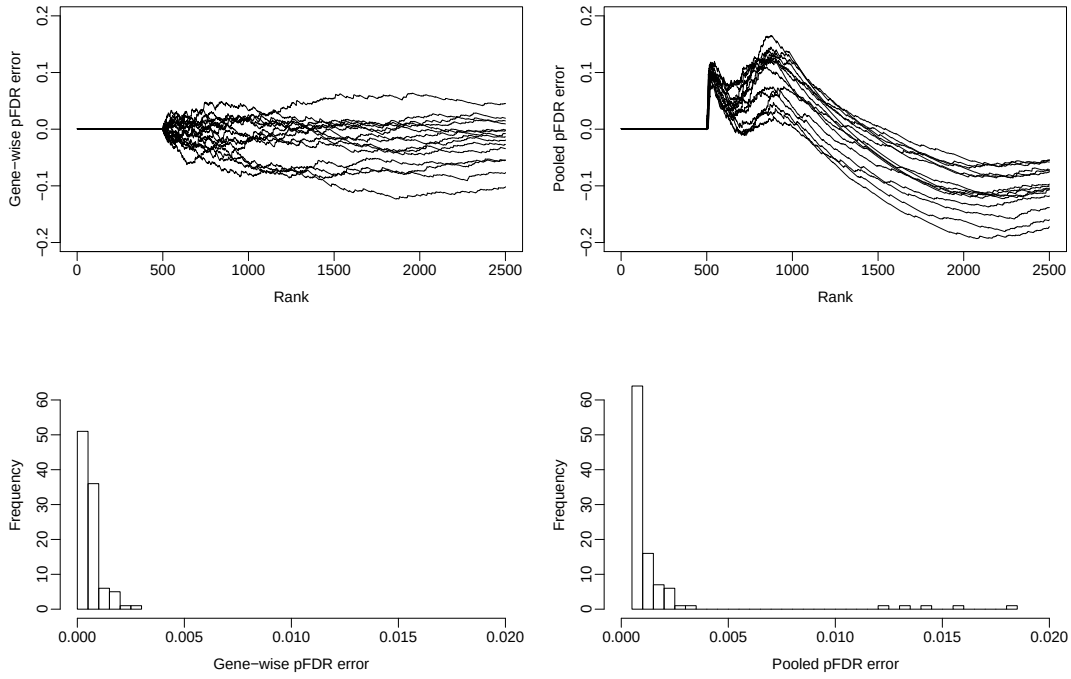

Figure 4.6: Differences between true and estimated false discovery rate in a simulation with 500 induced genes out of 2500. First row: shown are results of 15 simulation runs. P-values were computed in gene-wise (left) or pooled fashion (right). In both cases, the q-value estimates are accurate for the first 500 gene ranks. For ranks higher than 500, gene-wise computation on average leads to unbiased estimates with risk of under-estimation. Pooled computation leads to consistent behavior starting with conservative over-estimation of the true false discovery rate but running into severe under-estimation problems for higher ranks. Second row: close-up on differences of low-ranking q-values. Shown are frequencies of maximum differences found within ranks 1 to 500 in 100 simulation runs. With pooling, some larger differences occur.

which supports control. With pooled p-values, the estimates under-estimate the true values for higher ranks, which does not support control for higher FDR-cut-offs. However we used pooled p-values throughout the thesis to take advantage of faster computation and the fact that we do not concern control of error rates here. The introduced methods are based on the observed p-value *distribution* and not on single genes. Although the individual genes receive different rankings based on gene-wise or pooled p-values, the overall p-value distributions appear to be in good agreement, see Figure 4.7. In a Kolmogoroff-Smirnoff test, the null hypothesis of equal distribution functions had to be rejected ($p \approx 0.001$), which might be due to the large sample sizes of 12625 values each. However, use of the two p-value methods will lead to different interpretations of the individual local false discovery rates since the genes are ordered completely differently. Thus the esti-

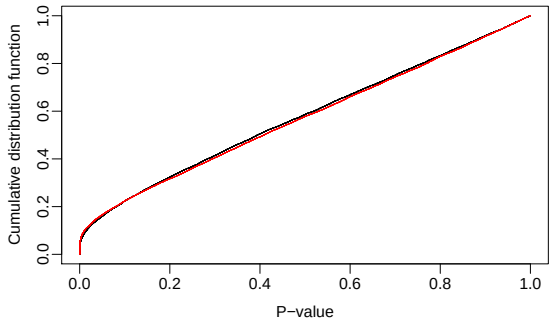

Figure 4.7: Cumulative distribution function of p-values in the ALL 1 comparison. P-values were derived by gene-wise (black curve) or pooled computation (red curve). The two computation methods lead to similar p-value distributions.

mates of a certain p-value level might agree well but the genes corresponding to this level might not be identical.