CHAPTER 1

Significance testing for small microarray experiments

Charles Kooperberg, Aaron Aragaki, Charles C. Carey, and Suzannah Rutherford
Fred Hutchinson Cancer Research Center, PO Box 19024, Seattle, WA 98109

1.1 Introduction

When there are many degrees of freedom it is sometimes less critical which significance test is carried out, as most analysis will give approximately the same result. However, when there are few degrees of freedom the choice of which significance test is being used can have a strong effect on the results of an analysis. Unfortunately, this is often the case for microarray experiments, as research laboratories often perform such experiments with only a few (say less than five) repeats, Reasons for the small number of repeats include availability of specimens and economics. Kooperberg et al. (2005) compared several approaches to significance testing for experiments with a small number of oligonucleotide (one-color) arrays. In this paper we summarize the results from that analysis, include a couple of additional methods, and describe a similar comparison for methods of carrying out significance testing for two-color (red-green) arrays.

The limited number of repeats, together with the large variability that even the best microarray platforms have, make 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 Bonferoni correction for 10000 genes requires a T-statistic of 20.6 or more for significance. The lack of degrees of freedom is really what drives the extremely large significance threshold for T-statistics: the same α and Bonferoni correction for 20 arrays requires a T-statistic of 6.3 or more while a normal distribution only requires a Z-statistic of 4.6 or more, on the other hand reducing the number of genes of interest on the original array from 10000 to 500 only reduces the required T-statistic to 11.3.

Nonparametric (Wilcoxon) or permutation tests are no easy way out. For example, for an experiment with k two-color (spotted) arrays, a P-value for a permutation test can be no smaller than 2^{-k}; if we want a two-sided test with $\alpha = 0.05$ and a Bonferoni correction for 10000 genes, we need k to be at least 19. Reducing the number of genes to 500 reduces the minimum k to 15. Similarly, for a one-color (oligonucleotide) array the P-value for a permutation tests with k cases and k controls a P-value cannot be smaller than $\binom{2k}{k}$; so for a two-sided test with $\alpha = 0.05$ and a Bonferoni correction for 10000 genes, we need at least $2k = 22$ arrays. Reducing the number of genes to 500 reduces the minimum number of arrays to 18.

As permutation tests are not going to help us, we need to obtain a better estimate for the residual variance to overcome the lack of repeats. There are two obvious choices: we can combine different genes in the same experiment or we can combine different experiments, if similar experiments were carried out. When genes are combined we can either choose to combine those genes for which the general expression level is similar as do, for example, Huang & Pan (2002) and Jain et al. (2003), or we can choose to combine all genes. An alternative approach to obtain more power with small experiments is to add a stabilizing constant to the estimate of the variance for each gene or to use some (Bayesian) model for the expression levels. SAM (Tusher, Tibshirani & Chu, 2001) is a methodology that adds a constant to the estimate the variance. The approaches by Baldi & Long (2001), Lönnstedt & Speed (2002), Smyth (2004), and Cui et al. (2005) are four related (empirical) Bayesian approaches. Wright & Simon (2002) discuss a closely related frequentist approach.

In this paper we do not control for multiple comparisons. In practice, when one carries out tests for many thousands of genes simultaneously, a multiple comparisons correction or a correction of the false discovery (FDR) rate is essential. See Dudoit, Shaffer & Boldrick (2003) for an extensive overview of multiple comparisons corrections. While several of these proposals use permutation arguments to correct for multiple comparisons, permutation typically either requires a substantial number of replicates (that are not available in small experiments), or they require implicit assumptions about similarities in the variational properties of different genes. In either scenario, we believe that only well calibrated marginal P-values are going to yield good multiple comparison corrected P-values.

P-values have the advantage that there are well established measures such as Type I error and power that can be used to judge the performance of a test. The FDR (Benjamini & Hochberg, 1995) does not have such a simple measure, to check whether estimates of the FDR are accurate on a single experiment In addition, just like for multiple comparison procedures, there are procedures to approximate the FDR from P-values.

1.2 Methods

Most of the methods that we compare in this paper can be used either for one-color (oligonucleotide) arrays or for two-color (spotted) arrays. We assume that the arrays have been properly normalized; see Section 1.7 for how we normalized our arrays.

1.2.1 Notation

Two-color spotted arrays For each gene and each two-color array we have an expression ratio x^m_{ijl} summarizing the (log-)expression ratio between experimental conditions $k = 1$ and $k = 2$ (that may be different between experiments) for gene $i = 1, \dots, n$ in experiment $j = 1, \dots J$ on replicate array $l = 1, \dots, L_j$. For each gene on each array we also have an estimate of the overall expression x^a_{ijl}, typically this will be the (geometric) average of the normalized expression for both channels of the array. Unless there is confusion we will write x_{ijl} instead of x^m_{ijl} for the log-expression ratios.

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

One-color oligonucleotide arrays Similarly, for each gene and each one-color array we have a (log-)expression x_{ijkl}, for experimental conditions $k = 1$ and $k = 2$, for gene $i = 1, \dots, n$ in experiment $j = 1, \dots J$ on replicate array $l = 1, \dots, L_{jk}$.

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

1.2.2 Significance Tests

All significance tests that we consider in this paper 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. Here $\widetilde{\sigma}_{ij}$ is an estimate of the variance of x_{ijl}. The methods that we discuss differ primarily in how the estimate $\widetilde{\sigma}_{ij}$ is obtained. The traditional test statistics estimate $\widetilde{\sigma}_{ij}$ uses only the data on gene i and experiment j. The approaches that inflate the variance and those that combine genes also use data on genes i^*, $i^* \neq i$; implicitly to estimate hyper-parameters for the empirical Bayes approach that inflates the variance, or explicitly to smooth the estimates for $\widetilde{\sigma}_{ij}$. Finally the approaches that combine experiments use data on experiments j^*, $j^* \neq j$. Most of the methods below have a proper reference distribution, but alternatively significance levels can be obtained using permutations (see Section 1.2.3); in fact, some of the authors recommend permutations as the method to obtain P-values.

Below we describe the test-statistics we are including in our comparison. We provide details for the two-color arrays, modifications for one-color arrays are indicated.

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. 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. In Kooperberg et al. (2005) it was shown that this approach has almost no power for small sample sizes, 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 a method they call "Local Pooled Error test" (LPE). As described in this paper, LPE only is applicable to one-color arrays. In their approach, let $\widehat{\sigma}_{ijk}$ 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 $\widehat{\sigma}_{ijk}$ 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. As the smoothing spline that is used effectively involves averaging a large number of genes, the authors use a normal reference distribution. In our study we have used the implementation by the authors, available in the R-package (Ihaka & Gentleman 1996) LPE, which is available from CRAN/Bioconductor* Since the method averages the variance separately for two conditions, it is currently only available for one-color arrays, where both experimental conditions are measured separately.

Loess Huang & Pan (2002) make several related proposals. The main difference between their approach and the approach by Jain et al. (2003) is that they first compute $\widehat{\sigma}_{ij}$ 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

* CRAN: The Comprehensive R Archive Network; see `http://www.r-project.org`.

show that, not unexpectedly, for the null-model a normal reference distribution is appropriate. We reimplemented their approach using a `loess` smoother.

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 some (hyper-)parameters, that can be used to stabilize the variance estimate. There are a variety of such methods with different motivations: ad-hoc (e.g. SAM, Tusher, Tibshirani & Chu 2001), using an (empirical) Bayes argument (e.g. Baldi & Long 2001, Lönnstedt & Speed 2002, Smyth 2004), a James-Stein type estimator (Cui et al. 2005), or a frequentist approach (Wright & Simon 2003).

The first three approaches that we discuss combine the sample variance $\widehat{\sigma}^2_{ij}$ with another estimate σ_{0ij} that has d_{ij} degrees of freedom, yielding a variance estimate of

$$\widetilde{\sigma}^2_{ij} = \frac{d_{ij}\sigma^2_{0ij} + (L_j - 1)\widehat{\sigma}^2_{ij}}{L_j + d_{ij} - 1}, \tag{1.1}$$

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

Cyber-T The Cyber-T approach of Baldi & Long (2001) is motivated as a fully Bayesian procedure. However as implemented in practice (see Section 5 of Baldi & Long 2001) the test is carried out using a T-test on (for two-color arrays) $L_j + \nu_0 - 1$ degrees of freedom, and an estimate of the variance (compare 1.1) of

$$\widetilde{\sigma}^2_{ij} = \frac{\nu_0\sigma^2_{0ij} + (L_j - 1)\widehat{\sigma}^2_{ij}}{L_j + \nu_0 - 1}, \tag{1.2}$$

where σ^2_{0ij} 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^a_{ij}. Thus the Cyber-T approach uses the average of a smoothed variance (like **LPE** and **Loess**, only using another smoother) with the regular variance of the **T-statistic**. A non-Bayesian interpretation of Cyber-T is thus that it combines a smoothed estimated (as in **Loess** and **LPE**) with a traditional estimate from the **T-test**.
We used the defaults $\nu_0 = 10$ and the window width $w = 101$ from the R-software available on `http://visitor.ics.uci.edu/genex/cybert`. Note that the paper of Baldi and Long mentions another default of $\nu_0 = 10 - L_j$.

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

$$\frac{1}{\sigma^2_{ij}} \sim \frac{1}{d_{0j}s^2_{0j}}\chi^2_{d_{0j}}.$$

(We include the index j for the parameters of the prior, as they may be different for different experiments $j = 1, \ldots, J$.) 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. Using methods of moments estimators estimates d_{0j}, s_{0j}^2, and a few other parameters are obtained. An inflated variance

$$\widetilde{\sigma}_{ij}^2 = \frac{d_{0j} s_{0j}^2 + (L_j - 1)\widehat{\sigma}_{ij}^2}{L_j + d_{0j} - 1}, \tag{1.3}$$

(compare 1.2) is used for a "moderated T-test" with $d_{0j} + L_j - 1$ degrees of freedom. Thus, a main difference between the approach of Smyth (2004) and the approach of Baldi & Long (2001) is that Limma uses one single estimate for the prior variance (s_{0j}^2) for all genes and it estimates the prior degrees of freedom d_{0j} based on the data, while the latter 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 available from the Bioconductor package Limma. We used Limma with the default options.

RVM The Random Variance Model (RVM) of Wright & Simon (2003) inflate the variance similar to Baldi & Long (2001) and Smyth (2004), and obtain a model similar to (1.1). They assume an inverse Gamma model for σ^2, and estimate the two parameters from this model using the method of maximum likelihood. Implementation of their approach would require estimating of two parameters of an F-distribution. We do not include RVM this method in our comparisons, as we could not locate publicly available software.

Shrinking Cui & 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 $\widehat{\sigma}_{ij}^2$ towards the mean variance $\sum_i \sigma_{ij}^2 / I$, 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). Note that the authors of this method recommend a permutation approach (see Section 1.2.3) to obtaining P-values. We still include this approach without permutations using a normal reference distribution, as well as using permutation P-values.

Methods combining experiments

Instead of combining different genes *within* one experiment, we can also combine expression levels of the same gene *between* experiments. This would potentially be useful if we have several smaller experiments, and it is thus 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.

For most of the other methods that we discussed it is, in principle also possible to pool different experiments in obtaining a single variance estimates. As all these methods already regularize the estimates for σ in some way, pooling typically has no effect, and the corresponding method behaves similar to the "parent" method, as was confirmed for the **Loess** approach in Kooperberg et al. (2005) and for **Limma** in unpublished results.

Note that methods whose implementation allows for general design matrices (e.g. **Limma**) can yield pooled estimates by setting up an appropriate design matrix and testing appropriate contrasts.

1.2.3 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 mth permutation. We can use

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

as an estimate of the P-value corresponding to T_i. If L is larger than, say, 8 we may want to sample permutations to save computing time; in this paper that is not an issue.

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 appears. When a parametric distribution is used the stronger assumption, that the distributions of each T_i under the null-hypothesis are the same as a particular parametric distribution, is made. The second assumption is much more severe, and it will lead to conservative P-values when in fact there are a substantial number of differentially expressed genes (Storey & Tibshirani 2003).

For one-color (oligonucleotide) 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.

Table 1.1 *Organization of the two-color (spotted) data for our analysis. Experiments whose code start with a D are expected to have differences between both groups, while those starting with an S are repeats, the digit "2" refers to the two-color (spotted) array type. 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 one	sample two	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

1.3 Data

For our analysis we use two sets of data. One comes from a one-color experiment, and is part of the data that was also used in Kooperberg et al. (2005), the other comes from a not yet published study on Drosophila.

The two-color experimental data that we use come from a series of spotted microarrays of *Drosophila melanogaster* that were grown in Suzannah Rutherford's lab at the Fred Hutchinson Cancer Research Center. The arrays are part of a larger set of experiments whose results have not yet been reported. The subset of arrays that we compare here include some experiments that are self-to-self hybridizations, and some experiments where both samples are genetically different, see Table 1.1. 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 the power of the tests.

For the SAM samples RNA from a large number of flies that were genetical identical, other than some being male and some being female, was combined and the RNA for the arrays was taken out of this large pool. For the D-recomb 304, D-recomb-220, D-pure, E-pure, and F-pure lines for each array samples from 15-30 flies that were genetical identical, other than some being male and some being female, was combined. In addition we included four unrelated Drosophila cell line arrays. We organized the experiments so that all experiments are "dye swapped": i.e. half of the arrays have sample one on the red channel, the other half have sample two on the red channel. There are 13,440 spots (genes) on each array.

Table 1.2 *Organization of the one-color (Affymetrix) data for our analysis. HD: Huntington's Disease mouse, WT: wildtype mouse. Experiments whose code start with a D are expected to have differences between both groups, while those starting with an S are repeats, the digit "1" refers to the one-color (Affymetrix) array type.*

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

One-color experimental data was obtained using Affymetrix Mu 11K-A microarrays generated for a series of experiments on Huntington's Disease mouse models. The results of these experiments were reported as a series of related papers (Chan et al. 2002; Luthi-Carter et al. 2002a; Luthi-Carter et al. 2002b). For this analysis we compare cerebellar gene expression in similarly aged mice carrying a wildtype or mutant form of the Huntington's gene. Every comparison reported in Chan et al. (2002), Luthi-Carter et al. (2002a), and Luthi-Carter et al. (2002b) showed some differentially expressed genes, although the amount of differentiation differed considerably between the experiments. For each of the experiments both groups had between 2 and 5 mice. Thus, all our repeats use different samples (sometimes referred to as "biological repeats") and are not repeat arrays using the same samples (sometimes refereed to as "technical repeats"), that could be expected to vary less. There are 6,595 probe sets (genes) on each array.

The experiments listed in Table 1.2 are the seven experiments comparing cerebellar tissue used in Kooperberg et al. (2005); the six experiments using striatum tissue used in that paper are not used here. As for the two-color experiments, some experiments are intended to establish that the tests have the right size and others are intended to establish the power of the tests.

1.4 Results

We analyze the experiments listed in Section 1.3 using the analysis methods described in Section 1.2.2. 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): at most $\alpha\%$ significant genes when tested at significance level α.

Typically 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) that were too close to background (see Section 1.7) . 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 (incorrect) assume independence for both array types 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.

1.4.1 Bandwidth selection for smoothers

Three 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.

In Table 1.3 we summarize the results for the two-color experiment for the **Loess** approach. The parameter `span` for the `loess()` function in R is approximately linear in the bandwidth for a local linear smoother. From this table we note 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 will increasingly lead to numerical problems in regions where there is less data.

We note that 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 that are called significant is much too large. The same was concluded in Kooperberg et al. (2005) for the one-color arrays.

In the remainder of our comparisons we use a `span` of 0.1, which yielded 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 will use **Cyber-T** and **LPE** with their default values.

1.4.2 Comparison of methods

In Tables 1.4 and 1.5 we show the results for seven of the methods described in Section 1.2.2 when applied to the two-color and one-color data described in Section 1.3, respectively. Results for the **LPE** method are not available for the two-color data. Cui et al. 2005 recommends permutations to obtain P-values for the **Shrinking** approach, as in Tables 1.6 and 1.7 and Figure 1.3 and 1.4. In Tables 1.4 and 1.5 and Figure 1.1 and 1.2 we use a normal reference distribution; which distribution is used has a substantial impact on the results.

Table 1.3 *Performance of the* **Loess** *approach for various values of the bandwidth (*`span`*) parameter for the two-color experiments. We report the percentage of genes that are 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

In Figure 1.1 we give a graphical display of how well these methods adhere to the significance levels, and in Figure 1.2 we display power. These figures are probability-probability plots on a logit-scale. That is, for a particular 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 we have N P-values. We plot the sorted P-values (horizontal) against $(1, \dots, n)/(N + 1)$. When the experiments that we consider are self-versus-self comparisons we would like these plots to 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 hurt us when we use the same method to analyze data where some genes are differentially expressed. Second, for groups of experiments where there is a difference between both samples we want the most horizontal curves, among the methods that did not generate a substantial number of false positives for the repeat experiments.

From Figure 1.1 we see that the **Loess** and **LPE** approach identify substantially more

Figure 1.1 *Performance of the various approaches to significance testing using an explicit reference distribution for small microarray experiments for the combined two-color and one-color self-versus-self experiments. For unbiased methods the curves should follow the identity line.*

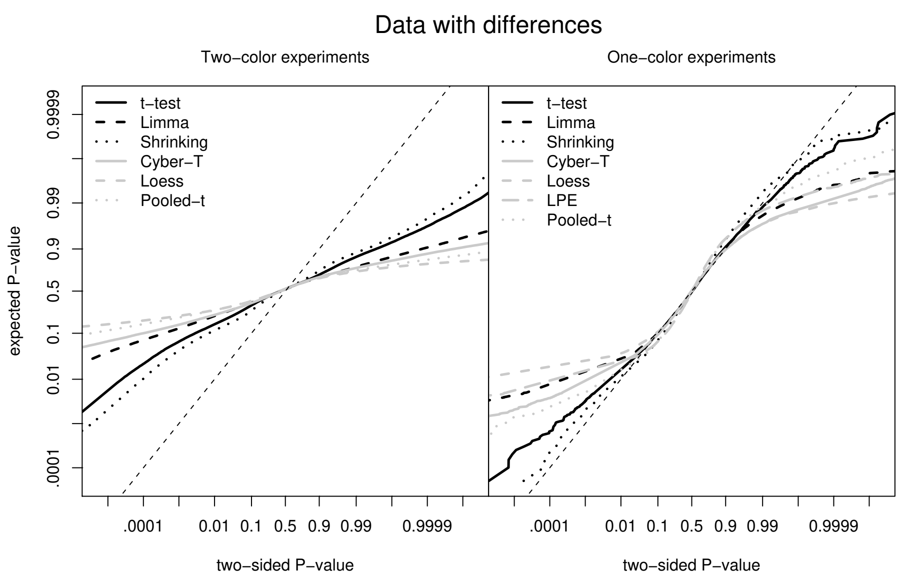

Figure 1.2 *Performance of the various approaches to significance testing using an explicit reference distribution for small microarray experiments for the combined two-color and one-color experiments that involve different samples. More horizontal curves correspond to more powerful methods.*

differentially expressed genes than the nominal levels for the experiments where in fact the two samples being compared are repeats. The **Cyber-T** approach shows a mild number of increases, and none of the other approaches shows serious bias. For both groups of experiments a normal reference distribution for the **Shrinking** approach appears too conservative.

Table 1.4 elaborates on this. At a significance level of $\alpha = 1\%$ only the **Loess** method shows a substantial bias, and it does that for five out of eight data sets. For microarray experiments the more stringent level $\alpha = 0.01\%$ is very relevant, as multiple comparisons corrections often will imply selecting genes at low significance levels. We note that the **Loess** again 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; in particular we notice that the bias is only substantial for one dataset (two-color experiment S2.2). The excess percentage of significant genes for the **Pooled-T** approach is minimal, and could just be due to chance.

From Figure 1.2 we note 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 actual 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. We do note from this figure though that the ordering of the methods is largely unchanged, suggesting that since our conclusions remain the same for two dramatically different experiments (different technologies, different amounts of differential genes) they are likely fairly robust and may well generalize to many other situations.

For both the two-color and the one-color experiments the **Loess** approach is the most powerful. This is not a surprise, since the method does not maintain significance levels for the experiments where both samples are repeats. Similarly, we are not surprised that the **LPE** method is quite powerful for the one-color experiments. This method also did not maintain significance levels for the experiments where both samples are repeats. Among the remaining methods, we note that the **Pooled-T** approach performs best for the two-color experiments, followed by the **Cyber-T** and **Limma** approach, while for the one-color experiments the **Cyber-T** and **Limma** approach seem slightly more powerful than the **Pooled-T** approach.

Table 1.5 confirms all these conclusions. Interestingly for the D2.x (two-color) experiments we notice that for those experiments with two arrays (D2.1, D2.2, and D2.3) the **Pooled-T** approach is particularly more powerful. Maybe this is not surprising: the borrowing of degrees of freedom between experiments, as the **Pooled-T** approach is doing, is particularly useful when the number of degrees of freedom is small.

Table 1.4 *Percentage of differentially expressed genes using various approaches to significance testing using an explicit reference distribution for small microarray experiments for the individual two-color and one-color self-versus-self experiments 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 1.5 *Percentage of differentially expressed genes using various approaches to significance testing using an explicits reference distribution for small microarray experiments for the individual two-color and one-color experiments that involve different samples at significance levels $\alpha = 1\%$ and $\alpha = 0.01\%$. The larger the percentage of differentially expressed genes, the more powerful a method is.*

$\alpha = 1\%$	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

1.4.3 Permutation P-values

As detailed in Section 1.2.3, an alternative approach to obtaining P-values is a permutation approach in which the test statistics for all genes are combined. In Figure 1.3 we give a graphical display of how well each of the methods adhere to the significance levels when P-values are determined using such an approach, and in Figure 1.4 we display power for these situations. We do not show permutation results for the **Pooled-T** approach: since this procedure combines arrays from different experiments a permutation procedure is less standard, besides that the results using a T-distribution already give satisfactory results.

The displays in Figures 1.3 and 1.4 are organized similar to Figures 1.1 and 1.2. We notice that the permutation approach for computing P-values yields approximately unbiased results for all approaches as all curves in Figure 1.3 follow the diagonal. However, as expected, the permutation approach reduces power for any of the approaches using randomization. In Figure 1.4 we note that the procedures based on permutation are considerably less powerful than the procedures that do not use permutation (as shown in Figure 1.2). In particular, we notice that the curves in Figure 1.4 all stay within a "band" of the diagonal. This is in fact 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 differentially expressed at any particular significance level thanks to the experimental design. This is explained in detail below in the discussion of Table 1.7.

Tables 1.6 and 1.7 for the permutation based procedures are organized similar to Tables 1.4 and 1.5 for the procedures using a reference distribution. From these tables we draw the same conclusions as from Figures 1.3 and 1.4: while the permutation approach does control the significance level α appropriately, it limits the power. We note from these tables that no methods and no data sets are exceptions. The part of Table 1.7 for the two-color (D2.x) experiments with different samples clearly illustrate an artifact of the permutation approach. As we have seen before, the D2.x experiments have very many differentially expressed genes (see Table 1.5). But in Table 1.7 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. Let's focuss on experimant D2.4. This is an experiment with 4 arrays. There are thus 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 1.5 suggest), and these 40% of the genes have very large test-statistics. There are about 10,000 genes on these arrays, thus 4,000 test-statistics are large, say larger than A. Now assume that among the 7 other permutations none of the test-statistics are larger than A. Then out of $8 \times 10,000 = 80,000$ test-statistics 4,000 are larger than A. However, at the $\alpha = 1\%$ level at most $0.01 \times 80,000 = 800$ can be called significant at $\alpha = 1\%$.

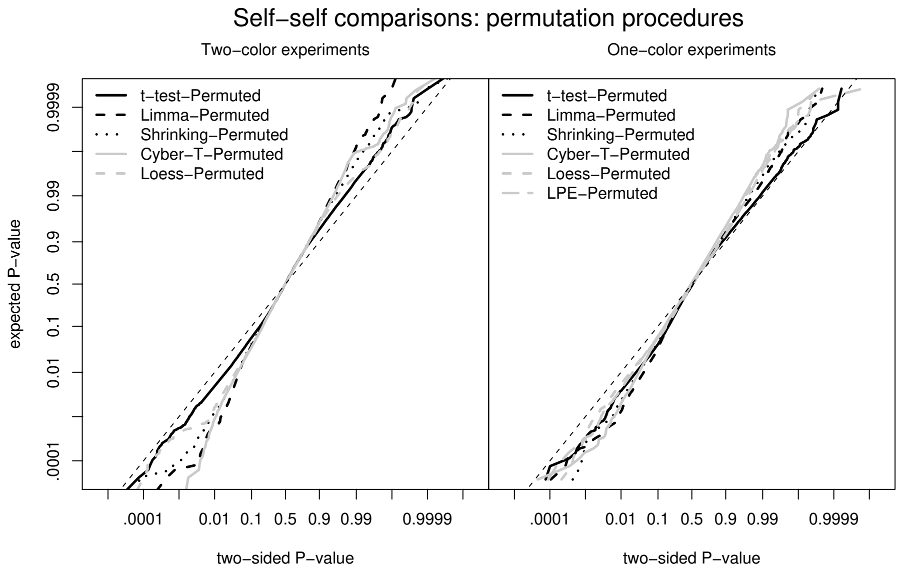

Figure 1.3 *Performance of the various approaches to significance testing using a permutation approach rather than a reference distribution for small microarray experiments for the combined two-color and one-color self-versus-self experiments. For unbiased methods the curves should follow the identity line.*

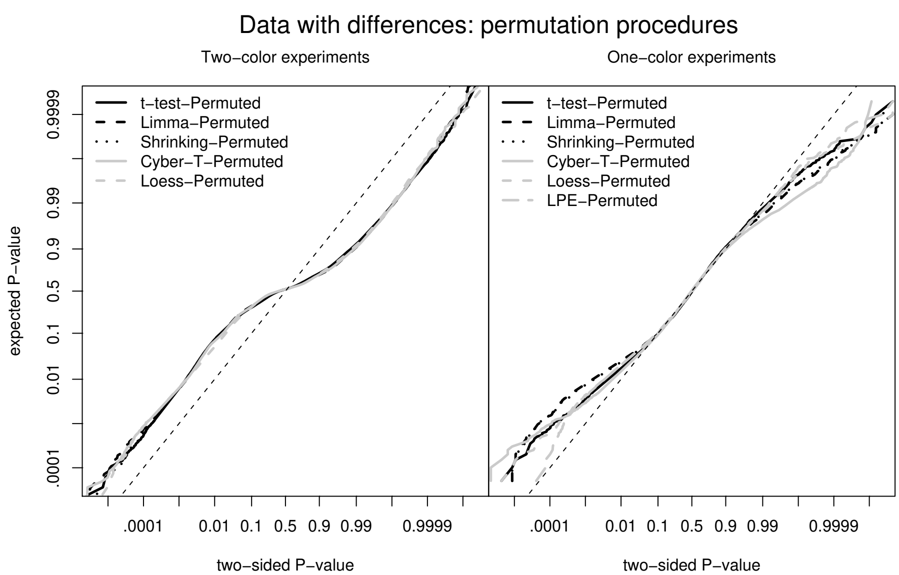

Figure 1.4 *Performance of the various approaches to significance testing using a permutation approach for small microarray experiments for the combined two-color and one-color experiments that involve different samples. More horizontal curves correspond to more powerful methods.*

Table 1.6 *Percentage of differentially expressed genes using various approaches to significance testing using a permutation approach rather than a reference distribution for small microarray experiments for the individual two-color and one-color self-versus-self experiments 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

Table 1.7 *Percentage of differentially expressed genes using various approaches to significance testing using a permutation approach rather than a reference distribution for small microarray experiments for the individual two-color and one-color experiments that involve different samples at significance levels $\alpha = 1\%$ and $\alpha = 0.01\%$. The larger the percentage of differentially expressed genes, the more powerful a method is.*

$\alpha = 1\%$	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

Which is 8%, rather than the 40% that are differentially expressed, of all the genes on the array. (In fact the percentage is slightly lower as a few rare permuted genes also have large statistics.) We could choose to ignore the "original" permutation in getting the percentiles of the permutation distribution, but this would violate the assumptions of exchangeability under the null-hypothesis of no differential expression. When the number of arrays increases, or when the number of differentially expressed genes is much smaller, this artifact clearly disappears.

1.5 Discussion

The choice of significance test in microarray experiments with low replication can dramatically influence the results. For both one-color and two-color arrays we set up our experiments so that we could both judge which approaches yield approximately unbiased P-values when the experimental conditions are identical, and which approaches are most powerful when both conditions differ. We focused on P-values, rather than for example the FDR, as we believe that a "good" P-value will yield a "good" multiple comparisons correction, and a multiple comparisons adjustment by itself can not save a procedure that yields badly calibrated P-values.

The two groups of experiments that we considered differed in another aspect besides technology: our one-color experiments had a modest number of differentially expressed genes, while our two-color experiments had many such genes. Given the difference between the two groups of experiments the similarity in results was striking.

Our main conclusions are:

- The **T-test** has almost no power when the sample size is small. When there are less than, say, six repeat arrays some of the alternative solutions are much more powerful. Kooperberg et al. (2005) concluded that the lack of power is even more extreme for the Welch statistic.
- 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 & Long 2001) appear very effective. Apparently such a regularization reduces the noise in the variance estimates effectively. Because of the similarity of the results for these two approach, and the much worse results for the smoothing approaches, we hypothesize that for the **Cyber-T** approach the running average estimate of σ_{0ij} is effectively estimating an overall variance, rather than a local variance. In our experiments **Limma** performed slightly better than **Cyber-T**.
- An approach which borrows degrees of freedom from other experiments **Pooled-T**, first proposed in Kooperberg et al. (2005), performs equally well as the **Limma** and **Cyber-T** approach. In fact, when the sample size is real small ($n = 2$) it seems to perform slightly better. Obviously for this approach the main question is "what to combine". In Kooperberg et al. (2005) a small simulation study was carried out suggesting that there can be a reasonable amount of experiment-to-experiment variation without seriously inflating the type-1 error. The fact that we

can without much problem combine cell-line experiments with RNA harvested from fruit-flies (as was done for the two-color experiments in this paper) confirms that conclusion.

- Methods which solely use a smoothed estimate of the variance, such as the **LPE** approach (Jain et al. 2003) and the **Loess** approach (inspired by Huang & 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 the **Loess** approach this was evident at $\alpha = 1\%$ and $\alpha = 0.01\%$, for the **LPE** approach it was only evident at $\alpha = 0.01\%$. However, since for microarray experiments often multiple comparisons corrections are carried out very small significance levels are in fact used, we would want to avoid methods that solely use smoothing approaches. A reason for the bias because of smoothing the variance may be due to the fact that with the normalization methods developed in recent years (see Section 1.7) the relation between variance and expression level has been considerably reduced.
  In particular, in Figure 1.5 for an individual two-color array and one of the two-color experiments and in Figure 1.6 for one of the one-color experiments we show the relation between the difference between the two signals (left side of Figure 1.5) or the variance and the average signal (other panels). As can be seen, the relation between average signal and variance is minimal, and 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). 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.
- A permutation approach to obtaining P-values severely reduces the number of genes that are identified as differentially expressed for experiments with a lot of differential expression. This limits our conclusions about the **Shrinking** approach (Cui et al. 2005), as for this approach it is the only suggested method to obtain P-values.

All approaches that we studied are either available in R-packages available from CRAN or Bioconductor, or are easily implemented in R code.

1.6 References

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

Y. Benjamini and T. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing", Journal of the Royal Statistical Society series B, 57, 289–300, 1995.

E. Y. Chan, R. Luthi-Carter, A. Strand, S. M. Solana, S. A. Hanson, M. M. DeJohn, C. Kooperberg, K. O. Chase, A. B. Young, B. R. Leavitt, J. J. Cha, N. Aronin, M. R. Hayden, and J. M.

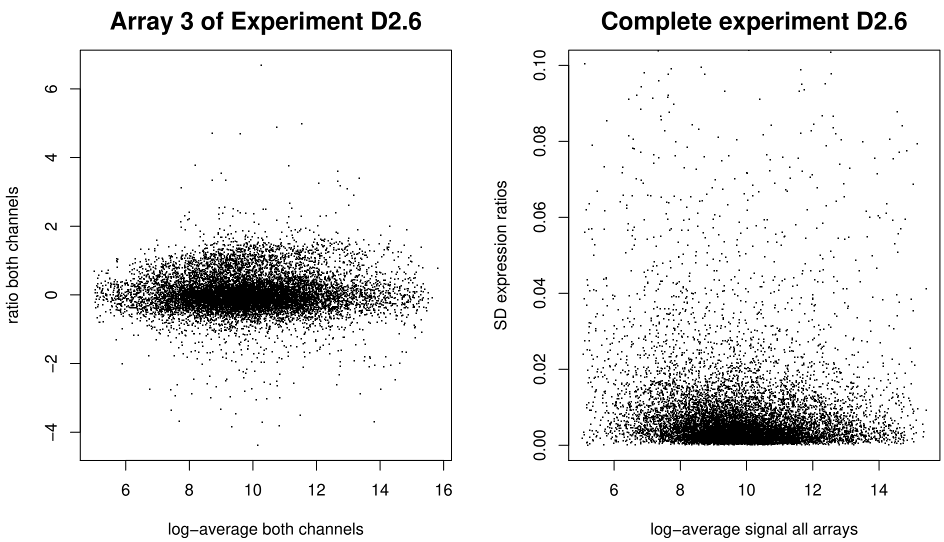

Figure 1.5 *Relation between log expression ratio and average log expression for one normalized two-color array, and the standard deviation of log expression ratios with the average log expression ratio for all arrays from that experiment.*

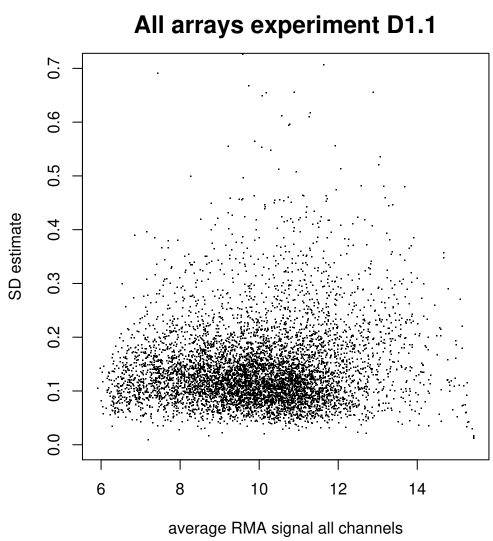

Figure 1.6 *Relation between the residual standard deviation and the average RMA normalized expression for one of the one-color experiments.*

Olson, “Increased huntingtin protein length reduces the number of polyglutamine-induced gene expression changes in mouse models of Huntington’s disease”, Human Molecular Genetics 11, 1939–1951, 2002.

X. Cui and G. A. Churchill, “Statistical tests for differential expression in cDNA microarray experiments”, Genome Biology 4, 210, 2003.

X. Cui, J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill, “Improved statistical tests for differential gene expression by shrinking variance components estimates”, Biostatistics, 6, 59–75, 2005.

S. Dudoit, J. P. Shaffer, and J. C. Boldrick, “Multiple hypothesis testing in microarray experiments”, Statistical Science, 18, 71–103, 2003.

X. Huang and W. Pan, “Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays”, Functional and Integrative Genomics, 2, 126–183, 2002.

R. Ihaka and R. Gentleman, “R: A Language for Data Analysis and Graphics”, Journal of Computational and Graphical Statistics, 4, 299–314, 1996.

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

N. Jain, J. Thatte, T. Braciale, K. Ley, M. O’Connell, and J. K. Lee, “Local pooled error test for identifying differentially expressed genes with a small number of replicated microarrays”, Bioinformatics, 19, 1945–1951, 2003.

C. Kooperberg, A. Aragaki, A. D. Strand, and J. M. Olson, “Significance testing for small sample microarray experiments”, Statistics in Medicine, in press, 2005.

I. Lönnstedt and T. P. Speed, “Replicated microarray data”, Statistica Sinica, 12, 31–46, 2002.

R. Luthi-Carter, S. A. Hanson, A. D. Strand, D. A. Bergstrom, W. Chun, N. Peters, A. M. Woods, E. Y. Chan, C. Kooperberg, A. B. Young, S. J. Tapscott, and J. M. Olson, “Dysregulation of gene expression in the R6/2 model of polyglutamine disease: parallel changes in muscle and brain”, Human Molecular Genetics, 11, 1911-1926, 2002a.

R. Luthi-Carter, A. D. Strand, S. A. Hanson, C. Kooperberg, G. Schilling, A. LaSpada, D. Merry, A. B. Young, C. A. Ross, D. R. Borchelt, and J. M. Olson, “Polyglutamine and transcription: gene expression changes shared by DRPLA and Huntington’s disease mouse models reveal context-independent effects”, Human Molecular Genetics, 11, 1927-1937, 2002b

G. K. Smyth, “Linear models and empirical Bayes methods for assessing differential expression in microarray experiments”, Statistical Applications in Genetics and Molecular Biology, 3 (1), 3, 2004.

J. D. Storey and R. Tibshirani, “Statistical significance for genome wide studies”, Proceedings of the National Academy of Sciences USA, 100, 9440–9445, 2003.

V. G. Tusher, R. Tibshirani, and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response”, Proceedings of the National Academy of Sciences USA, 98, 5116–5121, 2001.

B. L. Welch, “The significance of the difference between two means when the population variances are unequal”, Biometrika, 29, 350–362, 1938.

G. W. Wright and R. M. Simon, “A random variance model for detection of differential gene expression in small microarray experiments”, Bioinformatics, 19, 2448–2455, 2003.

1.7 Appendix: Normalization of arrays

Two-color arrays For the two-color arrays we first excluded all spots with a log base 2 expression 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 spots that do not hybridize well. In particular of the 13,440 spots on our arrays, 1,296 were excluded on all 36 arrays: of the remaining spots only about 2% were excluded. We then subtracted the background and used a print-tip loess correction using the Limma function `normalizeWithinArrays()` with defaults. Any spot that had at least two estimates for a particular experiment was included in our analysis. We employed various graphical QC tools, and felt that all arrays were of good quality.

One-color arrays For all methods we analyzed gene expressions that were normalized by the RMA algorithm of Irizarry et al. (2003). We also carried out the same analysis using the log of the MAS5 Average Difference summary and obtained essentially the same results. For RMA we normalized all arrays simultaneously; however when we analyzed each of the experiments separately, the results were again essentially the same. We employed various graphical QC tools, and felt that all arrays were of good quality.

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 grants GM 06673. The authors thank Andy Strand, Jim Olson, and the HDAG group for allowing us to use the one-color data.