CHAPTER 8

Significance testing for small microarray experiments

Charles Kooperberg, Aaron Aragaki, Charles C. Carey, and Suzannah Rutherford

8.1 Introduction

When a study has many degrees of freedom it is sometimes less critical which significance test is carried out, as most analyses will give approximately the same result. However, when there are few degrees of freedom the choice of which significance test to use can have a strong effect on the results of an analysis. Unfortunately, this small degrees of freedom situation is often the case for microarray experiments, as many research laboratories perform such experiments with only a few repeats. Reasons for the small number of repeats include specimen availability and economics. Kooperberg et al. (2005) compare several approaches to significance testing for experiments with a small number of oligonucleotide arrays (a one-color technology; see Section 1.4). This chapter summarizes results from that analysis and describes a similar comparison for methods of carrying out significance testing for two-color arrays (e.g., cDNA arrays).

The large variability that even the most precise microarray platforms have makes small-sample comparisons unattractive. A standard t-test for an experiment with six two-color arrays has, depending on whether other variables are controlled for, at most five degrees of freedom. The resulting two-sided test, with $\alpha = 0.05$ and a Bonferroni correction for 10,000 genes requires a t-statistic value of at least 20.6 for significance. The lack of degrees of freedom drives the extremely large significance threshold for t-statistics: the same α and Bonferroni correction for 20 arrays requires a t-statistic of 6.3 while a normal distribution only requires a Z-statistic of 4.6. On the other hand, reducing the number of genes of interest on the original array from 10,000 to 500 only reduces the required t-statistic to 11.3.

Nonparametric (Wilcoxon) or permutation tests do not provide a simple solution to the significance problem. For example, for an experiment with n two-color arrays, a p-value for a permutation test can be no smaller than 2^{-n}; a two-sided test with $\alpha = 0.05$ and a Bonferroni correction for 10,000 genes requires n to be at least 19.

Reducing the number of genes to 500 reduces the minimum n to only 15. Similarly, for a one-color array the p-value for a permutation test with n cases and n controls cannot be smaller than $\binom{2n}{n}n$; so for a two-sided test with $\alpha = 0.05$ and a Bonferroni correction for 10,000 genes, at least $2n = 22$ arrays are needed. Reducing the number of genes to 500 reduces the minimum number of arrays to 18.

There is thus a need for a better estimate of the residual variance to overcome the lack of repeats. Combining information can be helpful in this regards. There are two obvious choices available: combine different genes in the same experiment or combine different experiments, if similar experiments were carried out. For combining genes, we can choose either to combine those genes for which the general expression level is similar (see e.g., Huang and Pan (2002) and Jain et al. (2003)) or to combine all genes. Alternative approaches to obtain more power with small experiments are to add a stabilizing constant to the estimate of the variance for each gene or to use a (Bayesian) model for the expression levels. Significance Analysis of Microarrays (SAM; Tusher et al., 2001) is a methodology that adds a constant to the estimate of the SD. The approaches by Baldi and Long (2001), Lönnstedt and Speed (2002), Smyth (2004) and Cui et al. (2005) are four related (empirical) Bayesian approaches. Wright and Simon (2003) discuss a closely related frequentist approach.

In practice, when carrying out tests for many thousands of genes simultaneously, a multiple testing correction is essential (Section 7.2.3; see Dudoit et al. (2003) for an extensive overview). However, the focus here is on obtaining a well-calibrated marginal p-value, so we do not control for multiple comparisons.

8.2 Methods

Most of the methods that we compare here can be used either for one-color arrays or for two-color (spotted) arrays. We assume that the arrays have been properly normalized; see Section 8.6 for preprocessing details for the experiments we analyze here.

8.2.1 Notation

Two-color arrays. For each gene and each two-color array, the value x_{ijl}^{M} summarizes the ($\log_2$-)expression ratio (M-value; see Section 1.4.1) between experimental conditions $k = 1$ and $k = 2$ (these may be different between experiments) for gene $i = 1, \ldots, g$ in experiment $j = 1, \ldots J$ on replicate array $l = 1, \ldots, L_j$. For each gene on each array there is also an estimate of the overall expression level x_{ijl}^{A}, typically this will be the average of the normalized $\log_2$ expression for both channels of the array. Unless there is confusion we write x_{ijl} instead of x_{ijl}^{M} for the $\log_2$-expression ratios.

Let μ_{ij} be the "true" (mean) $\log_2$-expression ratio of gene i in experiment j for condition 1 relative to condition 2. Set $\widehat{\mu}_{ij} = \sum_l x_{ijl}/L_j$, $s_{ij}^2 = \sum_l (x_{ijl} - \widehat{\mu}_{ij})^2$, and $x_{ij}^{A} = \sum_l x_{ijl}^{A}/L_j$.

One-color arrays. Similarly, for each gene and each one-color array let x_{ijkl} be the ($\log_2$-)expression value for experimental condition $k = 1$ or $k = 2$, for gene $i = 1, \ldots, g$ in experiment $j = 1, \ldots J$ on replicate array $l = 1, \ldots, L_{jk}$.

Let μ_{ijk} be the "true" mean ($\log_2$-)expression level of gene i in experiment j under condition k. Set $\widehat{\mu}_{ijk} = \sum_l x_{ijkl}/L_{jk}$ and $s^2_{ijk} = \sum_l (x_{ijkl} - \widehat{\mu}_{ijk})^2$.

8.2.2 *Significance tests*

All significance tests that we consider can be written in the form

$$\frac{\widehat{\mu}_{ij}}{\widetilde{\sigma}_{ij}/\sqrt{L_j}}$$

for two-color arrays and

$$\frac{\widehat{\mu}_{ij1} - \widehat{\mu}_{ij2}}{\widetilde{\sigma}_{ij}\sqrt{\frac{1}{L_{j1}} + \frac{1}{L_{j2}}}}$$

for one-color arrays; $\widetilde{\sigma}^2_{ij}$ is an estimate of the variance of x_{ijl}, so $\widetilde{\sigma}_{ij}$ estimates the standard deviation (SD). The methods discussed here differ primarily in how the estimate $\widetilde{\sigma}_{ij}$ is obtained. The traditional test statistics estimate $\widetilde{\sigma}_{ij}$ based only on the data on gene i in experiment j. Approaches that inflate the variance or that combine genes also use data on genes i^*, $i^* \neq i$, either implicitly, to estimate hyperparameters for the empirical Bayes approach that inflates the variance, or explicitly, to smooth the estimates for $\widetilde{\sigma}^2_{ij}$. Finally, the approaches that combine experiments use data on experiments j^*, $j^* \neq j$. Most of the methods below have a defined reference (null) distribution, but alternatively significance levels can be obtained using permutations (see Section 8.2.3); in fact, some authors recommend permutations as the method to obtain p-values.

Below we describe the test statistics included in the comparison. We provide details for the two-color arrays; modifications for one-color arrays are indicated. All these approaches are either already implemented in `R` packages available from BioConductor (`http://www.bioconductor.org`) or CRAN (`http://cran.r-project.org`), or are easily programmed in `R` code.

Traditional single gene within-experiment method

t-statistic. The traditional t-statistic is

$$t_{ij} = \frac{\widehat{\mu}_{ij}}{\widehat{\sigma}_{ij}/\sqrt{L_j}},$$

where $\widehat{\sigma}^2_{ij} = s^2_{ij}/(L_j - 1)$, provided $L_j > 1$. The reference distribution is the t-distribution with $L_j - 1$ degrees of freedom, and the main assumption is that for each gene i and experiment j the x_{ijkl} are independent having a normal distribution with variance σ_{ij}, although the t-test is generally considered to be robust against departures from normality.

The two-sample t-statistic is the equivalent test for one-color arrays. Use of this statistic assumes that the variance for both experimental conditions is the same. An alternative is the Welch (1938) two-sample t-statistic that does not make that assumption. This approach has almost no power for small sample sizes (Kooperberg et al., 2005), and should probably be avoided for small microarray experiments.

Methods combining genes: smoothing the variance

There have been several proposals in the literature to combine the estimates of the variance for several genes to obtain better estimates, so that the resulting test has more degrees of freedom. Typically the assumption that is made is that genes with the same expression level have approximately the same variance. Under this assumption estimates for the variance can be obtained by smoothing the variance as a function of the expression level. For one-color arrays there are methods which smooth the variances jointly and methods which smooth variances separately for both experimental conditions.

LPE. Jain et al. (2003) describe the Local Pooled Error test method (LPE), applicable to one-color arrays where both experimental conditions are measured separately. This method is outlined here and described in detail by Lee, Cho, and O'Connell (Chapter 7 of this volume). In this approach, let $\widehat{\sigma}^2_{ijk}$ be the the sample variance of the x_{ijkl}, for $l = 1, \ldots, L_{jk}$. LPE regularizes these estimates for each j and k separately by smoothing the $\widehat{\sigma}^2_{ijk}$ versus $\widehat{\mu}_{ijk}$. The assumption being made here is that genes with the same expression level for the same experiment and the same condition have (approximately) the same variance. Since the smoothing spline that is used effectively involves averaging a large number of genes, the authors use a normal reference distribution.

Loess. Huang and Pan (2002) make several related proposals. The main difference between their approach and *LPE* is that they first compute $\widehat{\sigma}^2_{ij}$ and smooth these estimates against $\widehat{\mu}_{ij} = \widehat{\mu}_{ij1} + \widehat{\mu}_{ij2}$ for one-color experiments and against x^A_{ij} for two-color experiments. Their simulation results show that, not unexpectedly, for the null-model a normal reference distribution is appropriate.

Methods combining genes: (empirical) Bayesian model for σ Rather than smoothing the variance explicitly as a function of the expression level, we can include information from other genes for the analysis of a particular gene by making assumptions about the distribution of the variance for all genes. The information about the other genes then allows us to estimate (hyper)parameters that can be used to stabilize the variance estimate. There are several such methods, based on different motivations: *ad hoc* (Tusher et al., 2001), (empirical) Bayes argument (Baldi and Long, 2001; Lönnstedt and Speed, 2002; Smyth, 2004), James-Stein-type estimation (Cui et al., 2005), or a frequentist approach (Wright and Simon, 2003).

Some approaches combine the sample variance $\widehat{\sigma}^2_{ij}$ with another estimate σ^2_{0ij} that has d_{ij} degrees of freedom, yielding a variance estimate

$$\tilde{\sigma}_{ij}^2 = \frac{d_{ij}\sigma_{0ij}^2 + (L_j - 1)\hat{\sigma}_{ij}^2}{d_{ij} + L_j - 1}, \tag{8.1}$$

that can be used in a t-test with $d_{ij} + L_j - 1$ degrees of freedom. The methods *Cyber-T* and *Limma* use this approach; they differ primarily in the methods to obtain σ_{0ij}^2 and d_{ij}.

Cyber-T. The *Cyber-T* approach of Baldi and Long (2001) is motivated as a fully Bayesian procedure. However as implemented in practice (Baldi and Long, 2001, Section 5) the test is carried out using a t-test on (for two-color arrays) $\nu_0 + L_j - 1$ degrees of freedom, and an variance estimate (compare Equation 8.1)

$$\tilde{\sigma}_{ij}^2 = \frac{\nu_0\sigma_{0ij}^2 + (L_j - 1)\hat{\sigma}_{ij}^2}{\nu_0 + L_j - 1}, \tag{8.2}$$

where σ_{0ij}^2 is an estimate of the "prior variance" that is obtained as a running average of the variance estimates of the genes in a "window" of size w of similar x_{ij}^A. Thus, the *Cyber-T* approach uses the average of a smoothed variance (like *LPE* and *Loess*, just using a different smoother) with the regular variance of the t-statistic. A non-Bayesian interpretation of *Cyber-T* is thus that it combines a smoothed estimate (as in *Loess* and *LPE*) with a traditional estimate from the t-test.

We use the default values $\nu_0 = 10$ and window width $w = 101$ in the `R` software available at http://cybert.microarray.ics.uci.edu. (Note that in Baldi and Long (2001) a different default value of $\nu_0 = 10 - L_j$ is mentioned.)

Limma. Smyth (2004) generalizes the approach of Lönnstedt and Speed (2002). The main assumption in Smyth's model is a prior distribution on the variances σ_{ij}^2:

$$\frac{1}{\sigma_{ij}^2} \sim \frac{1}{d_{0j}s_{0j}^2}\chi_{d_{0j}}^2.$$

The model also includes priors on the coefficients for each gene in a linear regression model, which in the two-sample case reduces to the difference between the mean expression for the two groups. By the method of moments, estimates of d_{0j}, s_{0j}^2, and other parameters are obtained. An inflated variance

$$\tilde{\sigma}_{ij}^2 = \frac{d_{0j}s_{0j}^2 + (L_j - 1)\hat{\sigma}_{ij}^2}{L_j + d_{0j} - 1} \tag{8.3}$$

(compare Equation 8.2) is used for a "moderated t-test" with $d_{0j} + L_j - 1$ degrees of freedom. Thus, a main difference between the *Limma* approach of Smyth (2004) and the *Cyber-T* approach of Baldi and Long (2001) is that *Limma* uses one single estimate for the prior variance (s_{0j}^2) for all genes and estimates the prior degrees of freedom d_{0j} based on the data, while *Cyber-T* uses a smooth estimate for the prior variance σ_{0ij}^2, but it uses a fixed number of prior degrees of freedom ν_0. The approach of Smyth (2004) is implemented in the BioConductor package `limma` (Smyth, 2005).

Shrinking. Cui and Churchill (2003) and Cui et al. (2005) develop a James-Stein shrinkage estimate $\widetilde{\sigma}_{ij}^2$. After appropriate transformations this estimator "shrinks" the t-test estimate $\widehat{\sigma}_{ij}^2$ toward the mean variance $\sum_{i=1}^{n} \sigma_{ij}^2/In$, where the exact amount of shrinkage differs from gene to gene, and depends on the variability for that gene. Easy to implement formulas are given in Cui et al. (2005). The authors of this method recommend a permutation approach (see Section 8.2.3) to obtaining p-values. We include this approach without permutations using a normal reference distribution, as well as with the permutation p-values.

Methods combining experiments

Instead of simply combining different genes *within* one experiment, we can also combine expression levels of the same gene *between* experiments carried out using the same microarray platform. This would potentially be useful if there are several smaller experiments for which it is reasonable to assume that for each gene the variance in each experiment is approximately the same.

Pooled-t. We define the pooled t-test statistic, combining experiments, as

$$c_{ij} = \frac{\widehat{\mu}_{ij}}{\widehat{\sigma}_i\sqrt{\frac{1}{L_j}}},$$

where $\widehat{\sigma}_i^2 = \sum_j s_{ij}^2/L$ and $L = \sum_j (L_j - 1)$, provided $L > 0$. The reference distribution is the t-distribution with L degrees of freedom, and the main assumption is that the x_{ijl}^M are independent for each j and l, having a normal distribution with mean μ_{ij} and variance σ_i^2.

It is in principle also possible for the other methods discussed above to pool different experiments in obtaining a single variance estimates. Since these methods already regularize the estimates for σ^2 in some way, pooling typically has little or no effect, and the corresponding combined method behaves similarly to the "parent" method (Kooperberg et al., 2005).

8.2.3 Permutation p-values

Permutation of the arrays in an experiment can be an alternative to using a parametric reference distribution for a test statistic. Assume that we have a two-color experiment with L arrays, and that the test statistic for the ith gene is T_i. To compute the significance of T_i we also compute the test statistics for all genes for each of the $m = 1, \ldots, 2^L$ experiments that are obtained by "flipping" the signs of the x_{il}^m for some of the l. (We omit the index of experiment j.) Note that one of these permutations will be the original design. Let T_i^m be the test statistic for the ith gene for the m^{th} permutation. We estimate the p-value corresponding to T_i as

$$\sum_{i^*=1}^{n} \sum_{m=1}^{2^L} I(T_i < T_{i*}^m)/n2^L,$$

where $I(\cdot)$ is the indicator function. If L is larger than, say, eight, it may be preferred to sample permutations (rather than computing all possible permutations) to save computing time.

These estimates will be unbiased if (i) each T_i has the same distribution under the null-hypothesis, and (ii) no genes are differentially expressed. The first assumption is not as severe as it might appear, since no particular parametric form for the common distribution is assumed. The second assumption is much more severe, and it will lead to conservative p-values when in fact a substantial number of genes are differentially expressed (Storey and Tibshirani, 2003b).

For one-color arrays, we randomly rearrange the L_1 arrays with the first experimental condition and the L_2 arrays with the second experimental condition, and proceed in a similar manner.

8.3 Data

We analyze two sets of data. One comes from an unpublished study of Drosophila, and the other comes from a one-color experiment that is analyzed in Kooperberg et al. (2005).

Table 8.1 *Organization of the two-color data. Experiments whose code starts with a D (different) are expected to have differences between both groups, while those starting with an S (same) are repeats; the digit "2" refers to a two-color array. The arrays for experiments D2.3 and D2.4 and those for D2.5 and D2.6 are different; experiment S2.1 are arrays from a cell-line not used for the other experiments.*

Exp.	Sample 1	Sample 2	L_j	Different
S2.1	KC cell	KC cell	4	no
S2.2	SAM	SAM	2	no
S2.3	SAM	SAM	2	no
S2.4	SAM	SAM	4	no
D2.1	SAM	D-recomb 304	2	yes
D2.2	SAM	D-recomb 220	2	yes
D2.3	SAM	D-pure	2	yes
D2.4	SAM	D-pure	4	yes
D2.5	SAM	E-pure	4	yes
D2.6	SAM	E-pure	4	yes
D2.7	SAM	F-pure	6	yes

The two-color experimental data come from a series of spotted microarrays (13,440 spots) of *Drosophila melanogaster* that were grown in Suzannah Rutherford's lab at the Fred Hutchinson Cancer Research Center. All experiments are "dye-swapped": i.e., half of the arrays have sample one on the red channel (and therefore sample two in green), the other half have sample two on the red channel (with sample one in green). The arrays that we compare here include some experiments that are self-self hybridizations, and some experiments where both samples are genetically different (see Table 8.1). In a self-self hybridization the two labeled samples are from the same source, so no genes are in fact differentially expressed and those identified as such are false positives. Thus, the experiments S2.1, S2.2, S2.3, and S2.4 are intended to establish that the tests have the right size Type I error, and the experiments D2.1, D2.2, D2.3, D2.4, D2.5, D2.6, and D2.7 are intended to establish power properties of the tests.

One-color experimental data was obtained using Affymetrix Mu 11K-A microarrays (6,595 probe sets) generated for a series of experiments on Huntington's disease (HD) mouse models. The results of these experiments are reported in a series of related papers (Chan et al., 2002; Luthi-Carter et al., 2002a,b). For this analysis we compare cerebellar gene expression in similarly aged mice carrying either a wild type or mutant form of the HD gene. Every comparison reported in Chan et al. (2002), Luthi-Carter et al. (2002a) and Luthi-Carter et al. (2002b) shows some differentially expressed genes, although the amounts of differential expression differ considerably between the experiments. For each of the experiments both groups had between two and five mice. Thus, all the repeats use different samples (sometimes referred to as "biological replicates") and are not repeat arrays using the same samples (sometimes referred to as "technical replicates"). The one-color experiments are listed in Table 8.2. Again, test size is examined with the S experiments and power with the D experiments.

Table 8.2 *Organization of the one-color (Affymetrix) data. HD: Huntington's disease mouse, WT: wild type mouse. Experiments whose code starts with a D are expected to have differences between both groups, while those starting with an S are repeats; the digit "1" refers to a one-color (Affymetrix) array.*

Exp.	Tissue	Mouse	Group 1	Group 2	L_{j1}	L_{j2}	Different
S1.1	cerebellum	DRPLA 26Q	HD	HD	2	2	no
S1.2	cerebellum	DRPLA 26Q	WT	WT	2	2	no
S1.3	cerebellum	YAC	HD	HD	3	2	no
S1.4	cerebellum	YAC	WT	WT	3	2	no
D1.1	cerebellum	DRPLA 65Q	HD	WT	4	4	yes
D1.2	cerebellum	R6/2 12 weeks	HD	WT	2	2	yes
D1.3	cerebellum	N171	HD	WT	4	4	yes

8.4 Results

We analyze the experiments listed in Section 8.3 using the methods described in Section 8.2.2. For the experiments where both groups are different (D2.x and D1.x) we prefer methods with the largest percentage of significant genes (the largest power), provided that the method does have the correct percentage of significant genes in the experiments where both groups are the same (S2.x and S1.x): i.e., at most $\alpha\%$ significant genes when tested at significance level $\alpha\%$. This power comparison is fair when the Type I error rate is controlled at the same level α.

We show results for $\alpha = 1\%$ and $\alpha = 0.01\%$. For the two-color arrays there are approximately 11,000 genes after removal of spots (genes) whose intensities are too close to the background level (see Section 8.6). Assuming independence of genes, a 95% confidence interval for the percentage of significance genes based upon the binomial distribution is between 0.8 and 1.2% at $\alpha = 1\%$ and between 0 and 0.03% at $\alpha = 0.01\%$. For the one-color arrays there are 6,595 genes, thus these confidence intervals are slightly larger (0.75 through 1.25% at $\alpha = 1\%$ and 0 and 0.045% at $\alpha = 0.01\%$). When we average four experiments and (incorrectly) assume independence, we expect between about 0.9 and 1.1% significant genes at $\alpha = 1\%$ and between 0 and 0.025% at $\alpha = 0.01\%$ for both array types.

8.4.1 Bandwidth selection for smoothers

The methods *Cyber-T*, *LPE*, and *Loess* require the choice of a bandwidth or smoothing parameter. For *LPE* and *Loess* this determines over how many genes the variance is "averaged". For *Cyber-T* the averaged variance is combined with the variance for the individual genes.

Table 8.3 summarizes the results for the two-color experiment for the *Loess* approach. The parameter `span` for the `loess` function in `R` (Ihaka and Gentleman, 1996) is approximately linear in the bandwidth for a local linear smoother. Table 8.3 shows that the bandwidth has very little influence on the results. The explanation for this is that even for the smallest bandwidth the variances of several dozen genes are effectively averaged. Smaller values of `span` are not useful, as they lead to numerical problems in regions with little data.

For all four choices of `span` and for all S2.x experiments at $\alpha = 0.01\%$ and for two of the four of these experiments at $\alpha = 1\%$, the percentage of genes called significant is much too large. This was concluded by Kooperberg et al. (2005) for the one-color arrays.

For the remainder of the comparisons we use a `span` of 0.1, which yields the lowest average number of significant results for both $\alpha = 1\%$ and $\alpha = 0.01\%$ for the four S2.x experiments. As the influence of the bandwidth appears minimal, we use *Cyber-T* and *LPE* with their default values.

Table 8.3 *Performance of the Loess approach with varying bandwidth (*`span`*) for the two-color experiments. We report the percentage of genes called differentially expressed at levels* $\alpha = 1\%$ *and* $\alpha = 0.01\%$*. Ideally the four S2.x experiments would have* α *differentially expressed genes, while the seven D2.x would have many such genes.*

	$\alpha = 1\%$				$\alpha = 0.01\%$			
`span`	10	1	0.1	0.01	10	1	0.1	0.01
S2.1	1.1	1.1	0.7	0.7	0.340	0.306	0.198	0.159
S2.2	7.8	7.0	5.8	6.6	2.884	2.507	1.528	1.915
S2.3	2.2	2.1	2.0	2.0	0.984	0.922	0.982	0.942
S2.4	0.7	0.6	0.6	0.6	0.262	0.262	0.230	0.212
S2-ave	3.0	2.7	2.3	2.5	1.118	0.999	0.735	0.807
D2.1	25.8	25.9	26.8	27.1	11.941	11.994	12.698	12.827
D2.2	31.7	31.8	32.3	32.9	16.817	17.000	17.682	18.300
D2.3	53.5	53.6	53.8	53.8	38.170	38.354	38.368	38.457
D2.4	54.3	54.4	54.4	54.7	37.709	37.858	37.774	38.043
D2.5	43.3	43.5	43.5	44.2	28.006	28.190	28.225	28.574
D2.6	73.0	73.2	76.5	76.6	62.230	62.431	66.313	66.501
D2.7	62.1	62.3	64.3	64.3	47.863	48.003	50.124	50.471
D2-ave	49.1	49.2	50.2	50.5	34.677	34.833	35.883	36.168

8.4.2 Comparison of methods

Tables 8.4 and 8.5 show the results of the methods described in Section 8.2.2 when applied to the data described in Section 8.3 (results for the *LPE* method are not available for the two-color data). Cui et al. (2005) recommend permutations to obtain p-values for the *Shrinking* approach. In Tables 8.4 and 8.5 and Figures 8.1 and 8.2 we use a normal reference distribution. Tables 8.6 and 8.7 and Figures 8.3 and 8.4 use the permutation approach. The choice of distribution has a substantial impact on the results.

Figure 8.1 gives a graphical display of how well these methods adhere to the nominal significance levels, Figure 8.2 displays power results. These figures are probability-probability plots on a logit-scale. For a given method and a particular experiment let p_i be the two-sided (sometimes called signed) p-values; that is, if p_i is close to 0 there is evidence of under-expression and if p_i is close to 1 there is evidence of over-expression of group one relative to group two. We now combine all p_i for a group of experiments and sort them. Assume that there are N p-values. The sorted p-values are plotted on the horizontal axis, with $(1, \ldots, n)/(N + 1)$ on the vertical axis. For a self-self experiment, these plots should ideally follow the identity line, as that implies that the significance levels are "unbiased." Curves that flatten out are particularly worrisome, as they suggest significantly differentially expressed genes that are in fact false positives. Curves that are more vertical than the identity line suggest statistics that are too conservative: something that is not a concern when there is in fact no difference, but would likely be harmful when using the same method to analyze data where some genes are in fact differentially expressed. For the D experiments, where there is a difference between the two sample types, the ideal curve is more horizontal, as long as the method does not generate a substantial number of false positives in the S experiments.

Figure 8.1 shows that the *Loess* and *LPE* approaches identify substantially more differentially expressed genes than the nominal levels for the S experiments. The *Cyber-T* approach shows a mild number of increases, and none of the other approaches shows serious bias. For both groups of experiments, the *Shrinking* approach with a normal reference distribution appears too conservative.

Table 8.4 elaborates these observations. Although most methods appear to be rather conservative, at a significance level of $\alpha = 1\%$ the *Loess* method shows a substantial anticonservative bias, in five out of eight data sets. For microarray experiments, the more stringent level $\alpha = 0.01\%$ is very relevant, as multiple testing corrections generally imply selecting genes at low significance levels. Again, the *Loess* shows substantial bias. The *LPE* approach also indicates ten times more significant genes than the nominal value; this bias is present for three of the four data sets. At this significance level, the *Cyber-T* method shows a modest bias overall, being substantial for only one dataset (two-color experiment S2.2). The excess percentage of significant genes for the *Pooled-t* approach is minimal, and could be due to chance.

In Figure 8.2 it is seen that for all methods far more genes are identified as differentially expressed by the two-color experiments than by the one-color experiments, as the curves for the two-color experiments are much more horizontal than those for the one-color experiments. This is largely an effect of the particular data used, as the two-color Drosophila experiments involved substantially altered flies, while the differences between the mice involved in the one-color Huntington's disease experiments are much more subtle. This figure does indicate though that the ordering of the methods is largely unchanged, suggesting that since the conclusions remain the same for two dramatically different experiments (different technologies, different amounts of differential genes) they appear to be fairly robust and may well generalize to many other situations.

In both the two-color and the one-color experiments the *Loess* approach produces the most genes identified as differentially expressed. This is not a surprise, since the method does not maintain the correct significance levels in the self-self (S) experiments. Similarly, it is not surprising that the *LPE* method identifies more differential expression for the one-color experiments, since it also does not adequately control test size here. Among the remaining methods, which tend to maintain significant levels rather conservatively, the *Pooled-t* approach performs best for the two-color experiments, followed by *Cyber-T* and *Limma*, while for the one-color experiments *Cyber-T* and *Limma* approach seem slightly more powerful than *Pooled-t* (Table 8.5). Interestingly for the D2.x (two-color) experiments, *Pooled-t* seems more powerful in

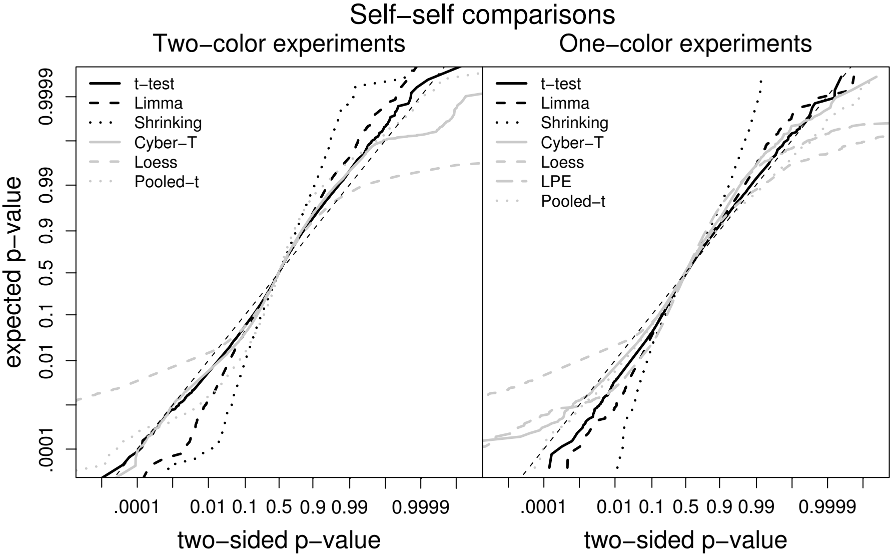

Figure 8.1 *Method performance using a defined null reference distribution in self-self (S) experiments. For unbiased methods the curves should follow the identity line.*

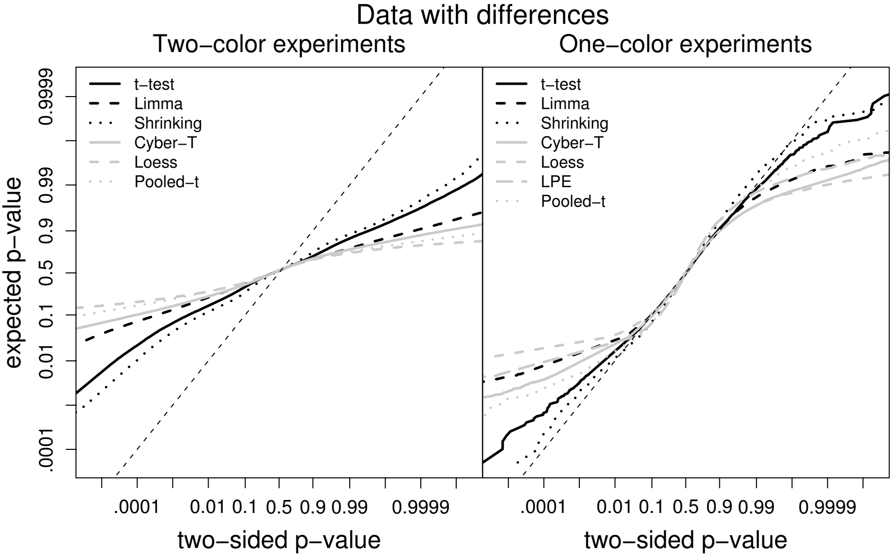

Figure 8.2 *Method performance using a defined null reference distribution in difference (D) experiments. If there is appropriate Type I error control, curves that are more horizontal correspond to more powerful methods.*

those experiments with two arrays (D2.1, D2.2, and D2.3). Maybe this is not surprising: borrowing degrees of freedom between experiments, as *Pooled-t* does, is particularly useful when the number of degrees of freedom is small.

8.4.3 Permutation p-values

As detailed in Section 8.2.3, an alternative to obtaining p-values is a permutation approach in which the test statistics for all genes are combined. Figure 8.3 gives a graphical display of how well each method adheres to the significance levels when p-values are determined using such an approach. Figure 8.4 displays curves related to power for these situations. We do not show permutation results for *Pooled-t*: since this procedure combines arrays from different experiments a permutation procedure is less standard, and in any case the results using a t-distribution are already satisfactory.

The permutation approach for computing p-values yields approximately unbiased, if somewhat conservative, results for all approaches since all curves in Figure 8.3 follow the diagonal. However, as expected, the permutation approach is associated with a reduction in the number of genes called differentially expressed. Figure 8.4 shows that the procedures based on permutation produce considerably fewer differentially expressed genes than the procedures that do not use permutation (Figure 8.2). In fact, the curves in Figure 8.4 all stay within a "band" of the diagonal. This result is a consequence of using the permutation approach with a small number of repeats: irrespective of the actual number of differentially expressed genes, there is a maximum number of genes that can be identified as differentially expressed at any particular significance level due to the experimental design. A detailed explanation is given below in the discussion of Table 8.7.

Tables 8.6 and 8.7 summarize results for the permutation-based procedures. Although the permutation approach does control the significance level α appropriately, there is correspondingly less differential expression identified for these data and methods. The part of Table 8.7 for the D2.x experiments clearly illustrates an artifact of the permutation approach. As already seen above, the D2.x experiments have very many genes identified as differentially expressed (see Table 8.5). But in Table 8.7 there seems to be a cap: at a significance level of $\alpha = 1\%$ for experiments D2.1, D2.2, and D2.3 all methods suggest at most 2% differentially expressed genes, for experiments D2.4, D2.5, and D2.6 all methods suggest at most 8% differentially expressed genes, and for experiments D2.7 all methods suggest at most 32% differentially expressed genes. We focus on experiment D2.4, which uses 4 arrays. There thus result at most $2^4 = 16$ permutations from "flipping" the arrays. Since each permutation arises twice (when all arrays are flipped relative to the first analysis), only 8 of these permutations are unique. Assume that for this experiment 40% of the genes are differentially expressed (as Table 8.5 suggests), and therefore that these 40% of the genes have very large test statistics. With about 10,000 genes on these arrays, there are thus about 4,000 large test statistics, say larger than a value A. Now

Table 8.4 *Percentage of differentially expressed genes in self-self (S) experiments identified using a defined null reference distribution at significance levels $\alpha = 1\%$ and $\alpha = 0.01\%$. For unbiased methods the percentage of differentially expressed genes should be close to α.*

$\alpha = 1\%$	*t-test*	*Limma*	*Shrinking*	*Cyber-T*	*Loess*	*LPE*	*Pooled-t*
S2.1	0.2	0.1	0.0	0.1	0.7	NA	0.3
S2.2	1.1	0.1	0.0	2.3	5.8	NA	0.3
S2.3	0.6	0.2	0.0	0.3	2.0	NA	0.4
S2.4	0.2	0.1	0.0	0.0	0.6	NA	0.1
S2-ave	0.5	0.1	0.0	0.7	2.3	NA	0.3
S1.1	0.4	0.2	0.0	0.4	0.7	0.4	0.0
S1.2	0.6	0.3	0.0	1.4	2.7	1.1	0.2
S1.3	0.8	0.1	0.0	0.3	3.9	0.3	3.2
S1.4	0.3	0.0	0.0	0.1	2.6	0.1	1.3
S1-ave	0.5	0.2	0.0	0.6	2.5	0.5	1.2

$\alpha = 0.01\%$	*t-test*	*Limma*	*Shrinking*	*Cyber-T*	*Loess*	*LPE*	*Pooled-t*
S2.1	0.000	0.000	0.000	0.000	0.198	NA	0.017
S2.2	0.009	0.000	0.000	0.277	1.528	NA	0.061
S2.3	0.018	0.000	0.000	0.000	0.982	NA	0.009
S2.4	0.000	0.000	0.000	0.000	0.230	NA	0.009
S2-ave	0.007	0.000	0.000	0.069	0.735	NA	0.024
S1.1	0.015	0.030	0.000	0.061	0.197	0.106	0.000
S1.2	0.000	0.000	0.000	0.045	0.697	0.243	0.000
S1.3	0.000	0.000	0.000	0.015	0.500	0.061	0.091
S1.4	0.000	0.000	0.000	0.000	0.728	0.000	0.000
S1-ave	0.004	0.008	0.000	0.030	0.531	0.102	0.023

Table 8.5 *Percentage of differentially expressed genes in difference (D) experiments identified using a defined null reference distribution at significance levels* $\alpha = 1\%$ *and* $\alpha = 0.01\%$. *If there is appropriate Type I error control, a larger percentage of differentially expressed genes corresponds to a more powerful method.*

$\alpha = 1\%$	*t-test*	*Limma*	*Shrinking*	*Cyber-T*	*Loess*	*LPE*	*Pooled-t*
D2.1	1.9	12.1	0.0	15.8	26.8	NA	30.9
D2.2	2.3	16.0	0.0	21.9	32.3	NA	28.9
D2.3	4.0	34.8	0.0	43.6	53.8	NA	48.2
D2.4	31.0	44.8	22.6	45.5	54.4	NA	62.7
D2.5	20.9	31.6	13.1	35.1	43.5	NA	52.4
D2.6	53.6	66.5	46.3	66.9	76.5	NA	58.6
D2.7	51.8	57.6	46.9	55.9	64.3	NA	56.3
D2-ave	23.7	37.6	18.4	40.7	50.2	NA	48.3
D1.1	2.6	3.4	2.0	4.0	6.4	2.7	3.3
D1.2	1.2	5.3	0.1	5.6	6.7	5.0	1.5
D1.3	1.6	1.6	1.0	1.6	3.0	0.9	0.8
D1-ave	1.8	3.4	1.1	3.7	5.4	2.9	1.9

$\alpha = 0.01\%$	*t-test*	*Limma*	*Shrinking*	*Cyber-T*	*Loess*	*LPE*	*Pooled-t*
D2.1	0.009	0.864	0.000	2.148	12.698	NA	10.835
D2.2	0.026	1.219	0.000	5.051	17.682	NA	11.928
D2.3	0.027	7.699	0.000	19.441	38.368	NA	26.722
D2.4	1.994	15.378	0.296	21.732	37.774	NA	44.632
D2.5	1.083	4.752	0.201	10.856	28.225	NA	31.806
D2.6	7.729	39.769	2.858	47.705	66.313	NA	40.295
D2.7	17.023	29.986	11.971	34.357	50.124	NA	38.347
D2-ave	3.984	14.238	2.189	20.184	35.883	NA	29.224
D1.1	0.121	0.349	0.030	1.046	2.593	0.788	0.516
D1.2	0.000	2.153	0.000	1.668	2.835	2.092	0.243
D1.3	0.106	0.243	0.061	0.379	1.410	0.288	0.182
D1-ave	0.076	0.915	0.030	1.031	2.280	1.056	0.313

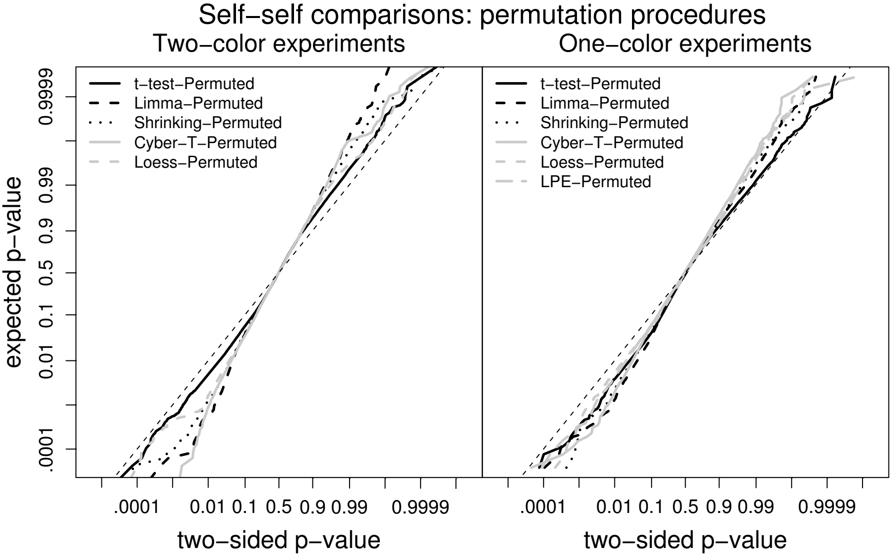

Figure 8.3 *Method performance using a permutation reference distribution in self-self experiments. For unbiased methods the curves should follow the identity line.*

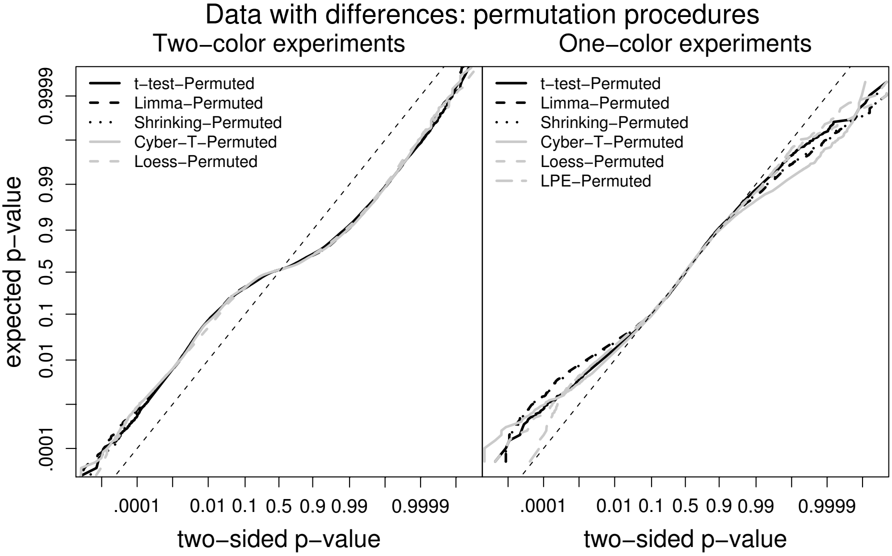

Figure 8.4 *Method performance using a permutation reference distribution in difference experiments. Curves that are more horizontal correspond to more powerful methods.*

Table 8.6 *Percentage of differentially expressed genes in self-self (S) experiments identified using a permutation distribution at significance levels $\alpha = 1\%$ and $\alpha = 0.01\%$. For unbiased methods the percentage of differentially expressed genes should be close to α.*

$\alpha = 1\%$	*t-test* permuted	*Limma* permuted	*Shrinking* permuted	*Cyber-T* permuted	*Loess* permuted	*LPE* permuted
S2.1	0.1	0.0	0.0	0.0	0.0	NA
S2.2	1.0	0.0	0.2	0.4	0.6	NA
S2.3	0.6	0.1	0.1	0.0	0.4	NA
S2.4	0.2	0.1	0.1	0.0	0.2	NA
S2-ave	0.5	0.1	0.1	0.1	0.3	NA
S1.1	0.3	0.1	0.1	0.1	0.1	0.1
S1.2	0.6	0.4	0.4	0.3	0.4	0.4
S1.3	1.1	0.5	0.4	0.2	0.5	0.5
S1.4	0.3	0.1	0.1	0.1	0.4	0.2
S1-ave	0.6	0.2	0.2	0.1	0.4	0.3

$\alpha = 0.01\%$	*t-test* permuted	*Limma* permuted	*Shrinking* permuted	*Cyber-T* permuted	*Loess* permuted	*LPE* permuted
S2.1	0.000	0.000	0.000	0.000	0.000	NA
S2.2	0.000	0.000	0.000	0.000	0.000	NA
S2.3	0.017	0.000	0.000	0.000	0.000	NA
S2.4	0.000	0.000	0.008	0.000	0.000	NA
S2-ave	0.004	0.000	0.002	0.000	0.000	NA
S1.1	0.000	0.000	0.000	0.000	0.000	0.000
S1.2	0.000	0.000	0.000	0.000	0.000	0.000
S1.3	0.000	0.000	0.000	0.015	0.000	0.015
S1.4	0.000	0.000	0.000	0.000	0.000	0.000
S1-ave	0.000	0.000	0.000	0.004	0.000	0.004

assume that among the 7 other permutations none of the test statistics is larger than A. Then out of $8 \times 10,000 = 80,000$ test statistics, there are 4,000 larger than A. However, at the $\alpha = 1\%$ level at most $0.01 \times 80,000 = 800$ can be called signif-

Table 8.7 *Percentage of differentially expressed genes in difference (D) experiments identified using a permutation distribution at significance levels $\alpha = 1\%$ and $\alpha = 0.01\%$. If there is appropriate Type I error control, a larger percentage of differentially expressed genes corresponds to a more powerful method.*

$\alpha = 1\%$	*t-test* permuted	*Limma* permuted	*Shrinking* permuted	*Cyber-T* permuted	*Loess* permuted	*LPE* permuted
D2.1	1.6	2.0	1.8	2.0	2.0	NA
D2.2	1.5	2.0	2.0	2.0	2.0	NA
D2.3	1.9	2.0	2.0	2.0	2.0	NA
D2.4	7.7	8.0	8.0	8.0	8.0	NA
D2.5	7.4	8.0	8.0	7.9	7.5	NA
D2.6	8.0	8.0	8.0	8.0	0.0	NA
D2.7	30.5	31.8	30.5	31.8	24.8	NA
D2-ave	8.4	8.8	8.6	8.8	7.8	
D1.1	2.8	3.8	3.8	3.6	2.8	2.8
D1.2	1.2	3.0	2.6	2.7	2.7	2.7
D1.3	1.9	1.8	1.8	1.4	1.3	1.0
D1-ave	2.0	2.9	2.7	2.6	2.3	2.1

$\alpha = 0.01\%$	*t-test* permuted	*Limma* permuted	*Shrinking* permuted	*Cyber-T* permuted	*Loess* permuted	*LPE* permuted
D2.1	0.008	0.008	0.008	0.008	0.017	NA
D2.2	0.017	0.017	0.017	0.017	0.026	NA
D2.3	0.009	0.008	0.000	0.009	0.018	NA
D2.4	0.068	0.076	0.076	0.068	0.079	NA
D2.5	0.075	0.083	0.059	0.084	0.079	NA
D2.6	0.075	0.075	0.075	0.025	0.068	NA
D2.7	0.308	0.315	0.283	0.308	0.314	NA
D2-ave	0.080	0.083	0.074	0.074	0.086	NA
D1.1	0.121	0.258	0.212	0.243	0.106	0.030
D1.2	0.000	0.000	0.015	0.015	0.015	0.015
D1.3	0.136	0.243	0.258	0.212	0.121	0.045
D1-ave	0.086	0.167	0.162	0.157	0.081	0.030

icant (at $\alpha = 1\%$) from the permutation distribution. This makes 8%, rather than the 40% that are differentially expressed, of all the genes on the array. (In fact the percentage called differentially expressed is slightly lower as a few rare permuted genes also have large statistics.) We cannot ignore the original permutation to obtain percentiles of the permutation distribution, since doing so violates the assumption of exchangeability under the null hypothesis of no differential expression. This artifact disappears when the number of differentially expressed genes is much smaller or when the number of arrays increases, since then more permutations can be created.

8.4.4 Relation between average signal and variance

The local smoothing approaches generally assume that genes with the same expression level have approximately the same variance, then estimate the variance by smoothing as a function of the expression level. We examine this relationship here. Figure 8.5(a) contains an MA plot for an individual two-color array, showing the relation between the difference between the logs of the two signals (i.e., the log-ratio, or M-value) and the average of the logs of the signals. For one of the two-color experiments (Figure 8.5) and one of the one-color experiments (Figure 8.6), the relation between the variance and the average signal is shown. As can be seen, the relation between average signal and variance is minimal. In fact, the correlation between the variance from one experiment to the next experiment for the same gene is much larger than the correlations in these figures (data not shown).

8.5 Discussion

We have seen here that the choice of significance test in microarray experiments with low replication can dramatically influence the results. We focus on p-values, rather than for example the false discovery rate (FDR), as we believe that an appropriately obtained p-value will yield a more reliable multiple testing correction, and that the multiple testing adjustment cannot by itself save a procedure that yields badly calibrated p-values.

The two groups of experiments analyzed here differ in another important aspect besides technology: the one-color experiments have a modest number of differentially expressed genes, while the two-color experiments have many such genes. Given this difference between the experiments, the similarity in results is striking.

The main conclusions are:

1. The t-test has almost no power when the sample size is small. When there are fewer than, say, six to eight repeat arrays some of the alternative solutions are much more powerful. Kooperberg et al. (2005) conclude that the lack of power is even more extreme for the Welch statistic, which suffers at least in part because the variance estimate is not pooled.

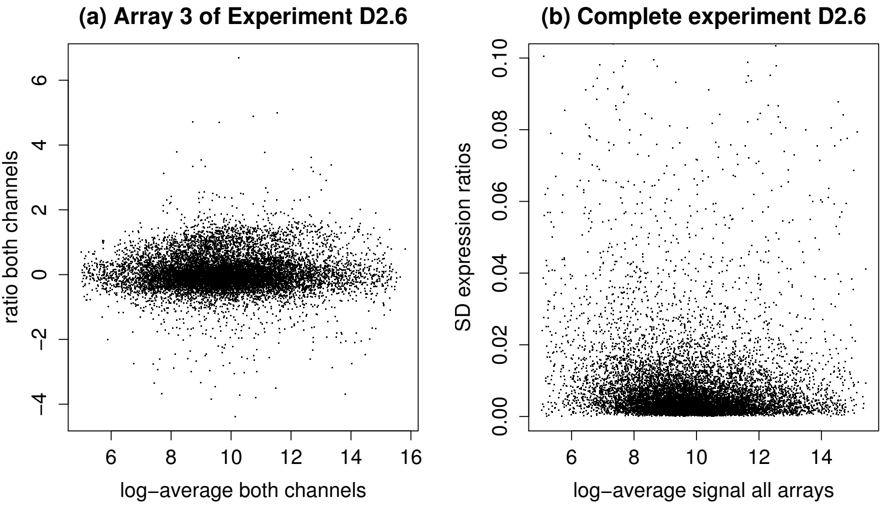

Figure 8.5 *(a) Relation between log expression ratio and average log expression for one normalized two-color array, and (b) SD of log expression ratios vs. average log expression ratio for all arrays from the experment (D2.6).*

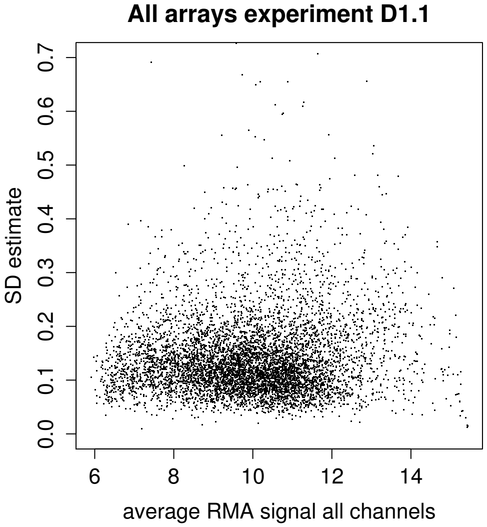

Figure 8.6 *Residual SD vs. average RMA value for one of the one-color experiments (D1.1).*

2. A permutation approach to obtaining p-values also severely reduces the number of genes that are identified as differentially expressed for small experiments with a lot of differential expression. This limits our conclusions about the *Shrinking* approach (Cui et al., 2005), since for this approach it is the only suggested method to obtain p-values.
3. Combining an estimate of the overall variance with an estimate of the individual variance, such as is done for *Limma* (Smyth, 2004) and *Cyber-T* (Baldi and Long, 2001), appears to be very effective. Apparently such a regularization reduces the noise in the variance estimates in an effective manner. Because of the similarity of the results for these two approaches, and the much worse results for the smoothing approaches, it appears that for the *Cyber-T* approach the running average estimate of σ^2_{0ij} is in effect estimating an overall variance, rather than a local variance. In the analyses here, *Limma* performs slightly better than *Cyber-T*.
4. The *Pooled-t* approach proposed by Kooperberg et al. (2005), which borrows degrees of freedom from other experiments, performs equally well as *Limma* and *Cyber-T*. In fact, when the sample size is minimal ($n = 2$) it seems to perform slightly better. An obvious question concerns which experiments to combine. A small simulation study carried out by Kooperberg et al. (2005) suggests that there can be a fair degree of experiment-to-experiment variation without seriously inflating the Type I error. The fact that we were able to combine here information on experiments carried out on such diverse material as cell-lines and RNA harvested from fruit flies lends support to this conclusion.
5. The approaches to combining information here do not all perform equivalently. Methods which use only a (locally) smoothed estimate of the variance, such as *LPE* (Jain et al., 2003) and *Loess* (Huang and Pan, 2002), can give severely biased results by inflating the percentage of significant genes well beyond a pre-specified level α when in fact there are no differences between the two samples. For *Loess* this is evident at $\alpha = 1\%$ and $\alpha = 0.01\%$, for *LPE* it is only evident at $\alpha = 0.01\%$. However, due to multiple testing in microarray experiments very small significance levels are generally used, so it would seem better to avoid methods relying solely on smoothing. One reason for this bias might be that with the improved normalization methods now available, the relation between variance and expression level has been considerably reduced (see Section 8.4.4). Thus, locally averaging the variances will sometimes yield variances that are too large and sometimes yield variances that are too small. When the variance is too small there is a substantial chance of incorrectly identifying a gene as differentially expressed. Another, more fundamental reason is due to the experimental design itself. The *LPE* approach is more appropriate for technical replicates, for which the error distributions are closer to Gaussian. The error distribution for biological replicates, such as those we analyze here, will confound technical variability with heterogeneous biological variability, leading to the observed

bias. Lee, Cho, and O'Connell (Chapter 7 in this volume) provide additional detail along with methods to address these issues.

8.6 Appendix: Array preprocessing

For all arrays we carried out a graphical quality assessment, which indicated that all arrays were of good quality.

Two-color arrays. For the two-color arrays we first exclude all spots with a $\log_2$-expression ratio of less than 5 and spots whose background level was higher than the foreground level for either channel. This excludes about 11.5% of the spots, primarily those that do not hybridize well. In particular, of the 13,440 spots on the arrays, 1,296 are excluded on all 36 arrays: of the remaining spots only about 2% are excluded. We then subtract the background and use a print-tip loess correction (Yang et al., 2002), carried out using the `limma` function `normalizeWithinArrays()` with defaults. Any spot that had at least two estimates for a particular experiment was included in the analysis.

One-color arrays. Gene expression is quantified using RMA (Irizarry et al., 2003b) on all arrays simultaneously. We also carried out the analyses using $\log_2$ of the MAS 5.0 summary (Affymetrix, 2002; see also Section 1.4) and again using RMA separately within each experiment. In both cases, the results are very similar to those reported in this chapter.

Acknowledgments

Charles Kooperberg was supported in part by NIH grants CA 74841, CA 53996, and HL 74745. Aaron Aragaki was supported in part by NIH grant CA 74841. Charles C. Carey was supported in part by NIH interdiciplinary training grant 2T32CA80416-06. Suzannah Rutherford was supported in part by NIH grant GM 06673. The authors thank Andy Strand, Jim Olson, and the HDAG group for allowing us to use the one-color data.