Method

Open Access

Exploratory differential gene expression analysis in microarray experiments with no or limited replication

Alexander V Loguinov*, I Saira Mian† and Chris D Vulpe*

Addresses: *Department of Nutritional Sciences and Toxicology, University of California at Berkeley, Morgan Hall, Berkeley, CA 94720, USA. †Life Sciences Division, Lawrence Berkeley National Laboratory, Cyclotron Road, Berkeley, CA 94720, USA.

Correspondence: Alexander V Loguinov. E-mail: Avl53@aol.com. Chris D Vulpe. E-mail: vulpe@uclink4.berkeley.edu

Published: 1 March 2004

Genome ***Biology*** 2004, **5:**R18

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/3/R18

Received: 24 June 2003
Revised: 1 December 2003
Accepted: 11 December 2003

Abstract

We describe an exploratory, data-oriented approach for identifying candidates for differential gene expression in cDNA microarray experiments in terms of α-outliers and outlier regions, using simultaneous tolerance intervals relative to the line of equivalence (*Cy*5 = *Cy*3). We demonstrate the improved performance of our approach over existing single-slide methods using public datasets and simulation studies.

Background

Multiple studies validate the utility of cDNA microarrays for comparing relative mRNA transcript levels between different biological samples. Both the biological systems under study and the technology itself contribute to the variability within and between microarrays [1-11]. A fundamental question is determining which of the potentially tens of thousands of genes assayed have transcript levels that differ significantly in the two samples. Experimental designs utilizing many levels of replication improve the ability to identify differentially-expressed genes [2-12]. However, the vast majority of studies utilize no or limited replication due to practical considerations of cost and feasibility. Thus, statistical techniques are needed for cDNA microarray studies with constraints on replication. A common strategy is to equate differentially-expressed genes with those genes having a ratio of hybridization intensity values greater, or less, than some user-defined threshold [13,14], such as two-fold change.

We describe a new approach for identifying differentially expressed gene candidates in cDNA microarray experiments without replication or with limited replication. We illustrate its utility by applying it to published data and demonstrate its advantages over current approaches. Microarray datasets are comprised of pairs of processed fluorescent intensity values, background corrected and normalized, for each of the N genes on the microarray. We discuss a model for such data in which the $\log_2(Cy5)$ and $\log_2(Cy3)$ values are linearly related and are samples drawn from a bivariate normal population 'contaminated' with outliers (see detailed definitions of the outlier-generating model in Methods) and possibly distorted due to heteroscedasticity. In a contaminated bivariate normal distribution, the main body of data is a sample from a bivariate normal distribution and constitutes the regular observations. The dataset also contains non-regular observations, 'outliers' or 'contaminants', which represent systematic deviations that, as we describe below, are candidates for differential expression. We check these underlying assumptions by applying exploratory data analysis tools (scatter plots with tolerance ellipses, Quantile-Quantile normal plots (QQNPs) with simulation envelopes and boxplots for residuals) to simulated and empirical datasets.

We formulate our method in terms of an α-outlier-generating model and outlier regions [15,16]. In a scatter plot of suitably normalized $\log_2(Cy5)$ versus $\log_2(Cy3)$ intensity values the

majority of data points lie in the vicinity of the line of equivalence (*Cy5* = *Cy3*). The line of equivalence can be estimated using a robust linear regression estimator and normalizing data by the regression fit making slope = 1 and intercept = 0. We subtract the fit from data to compute residuals [$\log_2(Cy5)$ - $\log_2(Cy3)$] which represent the vertical distances from the line of equivalence to the data points and are equivalent to the log-transformed ratios $\log_2(Cy5/Cy3)$. After these steps, we apply robust scatter plot smoothers to quantify and take into account the distortion of the data, if any, by heteroscedasticity. Data points far from the line of equivalence, 'outliers', are considered to be of greatest interest since they correspond to genes having noticeably different hybridization intensity values. An outlier could represent one of five circumstances: a gene with higher individual variability than the majority of genes; an outlier by chance; a sporadic technical or biological outlier; a systematic technical outlier (due to, for example, heteroscedasticity); or a systematic biological outlier due to differential expression. We assume that the further away from the line of equivalence an outlier is located, the more likely that it is genuinely 'up-' or 'down-regulated'. We compare our approach with some existing single-slide methods [13,14,17-20] and demonstrate that it works well in practice.

Results

We examined 10 published cDNA microarray experiments that compared 6,295 transcript levels in wild-type *Saccharomyces cerevisiae* and single gene deletion mutants pertinent to copper and iron metabolism [18]. The deleted genes were *mac1/YMR021C* (experiment number 96), *cin5/YOR028C* (26), *cup5/YEL027W* (36), *fre6/YLL051C* (64), *sod1/YJR104C* (162), *spf1/YEL031W* (163), *vma8/YEL051W* (189), *yap1/YML007W* (195), *yer033c* (214) and *ymr031c* (250).

Exploratory data analysis supports model for cDNA microarray data

We used exploratory data analysis tools to assess the assumptions underlying our method. We assume that biological and technical noise results in the majority of the measured expression levels changing randomly, independently, non-directionally and by a small amount. Thus, the $\log_2(Cy3)$ and $\log_2(Cy5)$ variables in cDNA microarray data should be linearly related and come from a contaminated bivariate normal distribution, possibly distorted due to heteroscedasticity. Using each tool, we compared the observed *mac1* data and datasets simulated as samples from a bivariate normal distribution with parameters corresponding to robust estimates of the location and variance-covariance matrix. Figure 1 shows concentration ellipses and QQNP for $\log_2(Cy5/Cy3)$ values for empirical and simulated datasets.

The $\log_2(Cy5)$ versus $\log_2(Cy3)$ scatter plots and concentration ellipses (Figure 1a,1b) provide a visual assessment of bivariate normality. The distribution of the *mac1* empirical data is similar to the simulated ('ideal') bivariate normal population except for the presence of strong outliers. The *mac1* data include significantly more unexpected events than might be expected for a sample from a bivariate normal population.

The QQNP for residuals ($\log_2(Cy5/Cy3)$) compares the quantiles of the empirical data with the quantiles of the standard normal distribution (Figures 1c,d). The *mac1* simulated data points lie along a straight line (the line for the standard normal distribution) except for some heavy tails due to finite sample size (Figure 1d). The empirical *mac1* data points (Figure 1c) also conform to a normal distribution except for longer tails (increased incidence of outliers and possible heteroscedasticity). Examination of the empirical data using exploratory data analysis tools supports our premise that the log-transformed channel intensities ($\log_2(Cy3)$ and $\log_2(Cy5)$) are linearly related and come from a contaminated bivariate normal distribution possibly distorted with heteroscedasticity.

The other nine datasets show a similar pattern (Figures 2,3,4,5,6,7,8,9,10). Figure 2 shows nine scatter plots with tolerance ellipses for the empirical log-transformed normalized channel intensities. There are strong bivariate outliers and differential gene expression candidates will be represented by *Y*-outliers. A data point which is an *X*-outlier or *Y*-*X*-outlier probably represents a technical gross error. Figure 3 represents scatter plots for simulated data produced using robust estimates of location and scale parameters for the corresponding empirical datasets. A 99.99% tolerance ellipse covers the simulated data points with no outliers. Figure 4 displays results after outlier removal from the empirical data using a simple cut-off and ignoring heteroscedasticity, if any. The majority of data points look like regular observations sampled from a bivariate normal population.

Ordinary QQNP results represented in Figures 5,6,7 are a different view of the data shown in Figures 2,3,4, comparing empirical, or simulated, quantiles with quantiles of the standard normal distribution. Outlying observations are all in the heavy tails. Figure 6 demonstrates the absence of strong outliers in the simulated data but heavy tails still persist due to finite sampling. Figure 7 shows that after the outlier removal, the main body of data ('regular observations') may be reasonably approximated with a normal distribution. Simulation envelopes for the QQNP support this conclusion, see text below.

Figures 8,9,10 illustrate the use of simulation envelopes. Clearly all simulated data points should be inside the envelopes as shown in Figure 9. Figure 10 confirms our expectations that after outlier removal the main body of regular data would be within the simulation envelopes. A hump-shaped deviation from the envelope limits in almost all the examples suggests a systematic technical error (the non-linear local bias is very reproducible and may be seen in the majority of

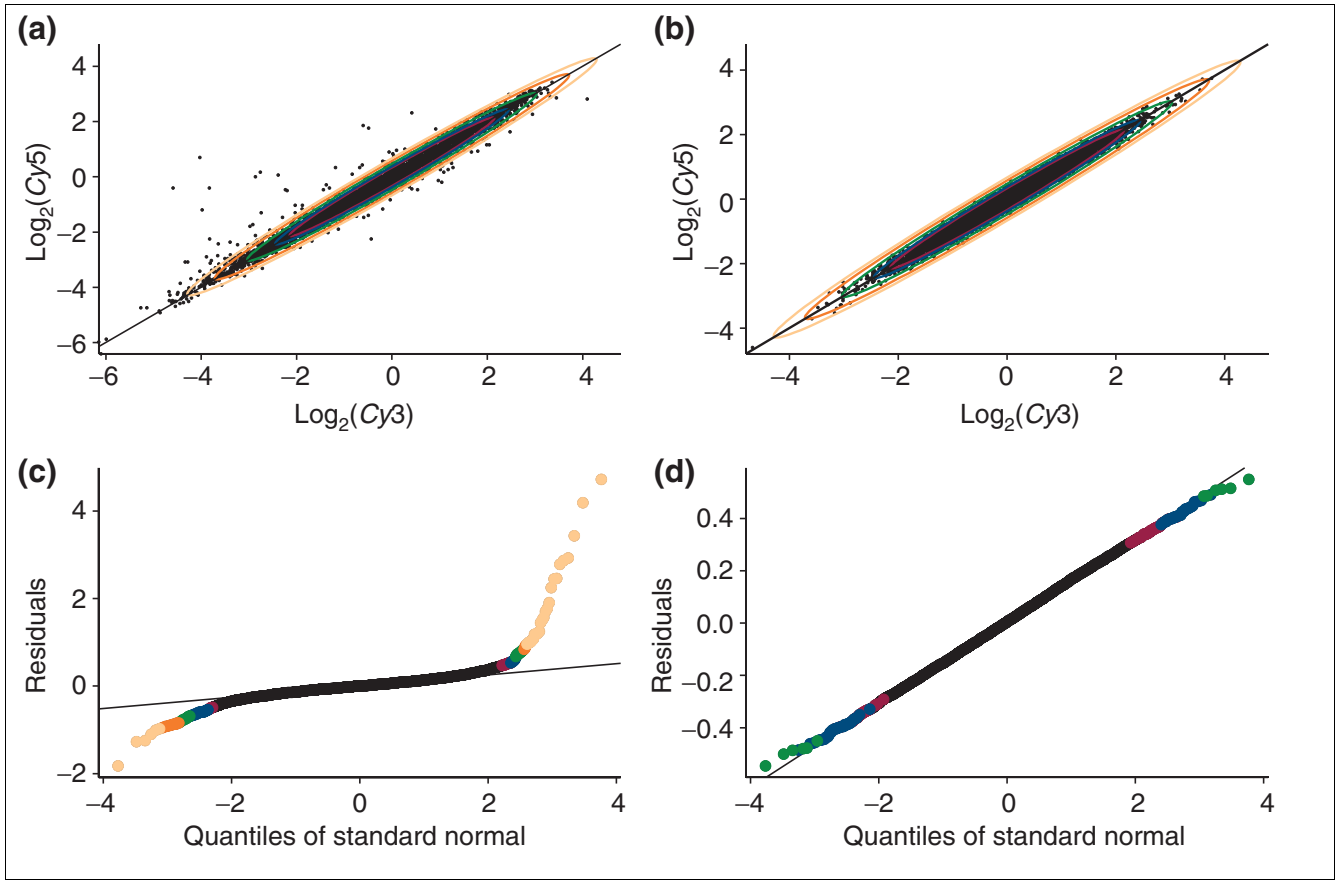

Figure 1
Visual tests of the underlying assumptions. Five concentration ellipses for **(a)** the standardized *mac1* dataset and **(b)** for a Monte Carlo simulated dataset with the same parameters of location and variance-covariance matrix (we used robust versions the location and scale estimators) as in *mac1* data. The tolerance ellipses cover 90% (red), 95% (blue), 99% (green), 99.9% (orange) and 99.99% (light orange) portions of the standard normal distribution to assist in visually testing the assumption of contaminated bivariate normality. QQNP of residuals for **(c)** *mac1* dataset and **(d)** for the corresponding Monte Carlo simulated dataset for comparison with (a) and (b). Outlying points are given in different colors in accordance with STIs in Figure 19b.

other scatter plots based on datasets from Hughes *et al.* [18]). Such a local non-linearity could be removed by applying a *lowess*-based normalization with appropriate smoothing parameters. The source of the systematic deviation is not known.

Detection of residual heteroscedasticity

In microarray data, the variance of residuals ($\log_2(Cy5/Cy3)$) is not a constant (homoscedasticity) but rather varies (heteroscedasticity) with intensity level ($\log_2(Cy5Cy3)/2$ or $\log_2(Cy3)$ or $\log_2(Cy5)$). The presence of residual heteroscedasticity argues strongly against arbitrary threshold methods to identify candidates for differential expression [10,11,19,21-24]. Approaches for assessing the heterogeneity of residual variance [10,11,25,26] include graphical, parametric and non-parametric methods. Here, we use non-parametric regression smoothing in 'absolute residuals versus $\log_2(Cy3)$' scatter plots to quantify residual variance. We compared S-plus/R *supsmu* (super smoother) and *lowess* (robust locally weighted regression) with other methods (scatter plots with tolerance ellipses, QQNPs with simulation envelopes, boxplots for residuals) and found that our regression smoother approach for absolute residuals performs well.

We used these smoothing methods to assess heteroscedasticity in the following way: grouping data into subsets of equal size and then applying regression smoothers to the median absolute residual in each group against median of $\log_2(Cy3)$ for that group [25]; same as this method but using boxplot for each group. One benefit of the latter approach is the fact that it can be used directly not only for residual diagnostics but also to take into account heteroscedasticity and estimate the number of candidates for differential gene expression. We also used Spearman rank correlation coefficients of the absolute residuals versus $\log_2(Cy3)$ to check the smoothing-based methods for consistency: positive values indicate increasing residual variance, negative ones indicate decreasing variance [25].

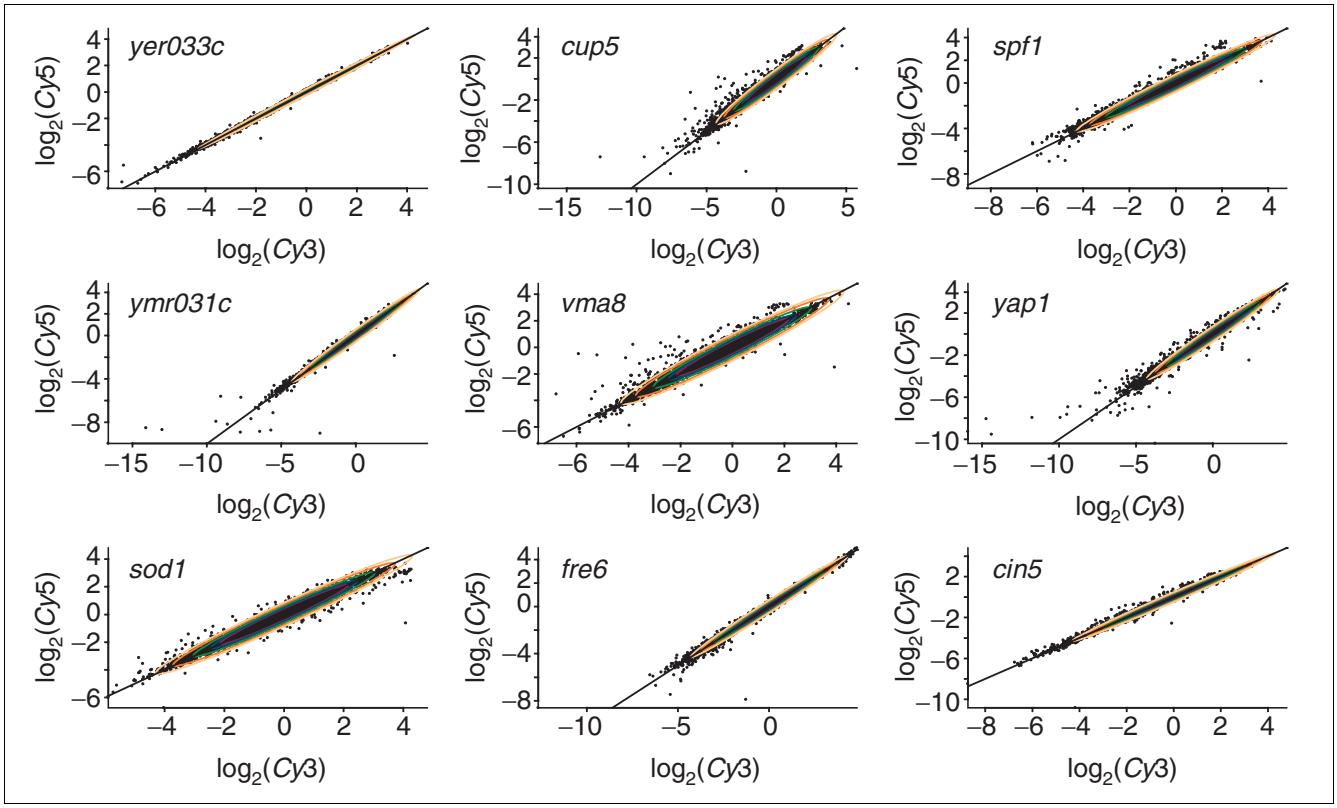

Figure 2
Overlay of concentration ellipses for the bivariate standard normal on real data. Scatter plots of nine datasets from Hughes *et al.* [18] overlaid with concentration ellipses for the standard normal distribution (see Figure 1 for the portions captured). Channel intensity values were log(base 2)-transformed, normalized and standardized.

We show an example of the results in Table 1 and Figures 11,12,13,14,15,16 for the *cup5* dataset. Figure 11 demonstrates the dependence between smoothed absolute residuals and smoothing parameters for *supsmu* and *lowess* procedures. As expected, *supsmu* procedure is more sensitive to prominent outliers in low intensity regions because it uses an automatically adjusted variable span. Prominent outliers in the low intensity area are both *Y*- and *X*-outliers and should be discarded. For the majority of *cup5* data, *supsmu* and *lowess* generate similar results. Figure 12 shows *supsmu* and *lowess* smoothing for 20 median-based sequential intervals of equal size and using different values of smoothing parameters at 'higher' resolution. Figure 13 does the same using *cup5* data in background. Figures 14 and 15 show boxplots for residuals using 10 and 20 equal size sequential intervals, respectively. They confirm the presence of heteroscedasticity as well. Boxplots for residuals with 20 subgroups of equal size using ±3*IQR*-based upper and lower extremes give an estimate for $k = 75$. This estimate is close to $k = 61$ identified by adjusted *supsmu*-based 99.998% simultaneous tolerance intervals (STIs) (Table 2). The difference in the estimates (14, or about 19%) can be explained by the fact that the ±3*IQR* rule generates about a 99.995% two-sided tolerance interval for a normally distributed population, while for a sample of finite size the corresponding upper and lower tolerance limits are wider (compare Equations 8 and 9) to cover 99.998% per residual group. Figure 16 is a smoothed version of Figure 14 using *supsmu* and *lowess* procedures for 3*IQR*-based extreme limits.

Statistical significance of outliers: ordinary and smoothed STIs

Our method equates contaminants of bivariate normal distributions (outliers) with candidates for differential expression. The outlier identification method has been developed previously for other applications [16,27,28]. We employ an approach based on the perspective of α-outliers and outlier regions [15,16]. In this approach, a point above the line of equivalence (that is, *Cy*3 = *Cy*5) is viewed as a candidate for an up-regulated gene, one below the line as a down-regulated gene and one in the vicinity of the line as an unchanged gene. Intuitively, points further away from the line - stronger outliers - are most likely to represent differentially-expressed genes. In other words, the probability that the observed difference in transcript level between the two samples might have arisen by chance decreases. To quantify these qualitative ideas, we applied statistical criteria to decide when points might result from no actual difference in expression (for

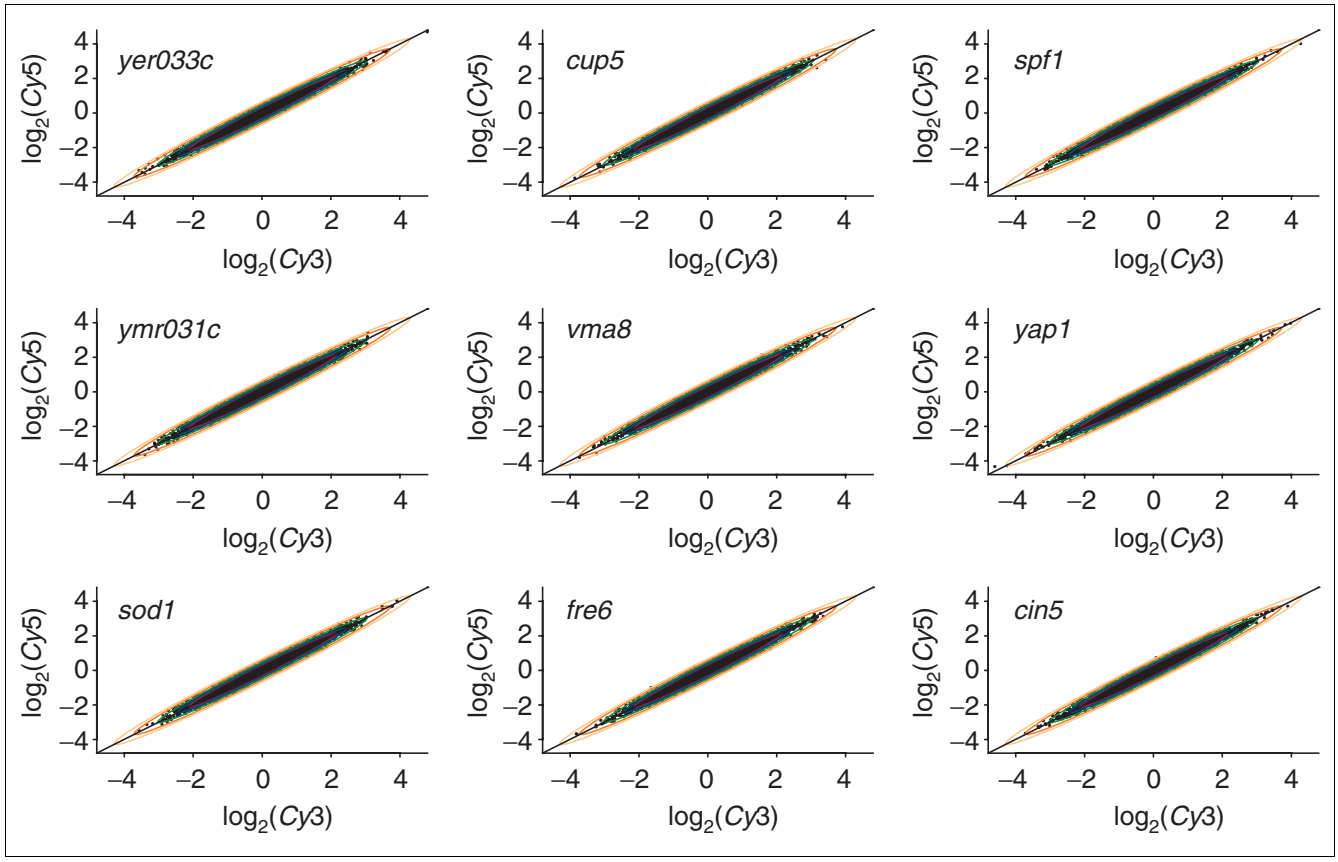

Figure 3
Overlay of concentration ellipses for the bivariate standard normal on simulated data. Scatter plots of nine simulated datasets (generated as random samples from a bivariate normal population) with overlaid concentration ellipses for the standard bivariate normal distribution (see Figure 1 for the portions captured). Channel intensity values were log(base 2)-transformed, normalized and standardized.

example, due to random fluctuations) versus those corresponding to genuine differential expression. We used a general α-outlier model for residuals (log-transformed normalized ratios) to identify candidates for differential expression.

In order to estimate the statistical significance of outliers, we used simultaneous tolerance intervals (STIs) based on Scheffé simultaneous confidence principles [29,30]. This approach guarantees a desired confidence level across the whole range of the predictor variable $X = \log_2(Cy3)$, $\log_2(Cy5)$ or $\log_2(Cy3Cy5)/2$ and for all $P = 100\%q$(the portion of the normal distribution covered by a certain STI, see Methods). We modified the approach using robust regression smoothers (*supsmu* or *lowess*) to approximate an unknown relationship, $s^2 = F(X)$, between residual variance and intensity. Five STIs for the *mac1* empirical and simulated datasets are shown (Figure 17a,b). For ordinary STIs, random fluctuations are seen to contribute to data points located away from the line of equivalence. However, the empirical data contain more and stronger outliers than the simulated data (Figure 17b). The five ordinary STIs were constructed under the assumption that the residual variance is constant (homoscedasticity) across the entire range of values for the predictor variable. We notice that for a large sample size (N = 6068, see Methods) ordinary STIs appear as straight lines (for small and moderate datasets they appear as hyperbolas; see Equations 2, 8, and 9 in Methods). Figure 17c shows residuals ($\log_2(Cy5/Cy3)$) as a function of $X = \log_2(Cy3)$ for the empirical *mac1* dataset. This plot and residual plots for other nine experiments (Figure 18) reveal that residual variance is not a constant (heteroscedasticity). Residual variance is commonly high for small values of X_i; it decreases to a minimum and may increase for large values of X_i, that is, the empirical dependence appears hyperbolic. We account for the heteroscedasticity by the use of smoothed STIs (Figure 17d). Accordingly, smoothed STIs appear as curves that are wider at low and high X_i values. Therefore, for a given portion of the normal distribution covered by a certain STI, points with X_i values at either extreme are further away from the line of equivalence.

For *mac1*, the width of the smoothed STIs is somewhat greater at low intensities compared to those at high

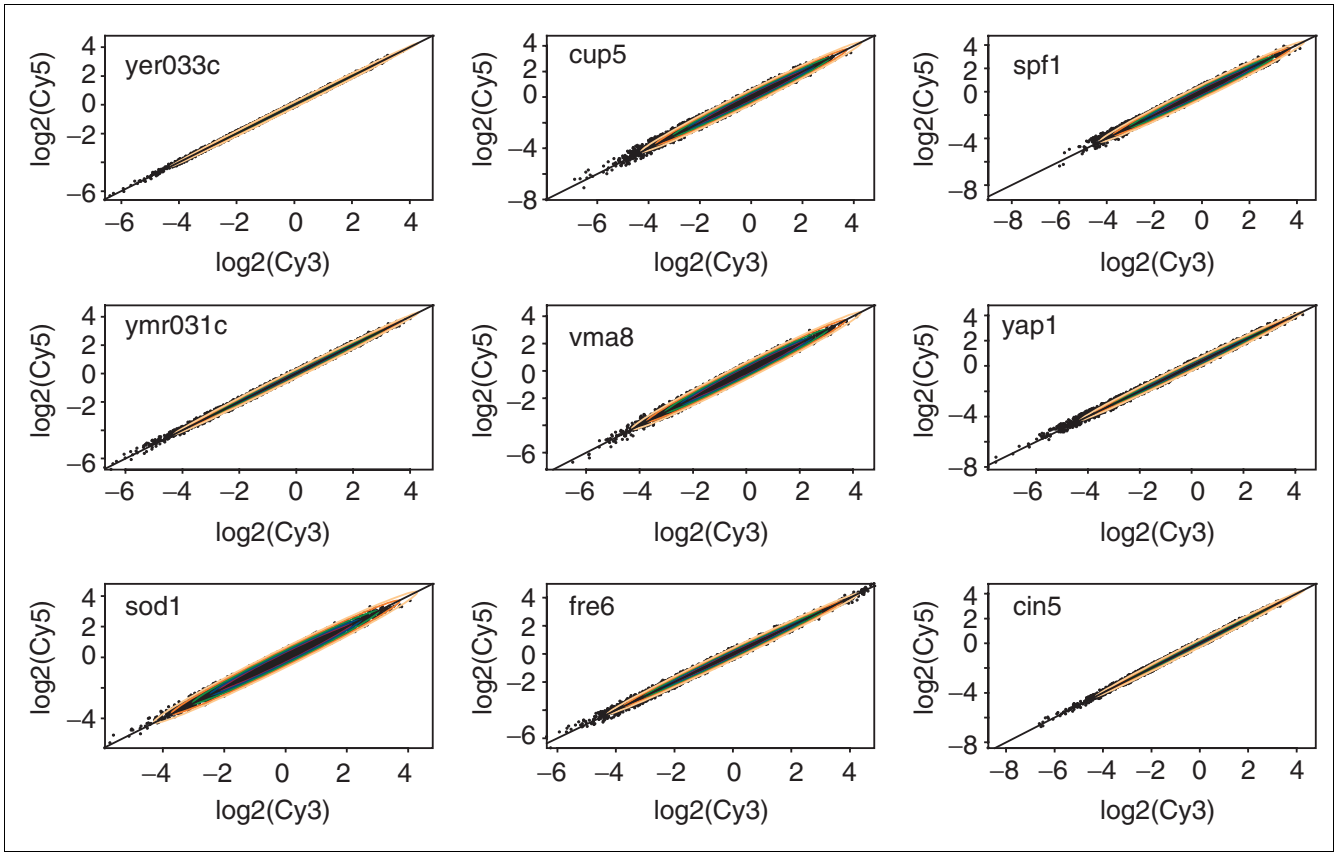

Figure 4
Overlay of concentration ellipses for the bivariate standard normal on real data with prominent outliers removed. Scatter plots of nine datasets from Hughes *et al.* [18] after outlier removal with concentration ellipses for the standard bivariate normal distribution (see Figure 1 for the portions captured). Two-sided 99.9% cut-off and robust measure of scale (median absolute deviation) for residuals were used to identify outliers. Channel intensity values were log(base 2)-transformed, normalized and standardized.

intensities. In the mid-range of X_i values, smoothed STIs lead to intervals that are narrower than ordinary STIs that do not consider heteroscedasticity. Therefore, candidates for differentially-expressed genes are more likely to be identified in the middle range of X_i values and are less likely to be defined at the extremes.

We evaluated the *lowess*- and *supsmu*-smoothing procedures by applying them to a simulated dataset taken from an 'ideal' bivariate normal population with the same parameters as the empirical *mac1* dataset. The robust scale estimates using the Huber τ-estimator for scale, *supsmu*- and *lowess*-based scale estimators are shown in Figure 19a. The smoothed scale estimators generate approximately straight lines parallel to the Huber τ-scale estimator.

Adjusted STI

An adjustment for Gaussian efficiency is necessary for the application of robust estimators [31] such as those we use for outlier identification in the presence of heteroscedasticity. We therefore adjust smoothed STIs to improve their accuracy. We calculate an adjustment constant (scale factor) to compensate for the difference between the Huber τ-estimator for scale and *supsmu*- or *lowess*-based scale estimators (see Methods for details). The adjusted constant is used as a scale factor for the smoothed STIs for empirical data. Adjusted smoothed STIs are shown in Figure 18 and Figure 19b. The dramatically different STIs amongst the ten datasets reflect their individual patterns of residual variance and demonstrate the necessity of tailored analysis of a dataset.

Candidates for differential expression

We identify candidates for differential expression by using STIs containing the $P = 100(1 - \alpha)$ portion of normal distribution covered with probability at least $1 - \gamma$ (see Methods). For *mac1*, no simulated data lies outside the 99.998% ordinary STIs (γ-level = 0.0001) suggesting that empirical data points outside the corresponding adjusted smoothed *supsmu*-based STIs are good candidates for differentially-expressed genes. For the *mac1* analysis (Table 3), 41 candidate genes for up-regulation and 20 candidates for down-regulation are identified using a 99.998% (γ-level = 0.0001) adjusted *supsmu*-based STI. For the ten datasets examined, up to approximately 2% of the genes were candidates for differential

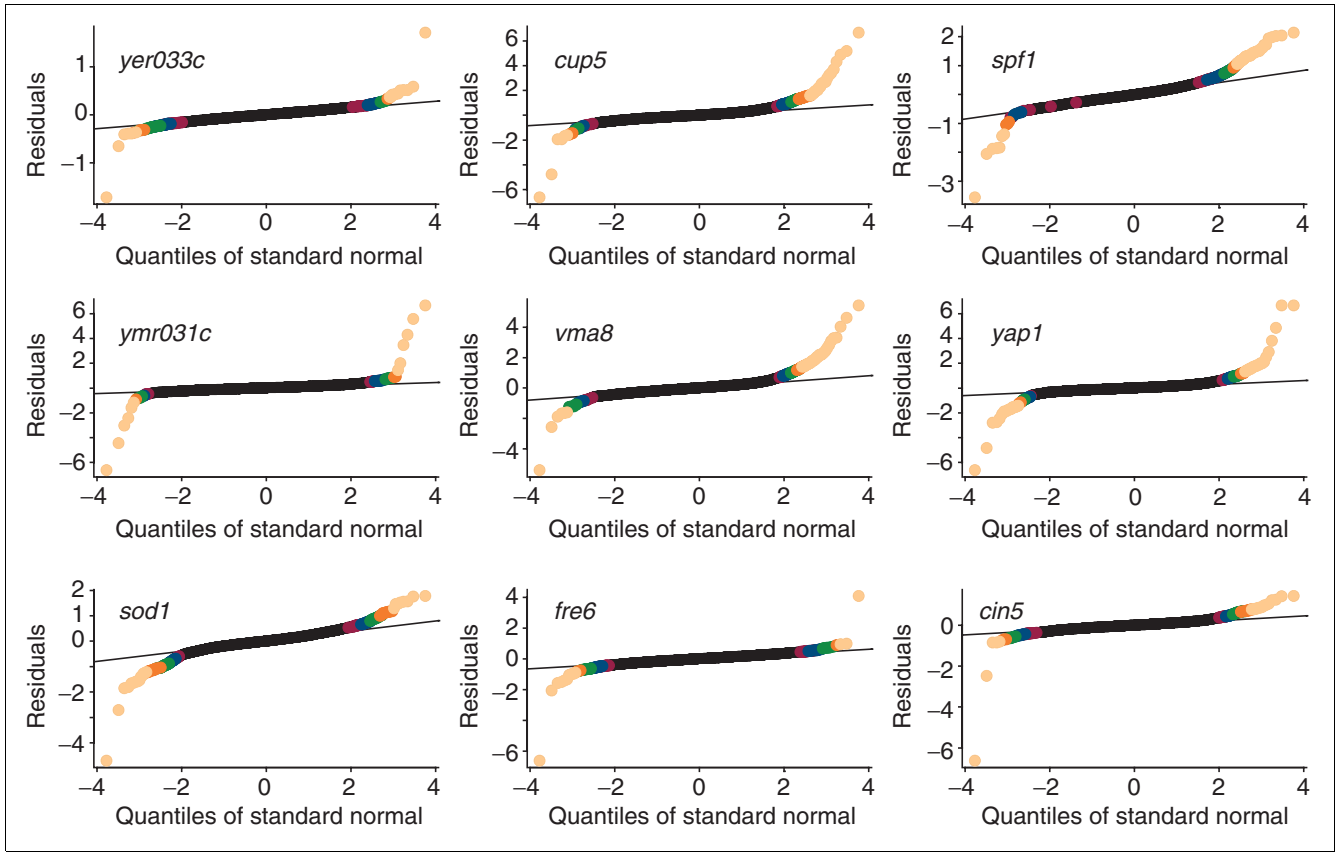

Figure 5
QQNP for real data: residuals of nine datasets from Hughes *et al.* [18]. Channel intensity values were log(base 2)-transformed and normalized. Compare with Figure 18 (colors for outliers match the tolerance band colors).

expression (see Table 2, Table 3 and Table 4 for *mac1* data and a comparative summary table for all ten datasets in Additional data). Overall, adjusted smoothed STIs provide a better balance between sensitivity and specificity across the whole range of predictor variable values ($\log_2(Cy3)$) and are thus more reliable than ordinary STIs. The approach takes into consideration multiplicity of comparisons, variation in the experimental response around the line of equivalence (or around zero for residuals) and intensity dependent variation in residual variance.

Differential expression in a single cDNA microarray: adjusted smoothed STIs and existing methods

We compared the adjusted smoothed STI technique for identifying differentially-expressed genes with other methods for single cDNA microarray data [13,14,17-19]. Within the framework of outlier detection analysis, the primary difference amongst these methods is the means used to define statistical intervals (see Discussion for the details of each model). In the arbitrary ratio approach, $Y_i/X_i = \log_2(Cy5/Cy3) = r_i$ defines a gene i as being differentially expressed if $r_i > t$ where t is a user-defined threshold [13,14,17]. The most frequently used value t corresponds to residuals of -1 (two-fold down) and 1 (two-fold up). Figure 20 compares differential expression in three different experiments using a ratio threshold and adjusted *supsmu*-based STIs. For $t = \pm 1$, any gene outside this cut-off would be deemed as up- or down-regulated. However, employing the criterion of genes higher than the 99.998% (γ-level = 0.0001) adjusted *supsmu*-based STIs would yield additional candidates for differential expression. Although $t = \pm 1$ seems more conservative for these datasets, it may be overly liberal for others.

Hughes *et al.* [18] developed an error model that made use of additional information about the variability of each gene based on 63 'same versus same' control experiments. Figure 21a and Table 4 compare differential expression in the *mac1* data as defined using the 'gene-specific' error model [18] and adjusted *supsmu*-based STIs. As we discuss below, some genes which were identified as differentially expressed using our adjusted *supsmu*-based STIs were not identified by the error model [18]. Our approach with four other models [17,19,20] (see also Discussion) in an outlier detection framework is compared in Figure 21a,b.

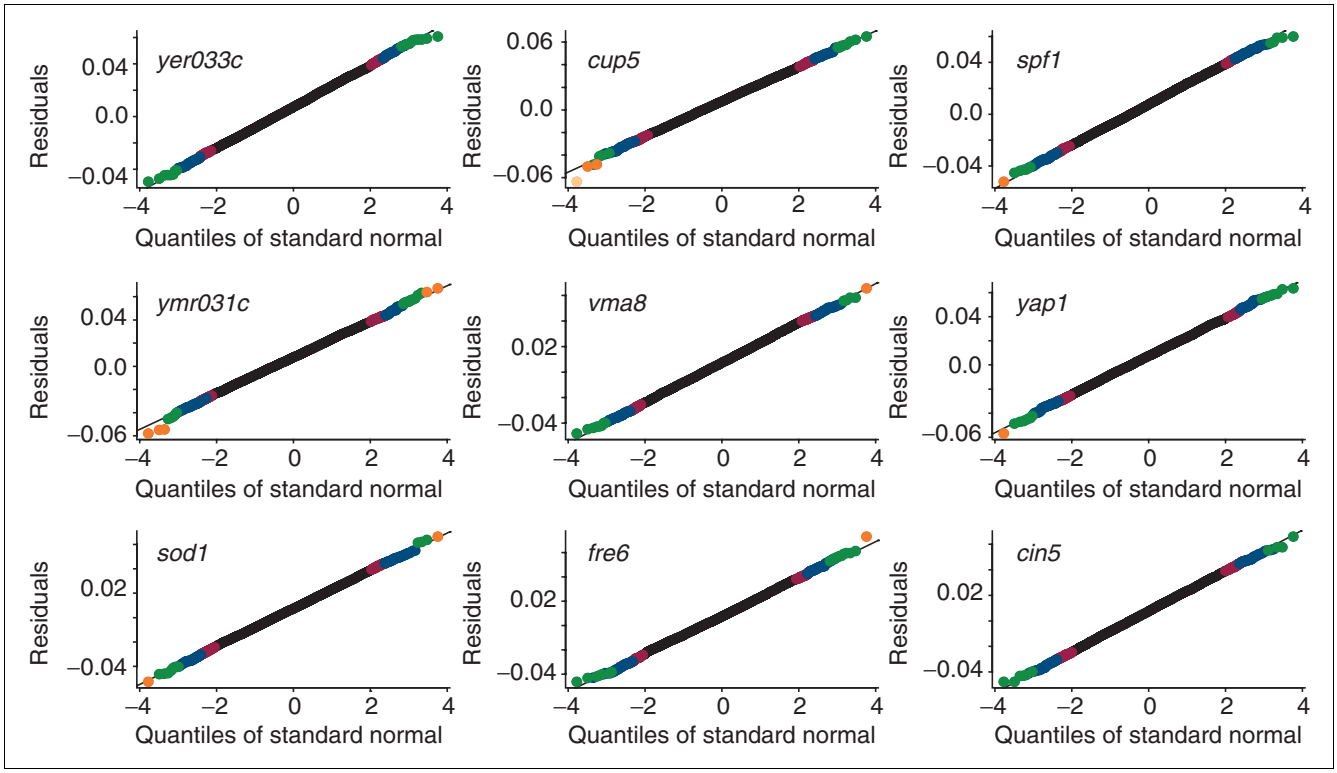

Figure 6
QQNP for residuals of nine simulated datasets (generated as random samples from a normal population). Channel intensity values were log(base 2)-transformed and normalized. Compare with Figure 18 (colors for outliers match the tolerance band colors).

In general, our adjusted smoothed STI method generates narrower bands in the mid-range of gene expression levels and broader bands in low and higher intensity areas. For *mac1*, the bands for Newton *et al.*'s method [19] and our method appear similar qualitatively. The STI-based measure of statistical significance takes into consideration the unique features and properties of empirical microarray datasets.

Simulation studies

We carried out simulation studies using sample parameter estimates from the *mac1* dataset to assess the performance of each of the single-slide methods. We created artificial datasets with 100 candidates (outliers) for differential expression. We simulated k = 100 non-regular observations and N-k = 6,068-100 = 5,968 regular observations (the main body of non-differentially-expressed genes). A random component was added to each outlier value using standard normal distribution with variance dependent on intensity. This set of 100 represents the 'true' outliers due to 'differential expression'. We simulated heteroscedasticity present in many datasets by including intensity-dependent variability in the low and high intensity levels for both non-regular and regular data points. We then compared the performance of each method to identify candidates for differential expression in multiple repeat runs of the simulation. (R code and data used for the simulations can be obtained from the authors upon request.) Figure 22 shows a plot of the simulated data with true outliers shown in red. We compared the performance of several different single-slide methods at ten 'cut-off' levels of relatively equivalent stringency as shown in Table 5 (except for Chen *et al.* [17] which use only two levels of significance). We compared PPV (positive predictive value), NPV (negative predictive value), sensitivity, specificity and likelihood ratios at each cut-off for each method (please see definitions of the test accuracy measures in [32] and in Additional data). We plotted these results in Figure 23 as a receiver operating characteristic (ROC) curve, a PPV curve, and a likelihood ratio curve. These results clearly demonstrate that our method outperforms existing single-slide methods with improved positive predictive values, likelihood ratio and higher ROC curves (greater area under the curve). These improved performance differences are most apparent at the most stringent significance levels which are likely to be most relevant in the context of multiple comparisons.

Comparison of biological significance of *mac1* results

The effects of the *mac1Δ* on the metabolism and gene expression in yeast are well documented. The absence of the Mac1p, a copper responsive transcription factor, results in down-regulation of copper uptake transporters and subsequent copper deficiency [33-36]. Copper is required for Fet3p which in turn is necessary for iron uptake in yeast. As a consequence,

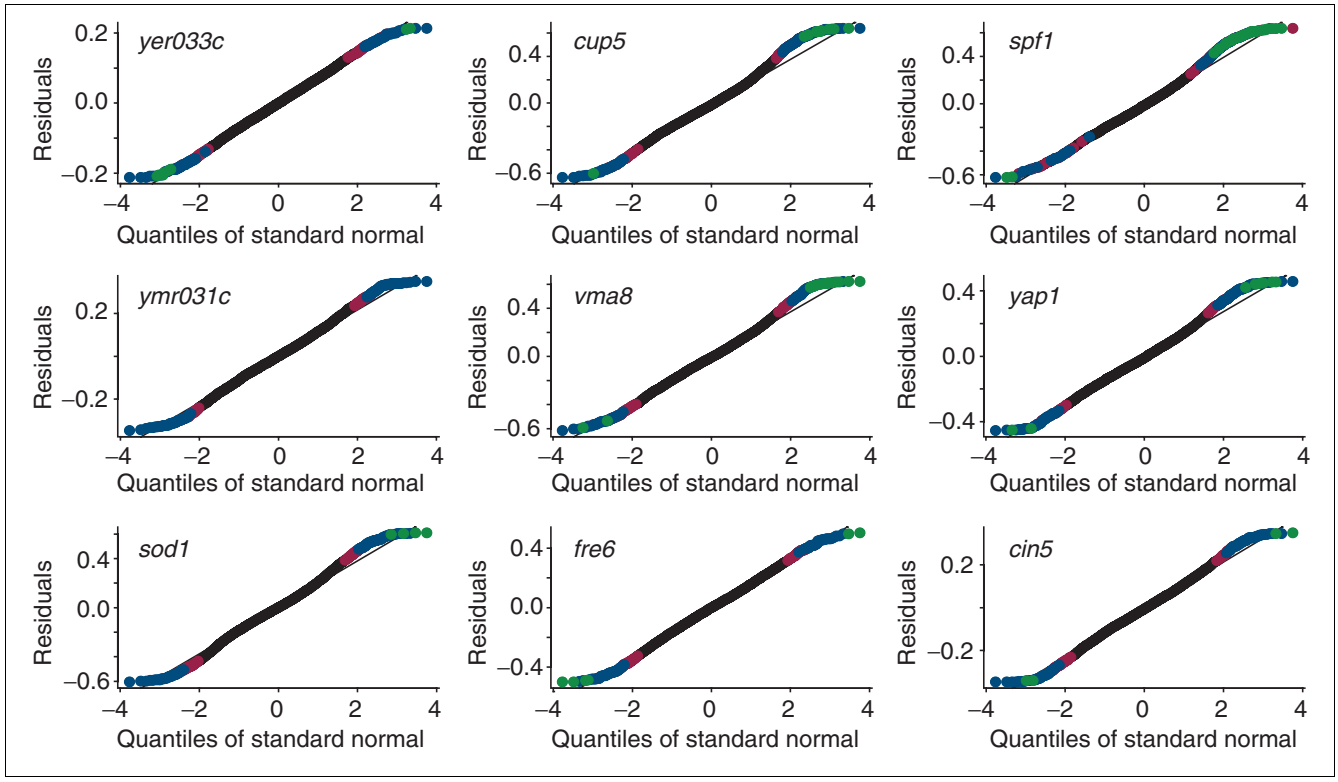

Figure 7
QQNP for residuals of nine datasets from Hughes *et al.* [18] after prominent outlier removal. Two-sided 99.9% region and robust measure of scale (median absolute deviation) for residuals were used to remove outliers. Channel intensity values were log(base 2)-transformed and normalized. Compare with Figure 18 (colors for outliers match the tolerance band colors).

copper deficiency results in secondary iron deficiency [37,38]. Iron deficiency leads to activation of iron responsive transcription factors, Aft1p and Aft2p, which induce transcription of a host of genes encoding proteins involved in iron uptake [39]. The identification of up-regulation of these target genes provides a reasonable biological standard for comparing the performance of the different methods. In addition, down-regulation of Mac1p targets might be expected in a *mac1Δ*. In Table 4, we present a comparison of the methods at relatively equivalently high levels of stringency for likely Aft1/2p and Mac1p targets present on the arrays and identified by at least one method as differentially expressed. We also include the two-fold cut-off for comparison. A total of 13 genes were not identified as up-regulated by Hughes *et al.* Two genes previously identified as a Mac1p target (*YFR055W*) [40] or down-regulated in *mac1Δ* (*CTT1*) [33] and *MAC1* itself were not identified by the Hughes *et al.* method. While identifying many of the Aft1/2p targets excluded by Hughes *et al.*, the other two methods did not identify *MRS4* or *SMF3*, which are regulated in response to iron deficiency, nor did they identify *YFR055W* and *CTT1* as down-regulated. We suggest that these results provide some biological validation of our approach and indicate increased performance of our method over the other methods at stringent significance levels - necessary given the multiplicity of comparisons.

Discussion

Multiple replications in the design of reagents (multiple spotting of each gene on a microarray) and experimental approach (multiple replicates of each hybridization) provide the soundest approach to confirm differential expression of genes (see, for example, [2-12]). However, experimental realities such as limited samples (for example, tumor specimen), a large number of samples (for example, time course experiments) and experimental cost have resulted in the vast majority of published cDNA microarray studies using limited or no replication. Several methods currently exist for the analysis of data from experiments with limited or no replication. Unfortunately, real microarray data generally violate the assumptions underlying these methods.

Limitations of underlying assumptions of current single-slide methods

Chen *et al.* [17] assumed that raw non-normalized and non log-transformed *Cy5* and *Cy3* intensities (raw intensities) are drawn from independent normal populations with common

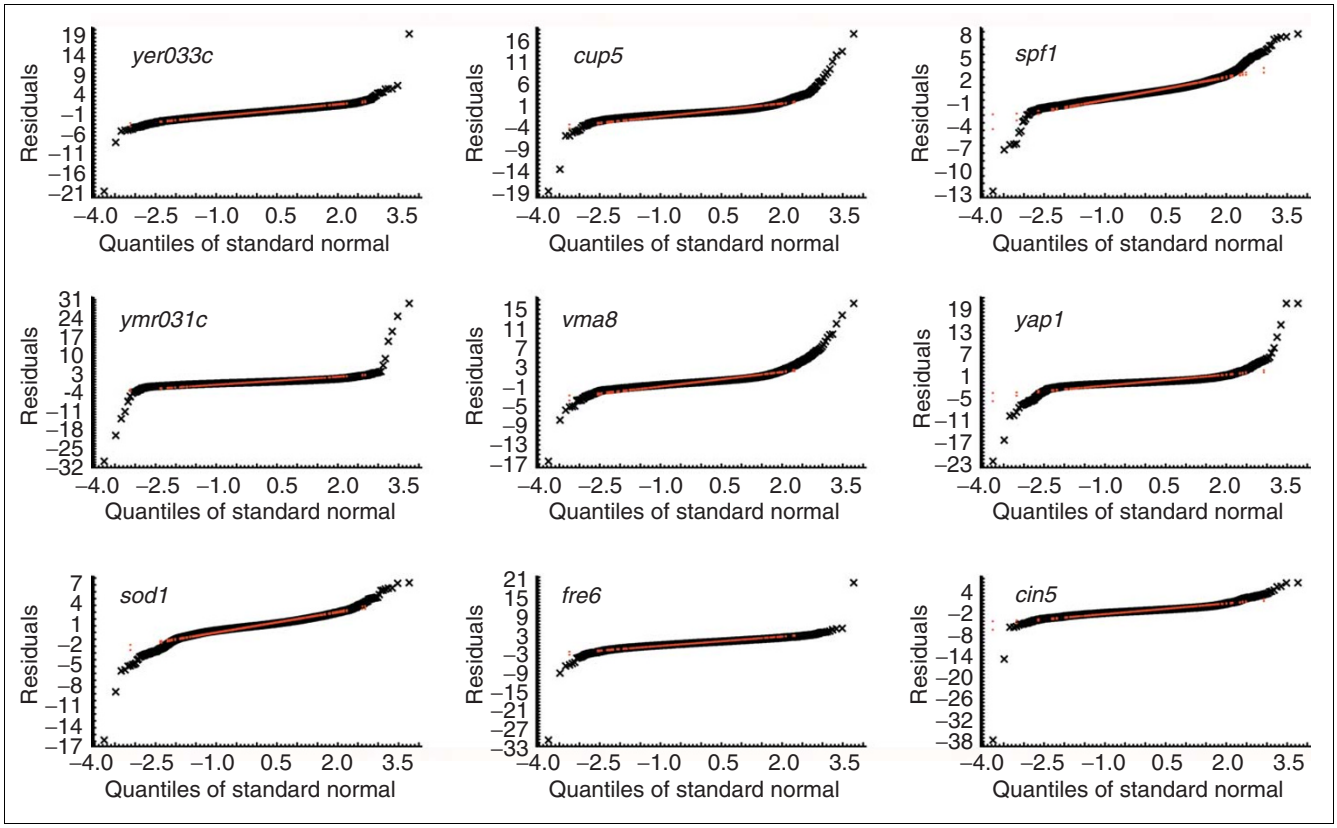

Figure 8
QQNP with simulation envelopes (based on 1,000 random samples from a normal population) for residuals of nine datasets from Hughes *et al.* [18]. Channel intensity values were log(base 2)-transformed and normalized. The envelopes are depicted as dashed red lines.

coefficient of variation. An asymmetric density function for raw ratios was derived. This results in asymmetric bands with the identification of more up-regulated than down-regulated genes irrespective of the dataset (compare Figure 4 in [19] and Figure 7 in [8]).

Newton *et al.* [19] assumed that because raw *Cy*5 and *Cy*3 intensities are always positive they can be considered as observations from a Gamma distribution with the same coefficient of variation (if *Cy*3 and *Cy*5 are independent, their joint distribution is a bivariate Beta distribution). Hierarchical Gamma-Gamma and Gamma-Gamma-Bernoulli models were formulated in which the posterior odds of change in expression were an additive (*Cy*5 + *Cy*3) and multiplicative (*Cy*5*Cy*3) functions of intensity. Contours of the posterior odds in ($X \equiv \log(Cy3)$, $Y \equiv \log(Cy5)$) scatter plots were used to identify differentially-expressed genes. In practical situations, it may be difficult to determine if data are log-normal or Gamma [41] but we argue that the former is more realistic for microarray data because the combination of biological and experimental noise results in the majority of the measured expression levels changing randomly, independently, non-directionally and for those changes to be small. The central limit theorem would therefore predict bivariate normality for the majority of log-transformed spot intensity values.

Sapir and Churchill [20] compute the posterior probability of differential expression using a mixture of orthogonal residuals derived from ordinary least squares regression of ($X \equiv \log_2(Cy3)$, $Y \equiv \log_2(Cy5)$). The approach assumes that differentially-expressed genes are drawn from populations with unknown distributions approximated with uniform distributions. This mixture model approach assumes that all outliers (k non-regular observations) follow the same distribution $D_0 = D_1 = \ldots = D_k$. Under a mixture model, contaminants are less separated from the regular observations than when using Ferguson-type model [16]. The method used to obtain orthogonal residuals in this approach is not resistant to outliers [42] and is redundant because the use of log-transformed ratios $\log_2(Cy5/Cy3)$ assumes normalization by linear regression to enforce slope equal to 1 and intercept equal to 0 (compare [8]). This approach does not take into consideration residual heteroscedasticity. Similarly, Yue *et al.* [1] described a

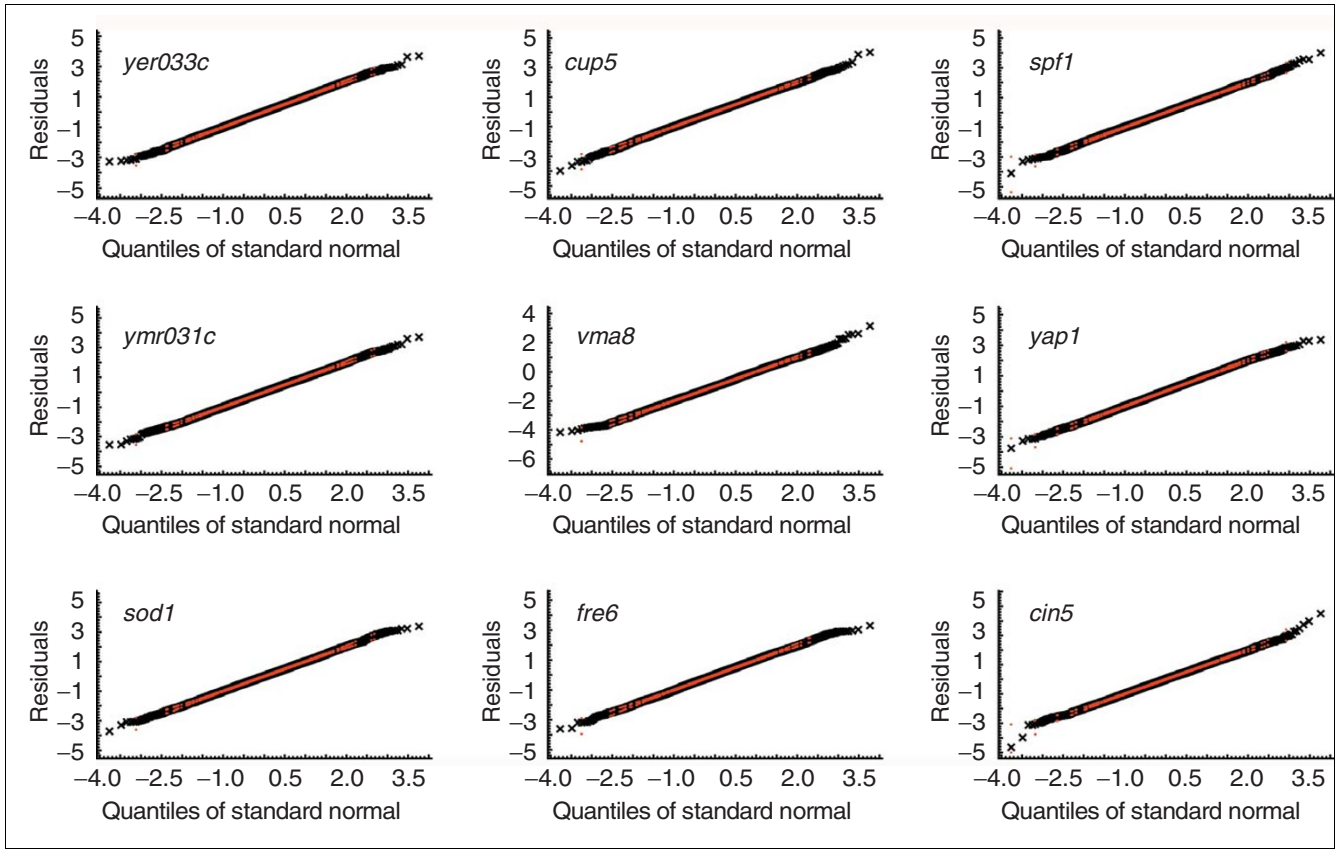

Figure 9
QQNP with simulation envelopes for Monte Carlo simulated data. QQNP with simulation envelopes (based on 1,000 random samples from a normal population) for residuals of nine simulated datasets (generated as random samples from a normal population). Channel intensity values were log(base 2)-transformed and normalized. The envelopes are depicted as dashed red lines.

method based on parametric two-sided tolerance intervals for ratios. This approach does not consider residual heteroscedasticity and the multiplicity of comparisons.

Hughes *et al.* [18] performed 63 control 'same versus same' hybridizations in addition to 300 'treatment versus control' cDNA microarray experiments, many in duplicate. They filtered candidates for differential expression on the basis of the information about individual variability in expression levels, and genes with unusually high variation were discarded (Figure 21a). One possible drawback is the assumption that the expression variance is the same in both samples. Hughes *et al.* [18] quantified residual heteroscedasticity using weighted location and scale estimators for 'same versus same' replicates. Their method used non-robust versions for the estimators - consequently the location estimates may have a bias and the scale estimates would be significantly inflated in the presence of outliers [43]. In addition, since this method uses extensive 'same versus same' hybridizations, it cannot be considered to be a single-slide method.

Advantages of α-outlier model and outlier identification method

In this work, we have described a *post hoc* (data-oriented) method, which makes fewer assumptions about the nature of the data, tests the assumptions to ensure their validity and produces computationally reasonable results. We showed that a reasonable model is one where the processed fluorescent intensity values are samples drawn from a bivariate normal population contaminated with outliers and possibly distorted due to heteroscedasticity. After a normalization by a robust linear regression fit to make slope equal to 1 and intercept equal to 0, in general, most data points in the $\log_2(Cy5)$ versus $\log_2(Cy3)$ scatter plot lie close to the line of equivalence ($\log_2(Cy5) = \log_2(Cy3)$) while a limited number of data points, outliers, lie outside the vicinity. The outliers are good candidates for differentially-expressed genes. The further an outlier is located from the line of equivalence the more likely it is to represent a systematic outlier rather than a chance observation. The α-outlier-generating model approach for identifying differentially expressed gene

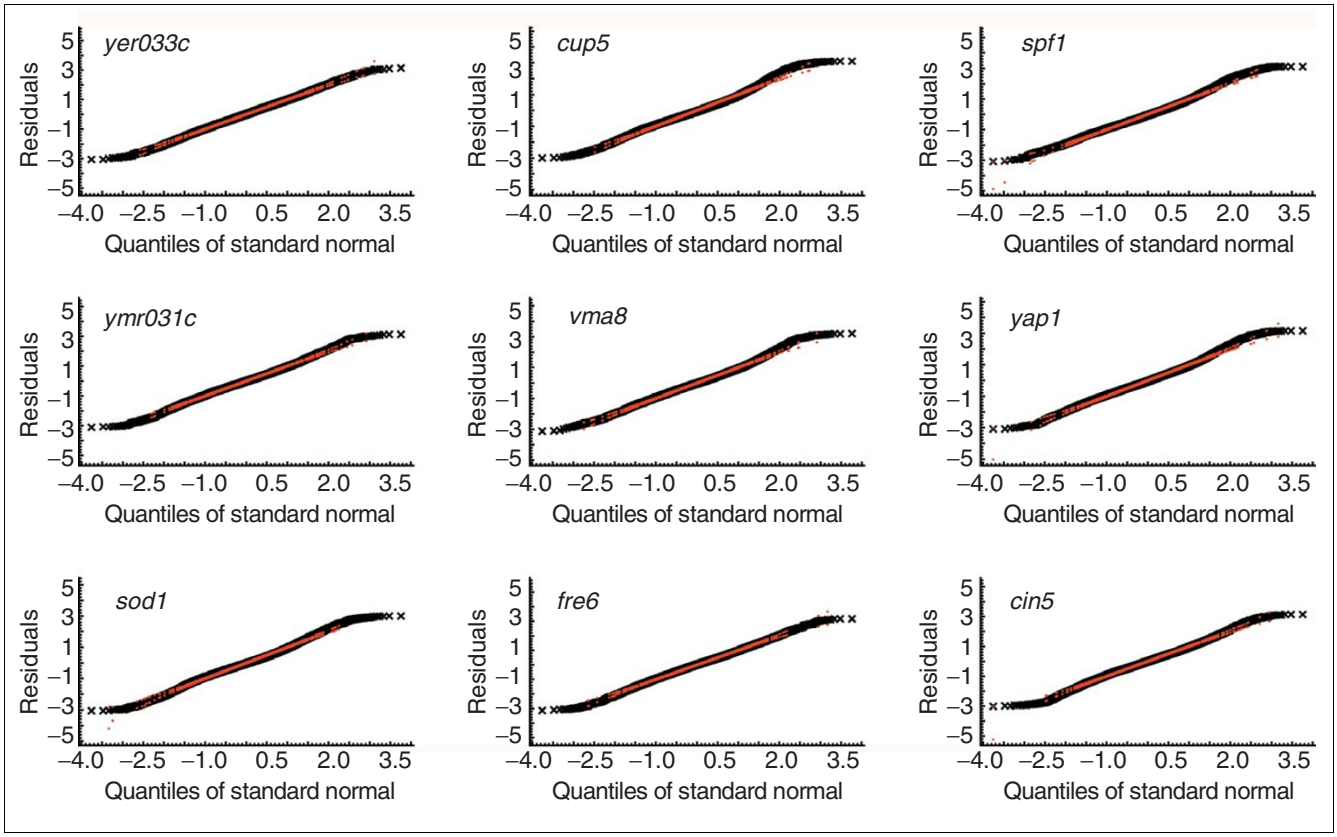

Figure 10
QQNP with simulation envelopes (based on 1,000 random samples from a normal population) for residuals of nine datasets from Hughes *et al.* [18] after prominent outlier removal. Two-sided 99.9% cut-off and robust measure of scale (median absolute deviation) for residuals were used to remove outliers. Channel intensity values were log(base 2)-transformed and normalized. The envelopes are depicted as dashed red lines.

Table 1

Detecting heteroscedasticity

Dataset	X_{min}	Rho_1	Rho_1 *p*-value	Rho_2	Rho_2 *p*-value
mac1	-1.50	-0.08	1.36E-03	0.16	0.00*E*+00
yer033c	-0.70	-0.10	2.30E-08	0.11	3.70E-09
cup5	-0.93	-0.17	0.00*E*+00	0.11	5.44E-10
spf1	-1.44	-0.18	0.00*E*+00	0.20	0.00*E*+00
ymr031c	-0.89	-0.17	0.00*E*+00	0.11	3.43E-10
vm8	-0.82	-0.10	7.98E-09	0.10	1.66E-08
yap1	-0.97	-0.23	0.00*E*+00	0.12	1.59E-10
sod1	-1.54	-0.09	3.68E-04	0.15	0.00*E*+00
fre6	-1.77	-0.13	2.69E-10	0.06	2.46E-04
cin5	-0.28	-0.12	2.17E-14	0.11	1.09E-07

Use of Spearman rank correlation for absolute residuals to detect heteroscedasticity [25] in ten datasets from Hughes *et al.* [18]. Empirical hyperbolas (here they are based on *supsmu* smoother) have minima around sample means. As a result, we use two subintervals to compute Spearman rank correlation coefficient: from minus infinity to X_{min} ($\log_2(Cy3)$ axis) and from X_{min} to plus infinity. We note that sign of Spearman rank correlation always coincides with the sign of first derivative for empirical hyperbolas at a given subinterval (compare Figure 20). Rho_1, Spearman coefficient of rank correlation for the former subinterval; Rho_1 *p*-value, *p*-values for values in column Rho_1; Rho_2, Spearman coefficient of rank correlation for the latter subinterval; Rho_2 *p*-value, *p*-values for values in column Rho_2 (*p*-values are given in scientific notation, *0.00E+00* means that the respective *p*-value was less than 10^{-16}).

Table 2

Candidate differential expressed genes with different approaches

RN	ORF name	STI1	STI2	smuSTI1	smuSTI2	adsmuSTI1	adsmuSTI2
1	YBR047W			+	+	+	
2	YBR207W	+	+	+	+	+	+
3	YBR295W			+	+	+	+
4	YDR264C	+		+	+	+	+
5	YDR269C			+	+		
6	YDR270W	+	+	+	+	+	+
7	YDR441C			+	+		
8	YDR476C	+		+	+	+	+
9	YDR534C	+	+	+	+	+	+
10	YEL065W	+	+	+	+	+	+
11	YER145C	+	+	+	+	+	+
12	YFL041W	+		+	+	+	+
13	YFR023W			+			
14	YFR024C-A			+			
15	YGL015C	+	+	+	+	+	+
16	YGL039W	+	+	+	+	+	+
17	YGL055W			+			
18	YGR065C			+	+	+	+
19	YGR079W			+	+	+	
20	YGR257C			+	+		
21	YHL035C	+	+	+	+	+	+
22	YHL040C	+	+	+	+	+	+
23	YHL047C	+	+	+	+	+	+
24	YHR042W			+	+	+	+
25	YHR175W	+	+	+	+	+	+
26	YJL145W			+			
27	YJL153C			+	+	+	+
28	YKL039W			+			
29	YKL220C	+	+	+	+	+	+
30	YKR052C			+	+	+	
31	YLL051C	+	+	+	+	+	+
32	YLL053C			+	+	+	
33	YLR034C			+	+	+	+
34	YLR046C			+	+	+	+
35	YLR056W			+	+	+	+
36	YLR126C	+	+	+	+	+	+
37	YLR127C			+	+		
38	YLR136C	+		+	+	+	+
39	YLR192C			+			
40	YLR205C	+	+	+	+	+	+
41	YLR214W	+	+	+	+	+	+
42	YMR006C			+	+	+	+
43	YMR011W			+	+	+	+
44	YMR058W	+	+	+	+	+	+
45	YMR251W	+	+	+	+	+	+
46	YMR319C			+	+		
47	YNL237W	+	+	+	+	+	+
48	YNL259C	+	+	+	+	+	+

Table 2 *(Continued)*

Candidate differential expressed genes with different approaches							
49	YNR056C	+	+	+	+	+	+
50	YNR060W	+	+	+	+	+	+
51	YOL153C	+	+	+	+	+	
52	YOL158C	+	+	+	+	+	+
53	YOR334W			+	+	+	+
54	YOR381W	+	+	+	+	+	+
55	YOR382W	+	+	+	+	+	+
56	YOR383C	+	+	+	+	+	+
57	YOR384W	+	+	+	+	+	+
58	YPL171C			+			
59	YBR250W			+			
60	YDR423C			+			
61	YER028C			+			
62	YHR199C			+			
63	YIL169C			+			
64	YJL149W			+	+	+	
65	YOL101C			+		+	
66	YOL164W			+		+	
RN	ORF name	STI1	STI2	smuSTI1	smuSTI2	adsmuSTI1	adsmuSTI2
1	YBR054W	+		+	+	+	+
2	YBR145W			+			
3	YBR147W			+	+	+	+
4	YCL030C	+		+	+	+	+
5	YDL171C	+		+	+	+	+
6	YDR035W			+	+	+	
7	YDR234W			+	+		
8	YEL039C			+	+	+	+
9	YER001W			+			
10	YER156C			+	+		
11	YER174C	+		+	+	+	+
12	YFL014W	+		+	+	+	+
13	YFR030W	+	+	+	+	+	+
14	YFR055W			+	+	+	+
15	YGL009C	+	+	+	+	+	+
16	YGL117W	+	+	+	+	+	+
17	YGR088W			+	+	+	+
18	YGR286C	+	+	+	+	+	+
19	YHL021C	+	+	+	+	+	+
20	YHL028W			+	+	+	
21	YHR018C	+		+	+	+	
22	YHR029C			+			
23	YHR045W			+			
24	YIL111W			+	+		
25	YJL048C			+	+	+	
26	YJL088W			+	+		
27	YJL089W			+	+		
28	YJL200C			+	+	+	
29	YJR016C			+			
30	YJR025C			+			

Table 2 *(Continued)*

Candidate differential expressed genes with different approaches							
31	YJR109C			+	+	+	
32	YJR137C	+		+	+	+	+
33	YKL062W			+	+	+	
34	YKL109W			+	+	+	
35	YKL141W			+	+	+	
36	YKL148C			+	+	+	+
37	YKL218C			+			
38	YKR066C			+	+	+	
39	YLL041C	+		+	+	+	+
40	YLR220W			+	+	+	
41	YLR304C	+		+	+	+	+
42	YMR021C	+	+	+	+	+	+
43	YMR022W			+	+	+	
44	YMR095C			+			
45	YMR096W			+	+	+	
46	YMR271C			+	+		
47	YNL160W			+			
48	YOL058W			+	+	+	
49	YOL064C			+	+		
50	YOR065W			+	+	+	
51	YOR195W			+	+		
52	YOR230W			+			
53	YOR356W			+	+	+	+
54	YPL092W			+	+	+	
55	YPR123C			+	+	+	
56	YPR160W			+	+		
57	YCR106W			+			
58	YGR052W			+			
59	YJR130C			+			
60	YOL119C			+			
61	YPL123C			+			
62	YPR167C			+			
Total		**47**	**34**	**128**	**98**	**84**	**61**

Candidates for differentially-expressed genes in the *mac1* cDNA microarray experiment based on three different approaches: STI1 is ordinary STI at 99.98% and STI2, at 99.998%; smuSTI1 is *supsmu*-based STI at 99.98% and smuSTI2, at 99.998%; adsmuSTI1 is adjusted *supsmu*-based STI at 99.98% and adsmuSTI2, at 99.998%. The study monitored transcripts in a *mac1* knockout and wild type *S. cerevisiae*. For the STIs, the above-mentioned captured portions of the respective normal distributions were covered with probability at least 99.99%.

candidates makes no assumptions about outlier distribution and dependency structure of the candidates. The mixture model of Sapir and Churchill [20] assumes that all candidates have the same distribution D_0 ($D_0 = D_1 = ... = D_k$, see Methods). Other approaches (for example, [17,19]) assume independence of channel intensities. The individual distributions are usually not known and could be different for each gene, since only replication allows estimation of the distributions. Multiple levels of dependence, such as co-regulated genes, are expected rather than unlikely in gene expression analysis. Since our model and analysis approach does not require such assumptions, which are likely to be violated by the data, we would argue that our approach is more realistic and generally applicable.

Accommodating heteroscedasticity in outlier identification

We compensate for a source of reproducible systematic technical error, heteroscedasticity, by using robust non-parametric regression smoothers to quantify the differences in the variability of gene expression values as a function of spot intensity levels. STIs corrected for heteroscedasticity and

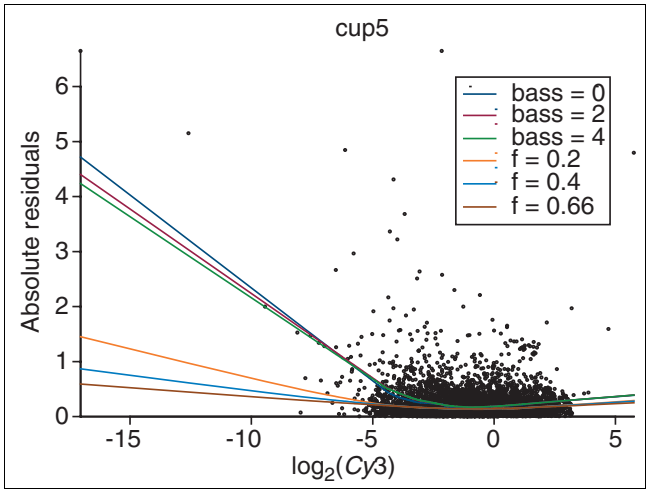

Figure 11
The use of smoothed absolute residuals to diagnose and quantify residual heteroscedasticity. 'Absolute residuals versus $\log_2(Cy3)$' scatter plot smoothed using *supsmu* and *lowess* and different values of the smoothing parameters *bass* and *f*, respectively. The figure illustrates the dependence of smoothing effect from magnitude of smoothing parameters. *bass* is control of the low frequency emphasis when using cross validation. The larger the value of *bass* (up to ten), the smoother the fit from automatic span selection [51,52,63]. *f* is fraction of the data used for smoothing at each $\log_2(Cy3)$ point. The larger the *f* value, the smoother the fit [51,52,63].

adjusted for Gaussian efficiency relative to the line of equivalence ($Cy5 = Cy3$) serve as a probabilistic tool for identifying outliers. Our approach uses robust scatter plot smoothing techniques to simultaneously diagnose and quantify the variance structure of the data and allow natural accommodation of heteroscedasticity in the identification of outliers. This

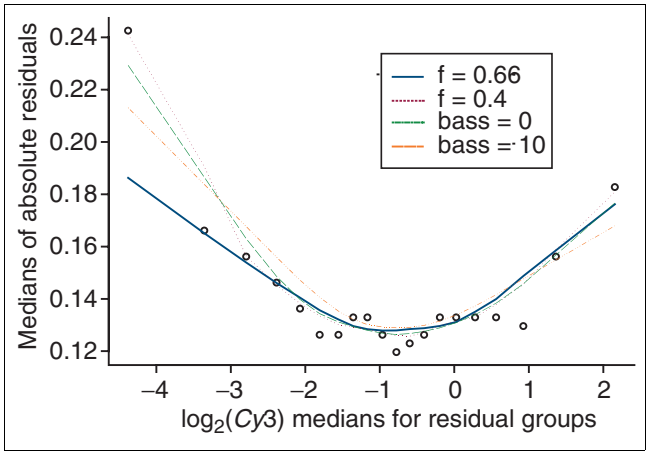

Figure 12
The use of smoothed absolute residuals for sequential intensity intervals. 'Absolute residuals versus $\log_2(Cy3)$' scatter plot based on *supsmu* and *lowess* for 20 sequential intervals of equal size and using different values of smoothing parameters (see legend to Figure 11 for details). Data are shown with higher resolution than in Figure 13.

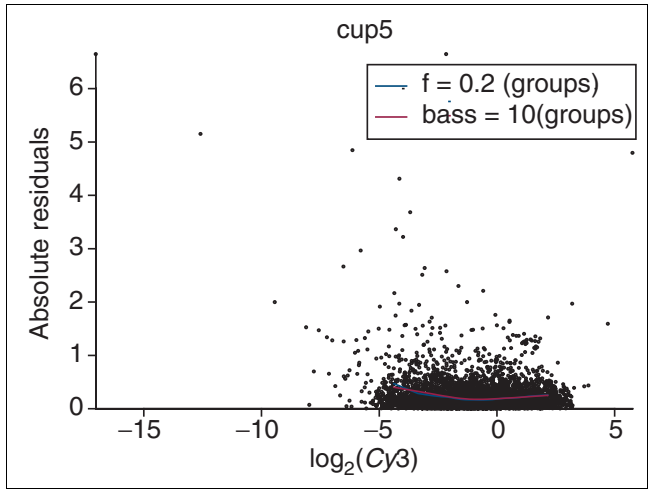

Figure 13
The use of smoothed absolute residuals for sequential intensity intervals. 'Absolute residuals versus $\log_2(Cy3)$' scatter plot with *supsmu* and *lowess* smoothing based on 20 sequential intervals of equal size with shown values of smoothing parameters (see legend to Figure 11 for details). Scale is different from Figure 12.

post hoc approach makes sense especially in view of the large sample size common in microarray experiments.

α-Outlier-generating model can be extended to multiple slide studies

We can extend our adjusted smoothed STI approach to datasets with multiple levels of replication. This provides a consistent method for experiments with and without replication. It is not clear how extant single-slide methods could be adapted for multiple-slide comparisons. Usually, two methods, each with their own data models and assumptions - one for single-slide and a second for a multiple-slide based method, are used.

Transformations versus interpretation of microarray datasets

Our method is based on limited data transformations (for example, background subtraction, log-transformation and global channel normalization) designed to preserve the data distribution and account for heteroscedasticity. A variety of non-linear transformation methods can be used to remove heteroscedasticity, for example, variance-stabilizing monotonic continuous non-linear transformations [21-23,44]. Equalizing residual variance in this manner does not guarantee that bivariate normality will be preserved for the majority of genes which are not differentially expressed. A model distribution assumption is especially important for statistical inference in the case of limited or no replication in the data. Non-linear transformation methods require preliminary research and computational experimentation with different types of transformations for each specific microarray dataset in order to make a choice between different transformations. Although transformation methods could represent a valuable

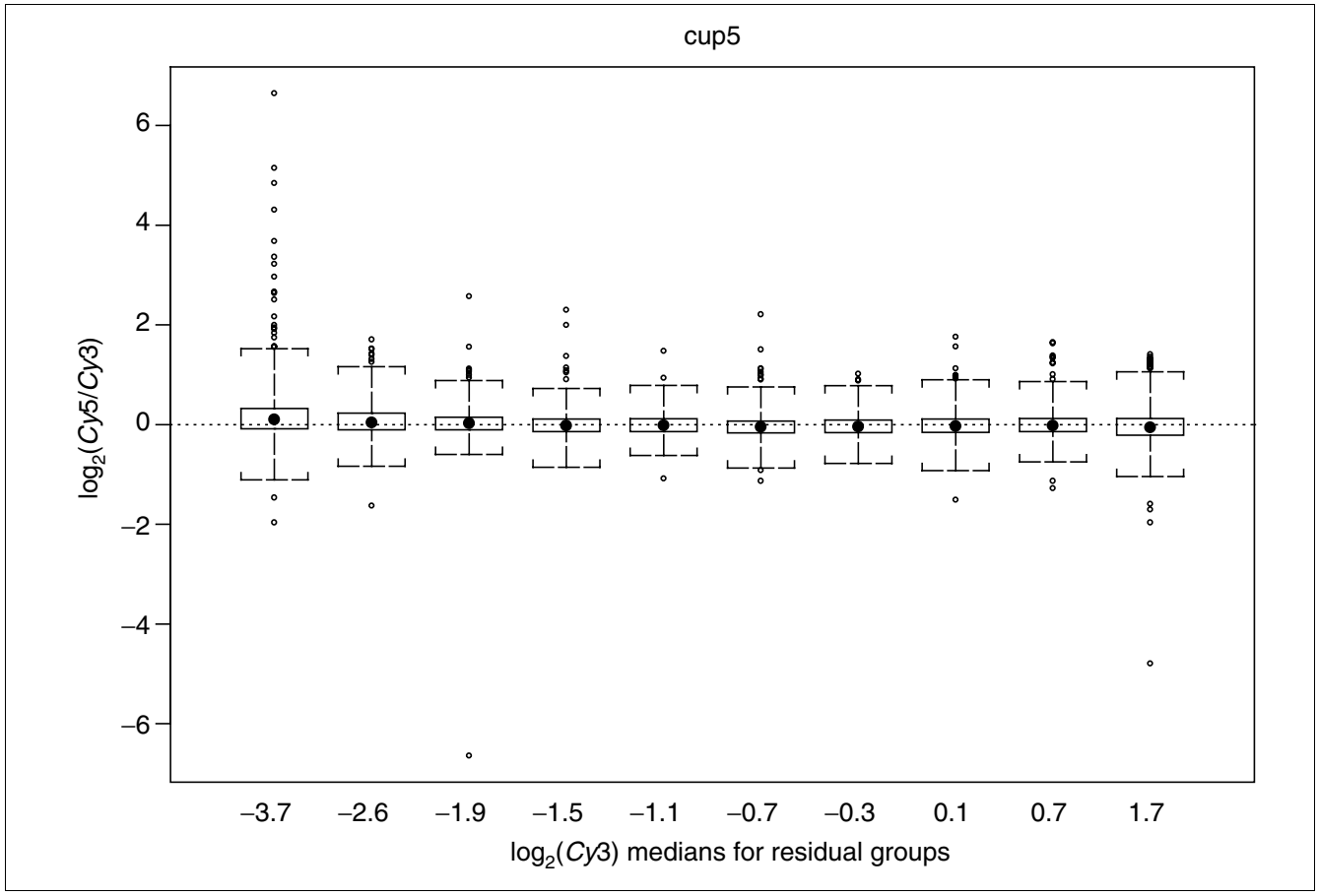

Figure 14
Boxplots for residuals using ten sequential intervals of equal size. A box corresponds to the *IQR* (inter-quartile range), the mid point is a sample median, and whiskers are 3*IQR* limits. Outliers (non-regular observations) are points outside the whiskers. Abscissa is based on medians for ten intervals of approximately equal size (the total sample size is 6,068, the first nine sets were 606, the tenth set was 614).

approach to microarray data analysis, any complex non-linear data transformation calls into question the validity of the transformations. Therefore, the application of these transformation methods requires trial and error followed by validation of each transformation for a particular experimental dataset [44]. We suggest that our approach, which relies on interpretation of existing data distributions including any heteroscedasticity rather than application of methods to change distributions, provides a reasonable alternative to variance-stabilizing methods.

Multiple comparisons in microarray data analysis

Typically, microarray data involve thousands of genes so clearly there is a problem of multiplicity of comparisons. Other model-based single-slide approaches do not consider this issue explicitly (see single-slide procedures described in [1,13,14,17,18]). First, we identify candidate outliers without correction to obtain unadjusted *p*-values (Table 3). A *p*-value is a probability to reject the null hypothesis when the null hypothesis is true and represents a measure of statistical significance in terms of false positive rate. One way to obtain adjusted *p*-values is to apply a Bonferroni correction based on *N* (the sample size of the entire dataset) which may be too conservative, so we examine two alternative corrections. In one alternative approach, we apply a multiplicity of comparison correction based on an estimate of *k* (number of non-regular observations) rather than the sample size of the entire dataset. This approach emphasizes stable outliers at the expense of other possible outliers (that is, *N*-*k*) which are inliers in the current single-slide experiment. Clearly, this Bonferroni correction by *k* provides a much less conservative result than the correction by *N* and we would argue more reasonable correction to identify true outliers. Other robust exploratory tools (see Methods) can be used to estimate *k*. In a more sophisticated approach to address these issues, the *q*-value is calculated from the ordered list of unadjusted *p*-values [45,46] (Figure 24). The *q*-value is the minimum false discovery rate [47] for a particular feature from a list of all

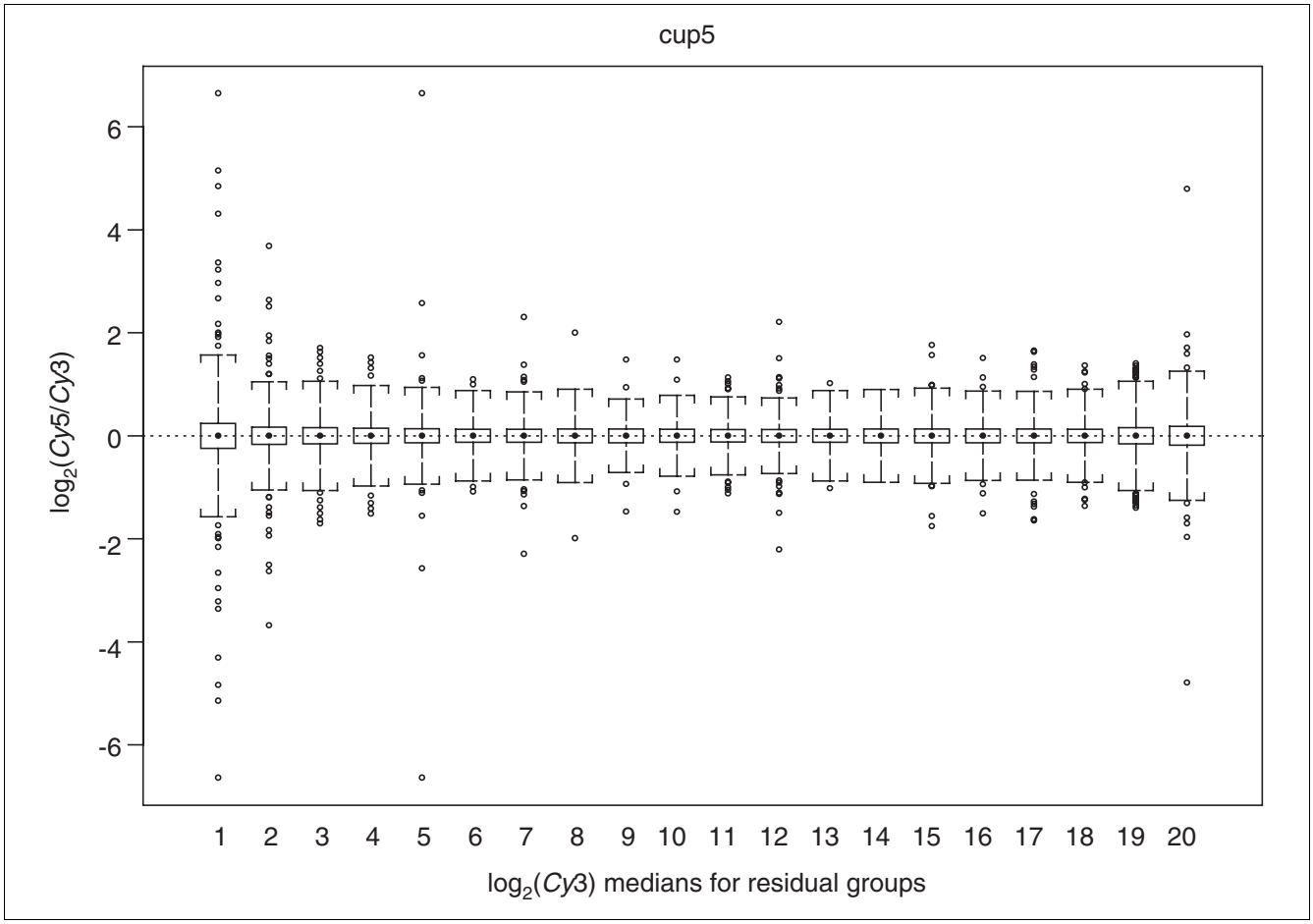

Figure 15
Boxplots for residuals using 20 sequential intervals of equal size. Details are the same as for Figure 14 but using 20 sequential intervals.

features [45,46]. The false discovery rate is the proportion of true null hypotheses among all null hypotheses which were found to be significant - for example, a false discovery rate of 1% means that among all candidates for differential expression found significant, 1% of these are true nulls on average [46].

Exploratory and confirmatory differential gene expression analysis

We suggest distinguishing explicitly between exploratory data analysis to identify candidates for differential gene expression and confirmatory analysis to identify differentially-expressed genes based on strict statistical inference. Exploratory differential gene expression analysis is appropriate for datasets with limited replication to identify the most likely candidates for differential expression. Clearly, additional independent experimental approaches or additional replicates are needed to confirm the exploratory analysis and distinguish outliers by chance from systematic outliers. Alternatively, confirmatory differential gene expression analysis with multiple layers of experimental and technical replication provides sound conclusions based solely on the microarray datasets. Nevertheless, exploratory microarray data analysis followed by independent confirmatory validation studies (for example, quantitative RT-PCR) represents a practical and cost effective solution for expression studies.

Methods

Transcript profiling data

The transcript profiling datasets examined in this study are from a published study that employed cDNA microarrays to compare gene expression in wild type *S. cerevisiae* and single gene deletion mutants [18]. Hughes *et al.* monitored 6,295 *S. cerevisiae* genes [18] in their study. For our analysis we used 6,068 of the 6,295 genes monitored. For each experiment, we used a Monte Carlo procedure to generate simulated datasets of N = 6,068 data points drawn from an 'ideal' bivariate

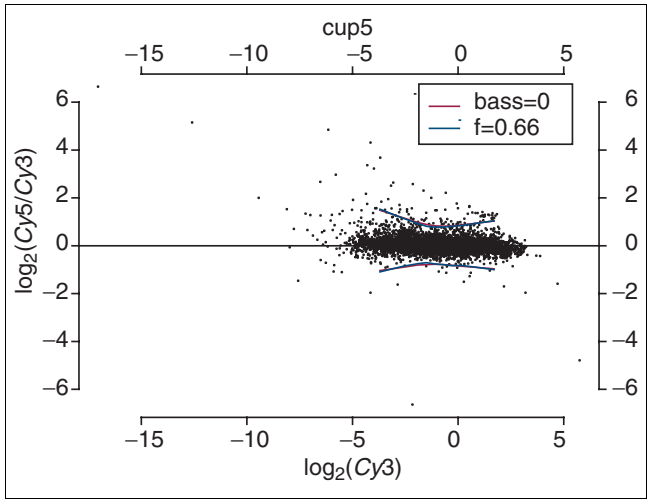

Figure 16
Diagnostics for heteroscedasticity pased on the upper and lower extremes computed as 3IQR for ten sequential intervals of equal size and using *supsmu* and *lowess* smoothers (compare with Figure 14).

normal population having the same parameters of location (means) and variance-covariance matrix as the empirical, observed data (the R/S-plus function *rmvnorm()*).

Data preprocessing

For each experiment, the observed data consisted of N = 6,068 pairs of *Cy3* and *Cy5* fluorescent intensity values which had been background corrected, tested for linearity, normalized by total signal intensity and log transformed [18], $\log_2(Cy3)_i$, $\log_2(Cy5)_i$, i = 1,... 6,068.

Exploratory data analysis

The empirical bivariate intensity data {$\log_2(Cy5)$, $\log_2(Cy3)$} and simulated datasets were examined using the following exploratory data analysis tools: concentration ellipses (*ellipse()* from package *ellipse* in R) and QQNP for residuals (*qqnorm()*).

Graphical tests for normality

Our approach assumes that the majority (more than 50%) of genes under consideration are not differentially expressed. We posit that changes in the expression levels of these genes are random, independent, non-directional and relatively small. These assumptions suggest that for the majority of data points, scatter plots '$\log_2(Cy5)$ versus $\log_2(Cy3)$' should exhibit bivariate normality, or univariate normality if $\log_2(Cy5/Cy3)$ is used as a measure of differential expression. Since we assume linearity (additivity), the data generally need to be normalized to remove non-linearity, if any, using global or print-tip group *lowess* method [48].

In addition to *mac1*, we examined data from nine other experiments, *mac1*, *cin5/YOR028C*, *cup5/YEL027W*, *fre6/YLL051C*, *sod1/YJR104C*, *spf1/YEL031W*, *vma8/YEL051W*, *yap1/YML007W*, *yer033c* and *ymr031c*. We tested these data for normality using three tools: scatter plots with concentration ellipses (tolerance ellipses), ordinary QQNPs for residuals, and QQNPs with simulation envelopes. The first and the third tools could be used to identify candidates for differential expression if the data are homoscedastic. We note that homoscedasticity may be imposed by applying variance-stabilizing transformations as discussed above (see Discussion).

For each experiment, we examined empirical and simulated data: firstly, log-transformed normalized channel intensities, $\log_2(Cy5)$ versus $\log_2(Cy3)$; secondly, channel intensities simulated as random samples from a bivariate normal population with the same location and scale parameters as in the first set of data; and thirdly, the same as the first set of data but after removing outliers defined as data points outside the 99.9% cut-off obtained using robust location and scale estimators. We used these tools rather than formal, analytical tests for univariate and bivariate normality [49,50] because the latter are sensitive to even small departures from normality if the sample size is large, that is, many thousands of data points.

Scatter plots with concentration ellipses (tolerance ellipses)

We overlay scatter plots of the experimental data with concentration ellipses for a bivariate normal distribution; the ellipses indicate curves of constant probability density for the standard bivariate normal population. The plots can be used as a visual test for bivariate normality and to investigate systematic and random deviations from it. As a diagnostic display for outliers [31], the plots can detect (projecting the data on the line $Y = -X$ [43]) regular observations, with internal X and well-fitting Y, and three types of outliers: only Y-outliers (vertical outliers with internal X and non-fitting Y), only X-outliers (outlying X and well-fitting Y) and both Y- and X-outliers (Y-X-outliers with outlying X and non-fitting Y).

In this work, we used sample medians and mads (medians of absolute deviations from the sample medians) to construct tolerance ellipses (Figures 2,3,4) but more sophisticated robust location and scale estimators can be utilized. For location, this includes a Huber M-estimator or Tukey's bi-square with 96% Gaussian efficiency [51,52] (*location.m()* in S-plus and *hubers()* from package *MASS* in R). For scale, this includes a Huber τ-estimate (*scale.tau()* in S-plus and *hubers()* in R), a bi-square A-estimate of scale (*scale.a()* in S-plus), which are 80% Gaussian efficient [51,52], or the use of MVE (*cov.mve()*/*plot.mve()*) and MCD (*cov.mcd()*/*plot.mcd()*) estimators.

QQNP for residuals

QQNP is a plot of the data sorted in ascending order compared with the corresponding quantiles of the standard

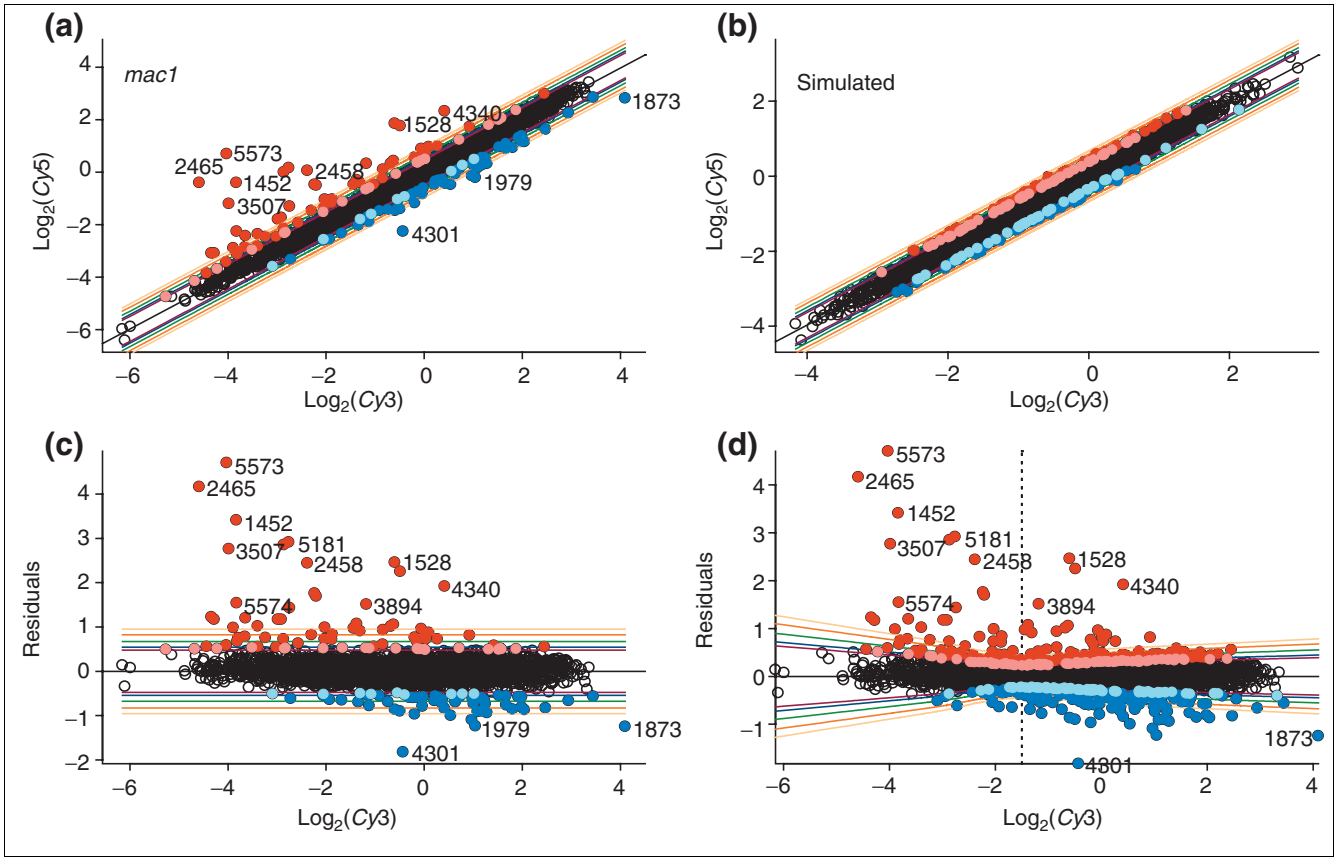

Figure 17
Five ordinary STIs for the **(a)** real and **(b)** simulated *mac1* datasets. The line of equivalence (black) has slope 1 and intercept 0 which corresponds to the case of *Cy*5 = *Cy*3. **(c,d)** Scatter plots of residuals for the real *mac1* dataset versus predictor variable. On this figure (c,d) and other figures 'residuals' mean '$\log_2(Cy5/Cy3)$'. The residuals are depicted in (c) and this is the same as (a) except that the linear trend has been subtracted resulting in a slope *A* = 0. The ordinary STIs assume residual homoscedasticity. (d) STIs shown corrected with the S-plus scatter plot smoother *supsmu* to reveal the dependence of residual variance on the value of the predictor variable. The *supsmu*-based STIs assume residual heteroscedasticity. Pink and cyan points lie in the interval between the upper and lower 95% and 99% STIs, respectively. Red and blue points lie above the upper and lower 99% STI, respectively. Black points lie below the upper and lower 95% STI. In all panels, the 95% (innermost), 99%, 99.8%, 99.98% and 99.998% (outermost) STIs are shown (covered with probability at least 0.9999). The vertical dotted line marks the location of the minima of the empirical hyperbolas. Therefore, red/pink, blue/cyan and black points represent up-regulated, down-regulated and unchanged genes, respectively.

normal distribution, that is, a normal distribution with mean zero and variance one [53,54]. An approximately linear plot signifies that the data are reasonably Gaussian. A U-shape suggests that the empirical distribution is skewed. A plot that is bent down on the left and bent up on the right denotes a distribution with 'heavier' (longer) tails than the standard normal. Although useful for the analysis of residual distribution, QQNP for residuals is not an effective tool for identifying outliers for three reasons. Firstly, there is no formal test to judge departures from the normal distribution [53]; secondly, residual heteroscedasticity, if any, is not considered; and thirdly, the residuals are linear combinations of random variables and tend to be more normal ('super-normality'), than the underlying error distribution. A possible solution for the latter is use of simulation envelopes [51,55,56].

QQNP with simulation envelopes for residuals

To enhance interpretation of QQNP, an approach based on simulation envelopes can be applied [51,55,56]. The simulation envelopes are obtained from randomly generated normal samples, which are standardized, sorted and then used to identify the maximum and minimum values to construct the upper and lower envelopes [51]. The simulation is repeated 1,000 times. For our implementation of the algorithm we used S-plus code developed by Venables and Ripley [51].

Data structure and outlier model

The term 'outlier' lacks a precise definition. Usually, outliers are interpreted as gross errors, or extreme, spurious, discordant, contaminating observations. In many contexts, outliers are undesirable data points. In our application, they are a

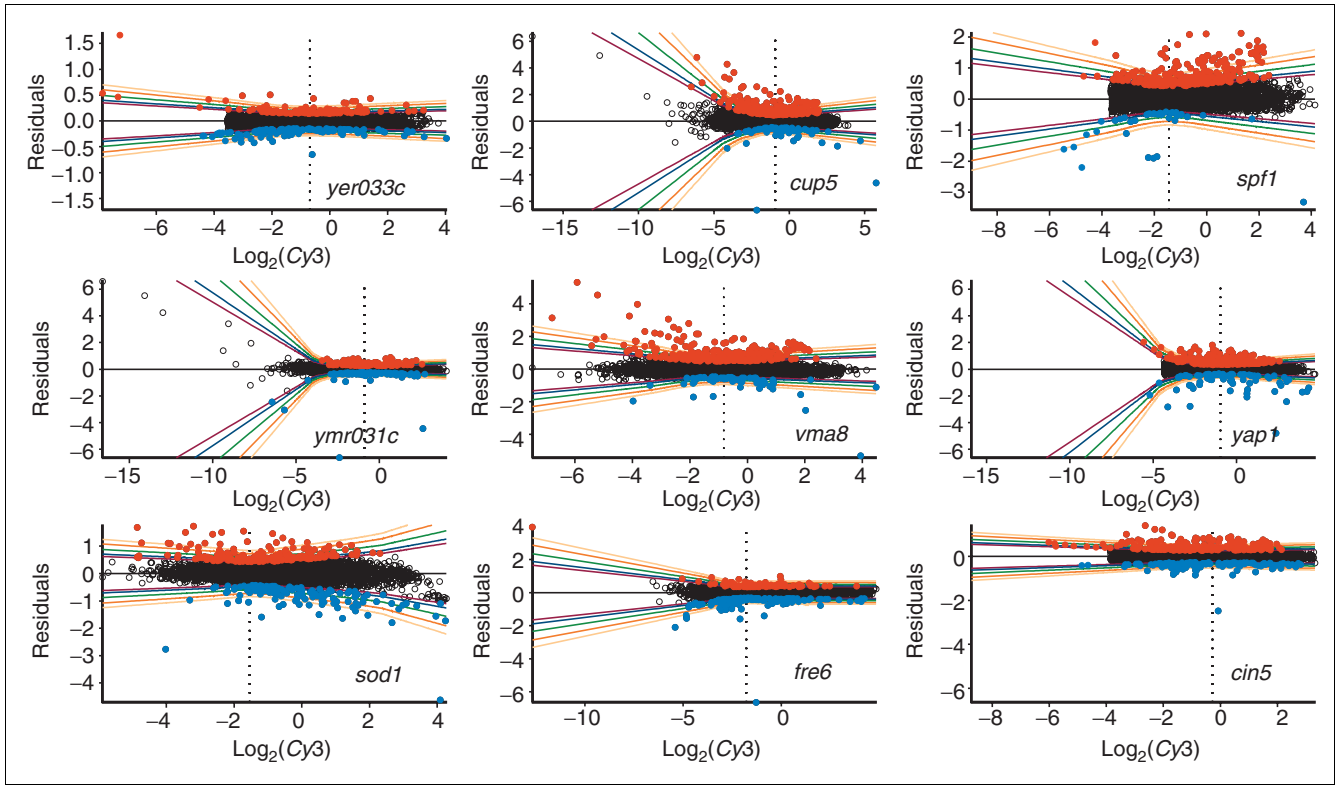

Figure 18
Residuals versus $X \equiv \log_2(Cy3)$ scatter plots for nine different cDNA microarray experiments. Each plot, *yer033c*, *cup5*, *spf1*, *ymr031c*, *vma8*, *yap1*, *sod1*, *fre6* and *cin5*, shows the 95% (innermost), 99%, 99.8%, 99.98% and 99.998% (outermost) adjusted *supsmu*-based STIs (covered with probability at least 0.9999). Red and blue dots mark up-regulated and down-regulated genes, respectively, with *p*-value ≤ 0.01 (≤ 0.05). The corresponding knockout genes are identified as the most prominent down-regulated ones (*cin5*, *sod1*, *spf1*, *vma8* and *yer033c*) or one of the most prominent down-regulated genes (*cup5*, *fre6*, *yap1* and *ymr031c*) with *p*-values always much less than 0.00002.

matter of considerable biological interest because they are candidates for differentially-expressed genes. Outlier-generating models [15,16] assume that a sample of size N contains N-k regular data points which are i.i.d. (independently and identically distributed) observations from a distribution F. The remaining k non-regular observations come from other distributions $D_1, ..., D_k$. If these distributions are defined as $D_i = F(\chi - \mu_i)$, or $D_i = F(\chi/\sigma_i)$ where $\mu_i > 0$, $\sigma_i > 1$, $i = 1, ..., k$, then the resulting models are known as location-slippage (Ferguson-type model) or scale-slippage, respectively. An alternative approach [15] connects outliers with their surprisingly extreme nature. Such a definition makes sense due to a logical relationship between extreme observations, outliers and contaminants [28]: extreme observations may or may not be outliers, outliers are always extreme observations, outliers may or may not be contaminants, and contaminants may or may not be outliers.

The following types of outliers in microarray experiments are possible: outliers by chance (due to finite sample sizes used), sporadic technical or biological outliers, and systematic outliers. Systematic outliers can be divided into reproducible technical outliers (for example, outliers due to heteroscedasticity), and biological outliers which contain both differentially-expressed genes, and genes with unusual high individual variability in expression. In general, only sufficient replication can distinguish differentially-expressed genes from other types of outliers. However, some types of outliers can be quantified in a single-slide experiment as we have shown in this article, in 'same versus same' hybridizations, or using loop design [4,5,57].

Data structure

The distribution F can be any unimodal symmetric distribution with positive density. We require that F be reasonably approximated by a normal distribution $N(0, \sigma^2(I))$ with a zero mean and an unknown intensity-dependent variance. Generally, the extreme contaminants may differ from the regular observations by their distributions but their location in the sample may overlap the bulk of regular data points. Outliers may depend on each other as well as on the regular observations.

A single-slide cDNA microarray experiment without spot replication generates $2N$ channel intensity values that can be reduced to N log-transformed ratios. Therefore, we can

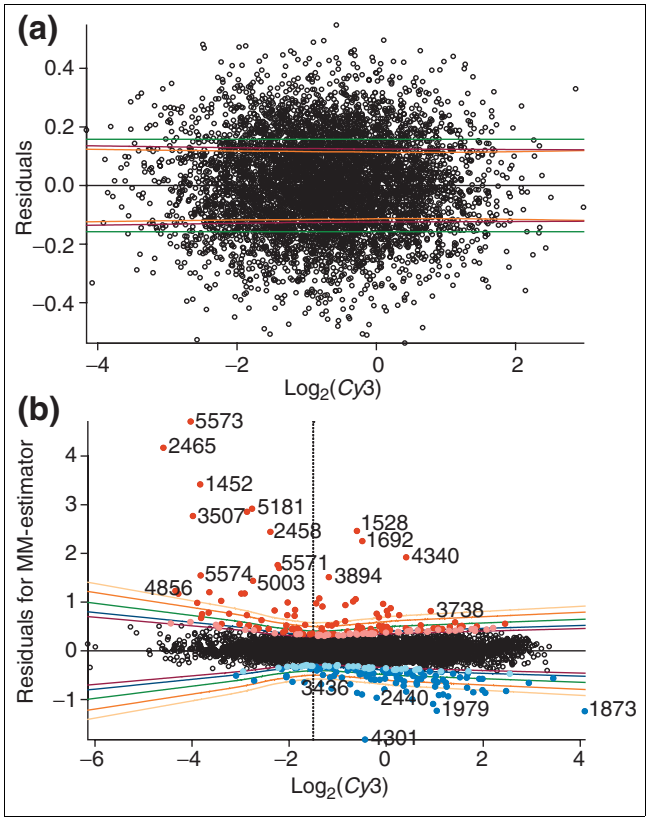

Figure 19

Scatter plots of residuals for *mac1* versus $\log_2(Cy3)$ as independent variables together with various STIs. **(a)** Residuals for artificial data drawn from a bivariate normal distribution with the same parameters as the real *mac1* data shown below. This plot is based on the same simulated dataset as in Figure 17b except that the linear trend has been subtracted resulting in slope = 0. A robust scale estimator (a Huber τ-estimate of scale) for residuals (green, outer), *supsmu* (purple, middle) and *lowess*-based (orange, inner) scale estimators are shown. Scale factors calculated to adjust *supsmu*- and *lowess*-based scale estimates were 1.25 and 1.35, respectively. **(b)** Adjusted *supsmu*-based STIs for the real data at the 95% (innermost), 99%, 99.8%, 99.98% and 99.998% (outermost) levels. These adjusted STIs take into account differences between the ordinary STIs and *supsmu*-based STIs for an artificial dataset having the same parameters (location, scale and coefficient of correlation) as its cognate real data. The vertical dotted line marks the location of the minima of the empirical hyperbolas. Red or pink and blue or cyan dots correspond to up-regulated and down-regulated genes, respectively, with *p*-value ≤ 0.01 (≤ 0.05). The most prominent down-regulated gene with case index number 4,301 is Mac1 (see also Table 3), which is not expressed in the *mac1* strain.

consider each single-slide outcome as a sample of N size. As a result, in single-slide experiments the number of genes is the sample size. In M replicated cDNA microarray experiments we have a sample of size M (log-transformed ratios) for every gene. Hence, we can consider a single-slide experiment without spot replication as a special case of the multiple-slide design with M technical replicates when $M = 1$. In this case, each gene could be considered as the sampling unit rather than each array.

Outlier model

We define the following outlier model [15,16], in which an observation is equivalent to $\log_2(Cy5/Cy3)$.

Consider a random sample of size N consisting of n regular observations and k non-regular observations or α'-outliers with respect to H_0, $N = n + k$, and $\alpha' = 1 - (1 - \alpha)^{1/N}$, and $k < N/2$. In other words, α is a significance level without adjustment for multiplicity of comparisons while α' is based on a correction for multiple comparisons. The $n = N - k$ regular observations are i.i.d. data points drawn from the same underlying distribution of a parametric family. For example, H_0 distribution could be normal $N(\mu, \sigma^2)$, for which the parameters μ and σ^2 are unknown, and k is unknown as well (with the only restriction that $k < N/2$). The k non-regular observations have unknown individual distributions: $D_1, \ldots, D_k$, that may depend on each other as well as the regular observations.

As an outlier identification rule we use a one-step procedure to detect an unknown number of outliers. The approach can be applied to univariate, bivariate and multivariate samples [16]. Given the random sample of N observations, the task is to determine whether the point is an α'-outlier with respect to the underlying normal distribution $N(\mu, \sigma^2)$. The estimated outlier region is specified by lower and upper bounds which are functions of N and α'. All points that are less (greater) than the lower (upper) bound will be in the outlier region being identified as α'-outliers. For the 'normalizing condition' [15,16]:

P (no outliers amongst N i.i.d. data points from a normal distribution) = $1 - \alpha$ (1)

Therefore, a regular observation can be identified as an α-outlier but with probability α only. For the univariate homoscedastic case, consider the value $t_i = |\log_2(Cy5/Cy3)_i - m|/s$, where m and s are location and scale estimators, respectively. An outlier can be defined as a point i for which $t_i \geq c_N(\alpha')$, where $c_N(\alpha')$ is a cut-off that is a function of N and α'. The choice of m and s is highly important for the performance and robust versions of these estimators are preferable [15,16].

A comparison of some specific outlier identification rules based on performance criteria revealed that one-step procedures with robust estimates for location and scale (outlier resistant rules) are superior at identifying outliers [16]).

The outlier model we used is only completely specified when we know distribution(s) for regular observations and distributions of the contaminants [16]. From this one can conclude that in the absence of information about the distributions of the contaminants, data analysis can be just explorative; for example, without replication it is impossible to differentiate between systematic outliers due to differential expression or other types of outliers.

Table 3

Candidate differentially-expressed genes

RN	ORF Name	Synonyms	CIN	$\log_2(Cy3)$	$\log_2(Cy5/Cy3)$	RSE	SR	*p*	*p*'	*p*''	*q*
Up-regulated genes											
1	YBR207W		427	-0.62	1.06	0.14	7.30	3.30E-13	2.00E-09	2.01E-11	1.11E-10
2	YBR295W	*PCA1, PAY2*	515	-2.30	0.83	0.16	5.26	1.50E-07	9.13E-04	9.18E-06	2.47E-05
3	YDR264C	*AKR1*	1186	0.10	0.87	0.17	5.22	1.81E-07	1.10E-03	1.11E-05	2.90E-05
4	YDR270W	*CCC2*	1192	-2.98	1.19	0.21	5.75	9.09E-09	5.52E-05	5.54E-07	1.62E-06
5	YDR476C		1394	-1.87	0.84	0.14	6.08	1.24E-09	7.51E-06	7.55E-08	2.78E-07
6	YDR534C		1452	-3.83	3.43	0.25	13.66	0.00E+00	0.00E+00	0.00E+00	0.00E+00
7	YEL065W	*SIT1*	1528	-0.60	2.46	0.15	16.89	0.00E+00	0.00E+00	0.00E+00	0.00E+00
8	YER145C	*FTR1*	1692	-0.49	2.25	0.15	15.11	0.00E+00	0.00E+00	0.00E+00	0.00E+00
9	YFL041W	*FET5*	1778	-0.84	0.94	0.14	6.76	1.54E-11	9.32E-08	9.37E-10	4.05E-09
10	YGL015C		1879	-3.40	1.04	0.23	4.53	6.07E-06	3.69E-02	3.71E-04	7.23E-04
11	YGL039W		1903	-0.68	1.01	0.14	7.06	1.80E-12	1.10E-08	1.10E-10	5.76E-10
12	YGR065C		2189	-0.19	0.76	0.16	4.82	1.49E-06	9.01E-03	9.06E-05	1.80E-04
13	YHL035C		2453	-2.20	1.70	0.15	11.07	0.00E+00	0.00E+00	0.00E+00	0.00E+00
14	YHL040C	*ARN1*	2458	-2.38	2.45	0.16	14.95	0.00E+00	0.00E+00	0.00E+00	0.00E+00
15	YHL047C		2465	-4.59	4.19	0.29	14.43	0.00E+00	0.00E+00	0.00E+00	0.00E+00
16	YHR042W	*NCP1*	2511	0.27	0.72	0.17	4.27	2.00E-05	1.21E-01	1.22E-03	2.01E-03
17	YHR175W	*CTR2*	2644	-1.38	1.08	0.13	8.32	0.00E+00	0.00E+00	0.00E+00	0.00E+00
18	YJL153C	*INO1, APR1*	3057	-2.03	0.74	0.15	5.06	4.33E-07	2.63E-03	2.64E-05	6.25E-05
19	YKL220C	*FRE2*	3507	-3.98	2.78	0.26	10.75	0.00E+00	0.00E+00	0.00E+00	0.00E+00
20	YLL051C	*FRE6*	3668	-1.44	0.97	0.13	7.55	4.97E-14	3.02E-10	3.03E-12	1.78E-11
21	YLR034C	*SMF3*	3716	0.04	0.72	0.16	4.41	1.04E-05	6.32E-02	6.35E-04	1.11E-03
22	YLR046C		3728	-1.88	0.72	0.14	5.19	2.21E-07	1.34E-03	1.35E-05	3.44E-05
23	YLR056W	*ERG3, SYR1*	3738	0.92	0.80	0.18	4.47	8.04E-06	4.88E-02	4.90E-04	8.71E-04
24	YLR126C		3808	-2.03	1.00	0.15	6.85	8.07E-12	4.90E-08	4.92E-10	2.23E-09
25	YLR136C	*TIS11, CTH2*	3818	-1.31	0.91	0.13	7.02	2.41E-12	1.46E-08	1.47E-10	7.30E-10
26	YLR205C		3885	-2.91	1.19	0.20	5.87	4.47E-09	2.71E-05	2.72E-07	8.47E-07
27	YLR214W	*FRE1*	3894	-1.18	1.51	0.13	11.62	0.00E+00	0.00E+00	0.00E+00	0.00E+00
28	YMR006C	*PLB2*	4286	-0.84	0.61	0.14	4.39	1.16E-05	7.06E-02	7.10E-04	1.22E-03
29	YMR011W	*HXT2*	4291	0.02	0.81	0.16	4.96	7.28E-07	4.42E-03	4.44E-05	1.00E-04
30	YMR058W	*FET3*	4340	0.42	1.91	0.17	11.10	0.00E+00	0.00E+00	0.00E+00	0.00E+00
31	YMR251W		4540	-4.28	1.19	0.27	4.33	1.48E-05	9.00E-02	9.05E-04	1.53E-03
32	YNL237W	*YTP1*	4856	-4.34	1.25	0.28	4.50	6.86E-06	4.16E-02	4.18E-04	7.95E-04
33	YNL259C	*ATX1*	4878	-0.06	0.96	0.16	5.95	2.90E-09	1.76E-05	1.77E-07	5.86E-07
34	YNR056C	*BIO5*	5003	-2.74	1.45	0.19	7.60	3.38E-14	2.05E-10	2.06E-12	1.28E-11
35	YNR060W	*FRE4*	5007	-3.64	1.22	0.24	5.03	4.96E-07	3.01E-03	3.02E-05	7.00E-05
36	YOL158C		5181	-2.76	2.93	0.19	15.28	0.00E+00	0.00E+00	0.00E+00	0.00E+00
37	YOR334W	*MRS2*	5524	-0.15	0.68	0.16	4.27	1.95E-05	1.18E-01	1.19E-03	1.97E-03
38	YOR381W	*FRE3*	5571	-2.23	1.77	0.16	11.37	0.00E+00	0.00E+00	0.00E+00	0.00E+00
39	YOR382W		5572	-2.86	2.86	0.20	14.42	0.00E+00	0.00E+00	0.00E+00	0.00E+00
40	YOR383C		5573	-4.03	4.72	0.26	18.07	0.00E+00	0.00E+00	0.00E+00	0.00E+00
41	YOR384W	*FRE5*	5574	-3.82	1.56	0.25	6.23	5.11E-10	3.10E-06	3.12E-08	1.29E-07
Down-regulated genes											
1	YBR054W	*YRO2*	273	1.28	-0.95	0.19	-5.12	3.11E-07	1.89E-03	1.90E-05	4.72E-05
2	YBR147W		366	-1.68	-0.64	0.13	-4.87	1.13E-06	6.86E-03	6.90E-05	1.40E-04
3	YCL030C	*HIS4*	547	1.80	-0.87	0.19	-4.50	6.95E-06	4.22E-02	4.24E-04	7.95E-04
4	YDL171C	*GLT1*	857	-0.59	-0.87	0.15	-5.95	2.81E-09	1.70E-05	1.71E-07	5.86E-07

Table 3 *(Continued)*

Candidate differentially-expressed genes

5	YEL039C	*CYC7*	1502	-1.34	-0.69	0.13	-5.31	1.14E-07	6.91E-04	6.95E-06	1.92E-05
6	YER174C	*GRX4*	1721	1.09	-0.92	0.18	-5.06	4.31E-07	2.61E-03	2.63E-05	6.25E-05
7	YFL014W	*HSP12, GLP1, HOR5*	1751	1.18	-0.91	0.18	-4.95	7.57E-07	4.60E-03	4.62E-05	1.02E-04
8	YFR030W	*MET10*	1836	0.54	-1.01	0.17	-5.79	7.61E-09	4.62E-05	4.64E-07	1.40E-06
9	YFR055W		1862	-0.74	-0.69	0.14	-4.90	1.00E-06	6.07E-03	6.11E-05	1.27E-04
10	YGL009C		1873	4.10	-1.27	0.23	-5.58	2.54E-08	1.54E-04	1.55E-06	4.40E-06
11	YGL117W		1979	1.05	-1.25	0.18	-6.85	8.05E-12	4.89E-08	4.91E-10	2.23E-09
12	YGR088W	*CTT1*	2212	-1.93	-0.63	0.14	-4.48	7.50E-06	4.55E-02	4.57E-04	8.34E-04
13	YGR286C	*BIO2*	2409	0.97	-1.10	0.18	-6.11	1.09E-09	6.61E-06	6.65E-08	2.54E-07
14	YHL021C		2440	-0.19	-0.97	0.16	-6.17	7.25E-10	4.40E-06	4.42E-08	1.76E-07
15	YJR137C	*ECM17, MET5*	3263	0.42	-0.85	0.17	-4.92	8.71E-07	5.28E-03	5.31E-05	1.12E-04
16	YKL148C	*SDH1*	3436	-1.12	-0.77	0.13	-5.91	3.56E-09	2.16E-05	2.17E-07	6.97E-07
17	YLL041C	*SDH2*	3658	-0.49	-0.90	0.15	-6.05	1.50E-09	9.10E-06	9.15E-08	3.25E-07
18	YLR304C	*ACO1, GLU1*	3985	2.05	-0.88	0.20	-4.48	7.56E-06	4.59E-02	4.61E-04	8.34E-04
19	YMR021C	*MAC1, CUA1*	4301	-0.43	-1.83	0.15	-12.17	0.00E+00	0.00E+00	0.00E+00	0.00E+00
20	YOR356W		5546	-0.03	-0.80	0.16	-4.93	8.45E-07	5.13E-03	5.16E-05	1.12E-04

Candidates for differentially-expressed genes in the *mac1* cDNA microarray experiment defined as those outside the adjusted *supsmu*-based 99.998%-STIs (unadjusted p-value ≤ 0.00002). The study monitored transcripts in a *mac1* knockout and wild-type *S. cerevisiae*. RN, row number; ORF name, systematic name for open reading frame; Synonyms, alternative gene names (if any); CIN, case index numbers used in scatter plots to identify data points; $\log_2(Cy3)$, predictor variable (log base 2 transformed *Cy*3 intensity value); $\log_2(Cy5/Cy3)$, log base 2 of *Cy*5/*Cy*3 ratio (here ratios are residuals); RSE, residual scale estimator (for data points represented in the table it is based on *supsmu* non-parametric regression smoother); SR, standardized residuals ($\log_2(Cy5/Cy3)$ divided by RSE); p, unadjusted p-values based on statistics for two-sided tolerance intervals; p', Bonferroni adjusted p-values using N = 6,068; p'', Bonferroni adjusted p-values using k = 61, q, q-values calculated using *qvalue()* function [45,46] (see Figure 24 supporting q-value calibration). p- and q-values are given in scientific notation and 0.00E+00 means that a value was less than 10^{-16}.

Ordinary STIs

The relationship between (residual) outlier test statistics and tolerance regions [27,29,30] has shown a need for using STIs to detect Y-outliers in regression. For probability at least (1 - γ), the central tolerance intervals (STIs are centered about 0) that are simultaneous in X and q can be determined using:

$$\pm s\{(2F(2,N-2,1-\gamma/2))^{1/2}[\frac{1}{N}+\frac{(X-\bar{X})^2}{\sum_{i=1}^{N}(X_i-\bar{X})^2}]^{1/2}+\lambda(q)(\frac{(N-2)}{\chi^2(N-2,\gamma/2)})^{1/2}\} \quad (2)$$

where $\lambda(q)$ is a two-sided quantile for the standard normal distribution, P = 100%q is portion of the population to be covered with the STI, γ is a confidence level, $\chi^2(N - 2, \gamma/2)$ is the lower $\gamma/2$ quantile of a χ^2-distribution with N - 2 degrees of freedom, $F(2, N - 2, 1 - \gamma/2)$ is the upper (1 - $\gamma/2$)-quantile of F-distribution with 2 and N - 2 degrees of freedom. Equation 2 uses the Bonferroni inequality to evaluate how far out on the tail of its distribution each observation lies. It is based on formula 6.5 discussed in [29,30]. The corresponding formulae for sample mean and residual variance are:

$$\bar{X}=\sum_{i=1}^{N}X_i/N \quad (3)$$

$$s^2=\sum_{i=1}^{N}Y_i^2/(N-2) \quad (4)$$

In our case, $X_i = \log_2(Cy3_i)$ and residuals $Y_i = \log_2(Cy5/Cy3)$, where i = 1,..., N. Because $\log_2(Cy5_i)$ and $\log_2(Cy3_i)$ are highly correlated, $X_i = \log_2(Cy5_i)$ or $X_i = \log_2(Cy3_i + Cy5_i)/2$ could be used as well. We used scatter plots 'residuals versus average spot intensity' to compute p-values. For figures we used 'residuals versus $\log_2(Cy3_i)$' scatter plots.

We used robust estimators for the location (robust options: sample median, Huber M-estimator, Tukey's bi-square) and for the scale (*supsmu* or *lowess* fits for $s^2 = f(X)$ adjusted with simple Monte Carlo simulations to guarantee approximate Gaussian efficiency). We computed five STIs with five different portions of the normal distribution: 95%, 99%, 99.8%, 99.98% and 99.998%, correspondingly, and covered with probability at least 1 - γ = 99.99%. The corresponding interval estimates for p-values using, for example, formula 8 to approximate sample distribution for residuals under the null hypothesis are: 0.05 <p, 0.01 <p ≤ 0.05, 0.002 <p ≤ 0.01, 0.0002 <p ≤ 0.002, 0.00002 <p ≤ 0.0002 and p ≤ 0.00002. 'p-value' here is the chance or probability that the tolerance interval constructed from a single sample will not include the true inlier or regular observation. Alternatively, it is the

Table 4

Comparison of genes identified as differentially expressed in *mac1Δ*

Gene	ORF	Loguinov *et al.* (this work) (0.0001)	Hughes *et al.* [18] (0.0001)	Churchill and Sapir [20] (0.9999)	Newton et al. [19] (0.9999)	2FC	Reference
FTH1	***YBR207W***	+	+	+	+	+	[67]
PCA1	***YBR295W***	+	a	+	+	a	[68]
AKR1	***YDR264C***	+	a	+	+	a	[69]
CCC2	***YDR270W***	+	+	+	+	+	[67]
FIT1	***YDR534C***	+	a	+	+	+	[70]
SIT1	***YEL065W***	+	+	+	+	+	[70]
FTR1	***YER145C***	+	+	+	+	+	[67]
FET5	***YFL041W***	+	a	+	+		[71]
	YFR055W*	-	b	b	b	b	[72]
CTT1	***YGR088W****	-	b	b	b	b	[33]
VMR1	***YHL035C***	+	+	+	+	+	[69]
ARN1	***YHL040C***	+	+	+	+	+	[70]
ARN2	***YHL047C***	+	+	+	+	+	[70]
FRE2	***YKL220C***	+	a	+	+	+	[67]
MRS4	***YKR052C****	+	a	a	a	a	[73]
FRE6	***YLL051C***	+	a	+	+	a	[74]
SMF3	***YLR034C****	+	a	a	a	a	[75]
TIS11	***YLR136C***	+	a	+	+	a	[76]
HMX1	***YLR205C***	+	+	+	+	+	[77]
FRE1	***YLR214W***	+	+	+	+	+	[67]
MAC1	***YMR021C***	-	a	-	-	-	[33]
FET3	***YMR058W***	+	+	+	+	+	[67]
ATX1	***YNL259C***	+	a	+	+	a	[78]
BIO5	***YNR056C***	+	a	+	+	+	[69]
FRE4	***YNR060W***	+	a	+	+	+	[74]
ARN4	***YOL158C***	+	+	+	+	+	[70]
FRE3	***YOR381W***	+	+	+	+	+	[74]
FIT2	***YOR382W***	+	+	+	+	+	[79]
FIT3	***YOR383C***	+	+	+	+	+	[79]
FRE5	***YOR384W***	+	+	+	+	+	[74]

The performance of each method (except Chen *et al.* [17]) at equivalently high stringency levels is presented for selected genes likely to be differentially expressed in *mac1Δ* as described in text. For Sapir and Churchill [20], the cut-off (0.9999) corresponds to posterior probability of being differentially expressed and for Newton *et al.* [19], the cut-off 0.9999 is posterior probability of true differential expression. For this work, 0.0001 is a cut-off for *q*-values. For Hughes *et al.* [18], 0.0001 is a cut-off for *p*-values. 2FC is indicates two-fold change approach. A (+) sign indicates that a gene was identified as up-regulated by a method at this level of stringency while a (-) sign indicates that it was identified as down-regulated. Genes not identified by one method that were identified by other methods are indicated by a (+) or b (-) in the appropriate column. Genes identified only by this work are asterisked.

probability of having observed our data, or more extreme data, when the null hypothesis is true. Therefore, the null hypothesis here is: 'a data point is an inlier from i.i.d. random sample of size *N*', and we assume the corresponding null distribution for inlying residuals to be a normal distribution with mean = 0 and unknown variance, which may not be a constant.

Simultaneous tolerance intervals for data with replication

If replicated data are available, we have several (for example, n_i) observations on Y at each point X : $Y_{ij} = \log_2 (Cy5/Cy3)_{ij}$, $i = 1,..., K, j = 1,..., n_i$. In terms of ANOVA it can be described as two-way layout (if all n_i are equal) [58,59].

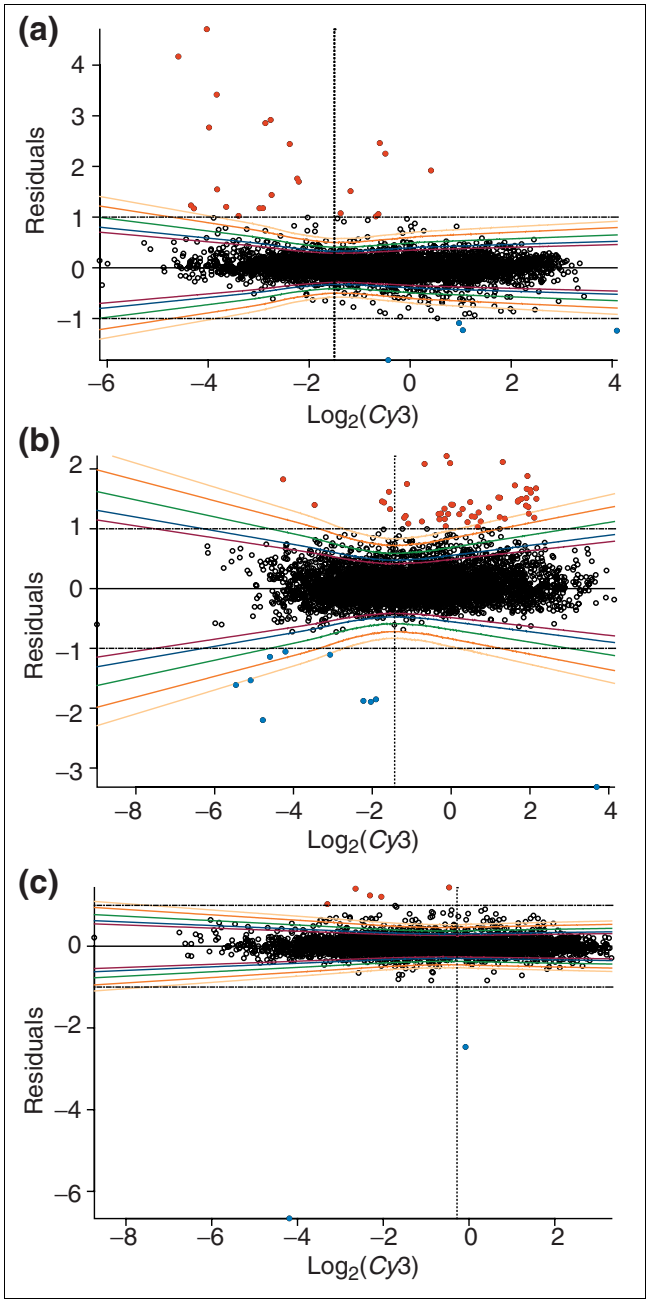

Figure 20
Comparison with a two-fold cut-off. Candidates for differentially-expressed genes defined using adjusted *supsmu*-based STIs or a threshold for ratios from the **(a)** *mac1*, **(b)** *spf1* and **(c)** *cin5* experiments. The dashed horizontal lines with intercepts of -1 and +1 correspond to two-fold changes in log-transformed (base 2) ratio. Red and blue dots denote genes up-regulated and down-regulated, respectively, according to this criterion. Moving away from the zero line, the 95%, 99%, 99.8%, 99.98% and 99.998% adjusted *supsmu*-based STIs are shown (covered with probability at least 0.9999). Further details about the *mac1* genes identified with case index numbers can be found in Table 3 (see also Table 2 for comparison of the adjusted *supsmu*-based STIs with other procedures).

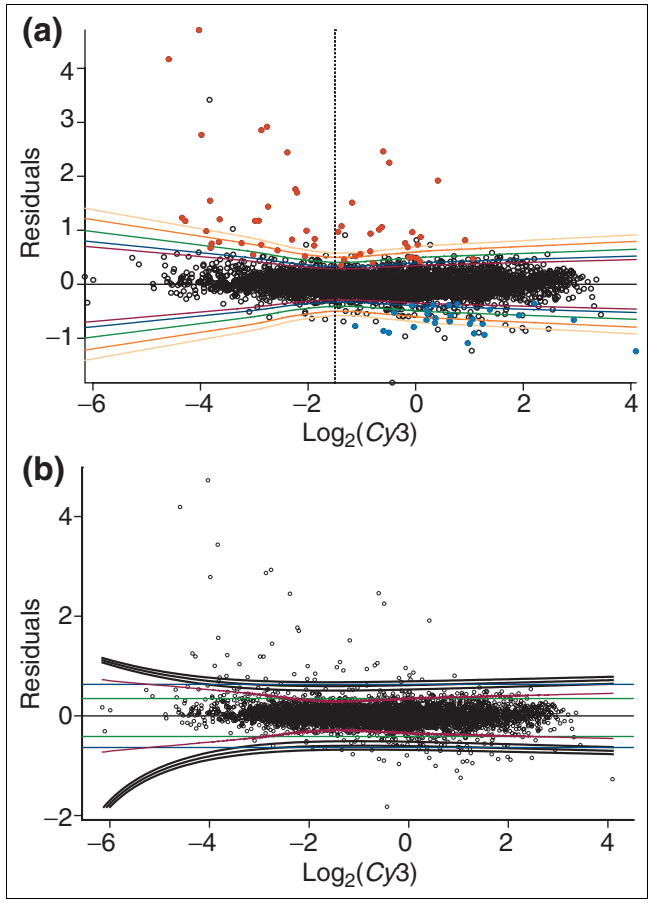

Figure 21
Comparison with other methods. **(a)** Comparison of *mac1* candidates for differentially-expressed genes defined using adjusted *supsmu*-based STIs and by Hughes *et al.* [18]. Red and blue dots denote genes designated as being up-regulated and down-regulated, respectively, by the Hughes *et al.* 'gene-specific' error model at p-value ≤ 0.05. The 95%, 99%, 99.8%, 99.98% and 99.998% adjusted *supsmu*-based STIs are shown (covered with probability at least 0.9999). The vertical dotted line marks the location of the minima of the empirical hyperbolas. **(b)** Comparison 95% adjusted *supsmu*-based STIs (red) with statistical intervals based on three other single-slide methods: a hierarchical Gamma-Gamma-Bernoulli model [19] with the posterior odds of change in expression 1:1 (innermost), 1:10 (middle) and 1:100 (outmost); mixture of orthogonal residuals with posterior probability of differential expression 95% (blue) based on approach from Sapir and Churchill [20]; and 95% band using asymmetric density function for raw ratios (*Cy*5/*Cy*3) assuming the same coefficient of variation for both fluors [17].

This information can be used when building STIs following an analogy with simultaneous prediction interval [58,59] (see also Equations 12 and 13 below):

$$\pm s\{(2F(2,N-2,1-\gamma/2))^{1/2}[\frac{1}{N}+\frac{(X-\bar{X})^2}{\sum_{i=1}^{K} n_i(X_i-\bar{X})^2}]^{1/2}+\lambda(q)(\frac{(N-2)}{\chi^2(N-2,\gamma/2)})^{1/2}\} \quad (5)$$

where

$$N=\sum_{i=1}^{K} n_i$$

Table 5

Comparison of performance with simulated datasets with 100 true positives

Cut-off points	PPV	NPV	Specificity	Sensitivity	Likelihood ratio
Sapir and Churchill [20]					
0.5	0.54	1.00	0.99	0.85	70.46
0.6	0.56	1.00	0.99	0.80	75.78
0.7	0.56	1.00	0.99	0.78	76.31
0.8	0.56	1.00	0.99	0.73	76.43
0.9	0.55	0.99	0.99	0.64	73.45
0.95	0.55	0.99	0.99	0.60	71.62
0.99	0.54	0.99	0.99	0.53	70.29
0.998	0.57	0.99	0.99	0.48	79.57
0.9998	0.58	0.99	0.99	0.43	82.78
0.99998	0.57	0.99	1.00	0.34	78.04
Newton *et al.* [19]					
0.43 (0.3)	0.62	1.00	0.99	0.76	96.50
0.67 (0.4)	0.61	0.99	0.99	0.70	92.84
1 (0.5)	0.61	0.99	0.99	0.66	91.60
2.33 (0.7)	0.61	0.99	0.99	0.59	95.17
4 (0.8)	0.60	0.99	0.99	0.52	91.28
5.67 (0.85)	0.60	0.99	0.99	0.50	90.42
9 (0.9)	0.61	0.99	0.99	0.49	94.33
19 (0.95)	0.62	0.99	0.99	0.49	97.48
99 (0.99)	0.63	0.99	1.00	0.44	101.00
499 (0.998)	0.60	0.99	1.00	0.36	89.52
Loguinov *et al.* (this work)					
0.5	0.40	1.00	0.98	0.99	39.65
0.25	0.60	1.00	0.99	0.93	90.99
0.2	0.63	1.00	0.99	0.91	102.47
0.15	0.68	1.00	0.99	0.90	127.89
0.1	0.72	1.00	0.99	0.87	157.34
0.05	0.76	1.00	1.00	0.82	188.22
0.01	0.79	1.00	1.00	0.76	226.78
0.002	0.80	0.99	1.00	0.60	238.72
0.0002	0.88	0.99	1.00	0.44	437.65
0.00002	0.89	0.99	1.00	0.40	477.44
Chen *et al.* [17]					
0.05	0.3	1.00	.97	1.00	37.30
0.01	0.45	1.00	.98	0.88	49.55

The first column lists the cut-off points used for each method for the performance comparison on simulated data as described in the text. For Sapir and Churchill [20] the cut-offs correspond to posterior probabilities of being differentially expressed. Similarly for Newton *et al.* [19] they correspond to posterior odds (probabilities) of true differential expression. For this work, the cut-offs for *q*-values are shown. Chen *et al.* [17] have two cut-offs which are approximated with a polynomial fit and are not shown in Figure 24 because there are no approximations available for other cut-offs. PPV is positive predictive value which equals TP (true positive)/TP + FP (false positive). NPV is negative predictive value which equals TN (true negative)/TN + FN (false negative). Sensitivity is TP/(TP + FN). Specificity is TN/(FP + TN). The likelihood ratio (Bayes' factor) is Sensitivity/(1-Specificity). For computations we used simulated data shown in Figure 22.

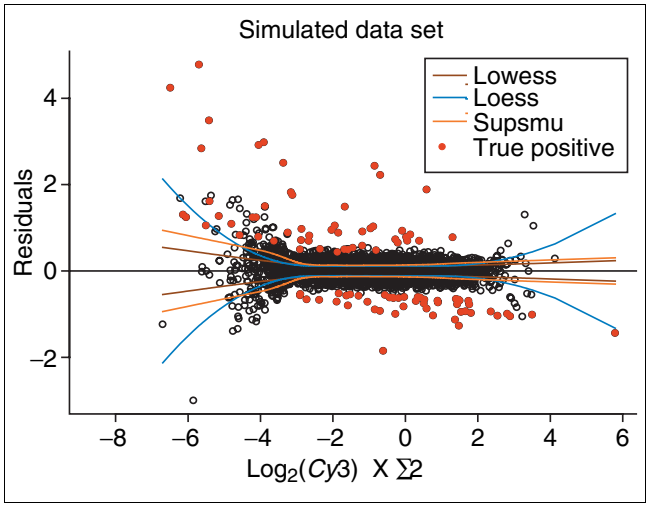

Figure 22
Simulated dataset for method performance testing. Non-regular observations (k = 100) and regular observations (N-k = 6,068-100 = 5,968) as the main body of non-differentially-expressed genes were used for the simulations. Sample parameters are taken from the *mac1* dataset. A random component was added to each outlier value using standard normal distribution with variance dependent on intensity. Heteroscedasticity for regular observations was also simulated by including intensity dependent variability in the low and high intensity levels. Three non-parametric smoothing methods were used to check absolute residuals for heteroscedasticity: *supsmu*, *lowess* and *loess* (the latter was based on a locally-quadratic fitting).

is the total number of observations, n_i is the number of replicates at point X_i,

$$\bar{X} = \sum_{i=1}^{K} n_i X_i / \sum_{i=1}^{K} n_i \qquad (6)$$

$$s^2 = (\sum_{i=1}^{K}\sum_{j=1}^{n_i} Y_{ij}^2 / (\sum_{i=1}^{K} n_i - 2)) \qquad (7)$$

where $X_i = \log_2(Cy3_i)$, and residuals $Y_{ij} = \log_2(Cy5_{ij}/Cy3_{ij})$, i = 1,..., K and j = 1,..., n_i. If $n_i \equiv 1$ then $K \equiv N$ and Equation 5 coincides with Equation 2.

Given replicated data, the fitted linear model can be checked by breaking the residual sum of squares into two components, lack of fit sum of squares and pure error sum of squares provided that the pure error is approximately the same throughout the data [58].

Other approximations for tolerance intervals

A two-sided β-expectation tolerance interval is given by [60]:

$$\pm (s(I)\, K(N, \beta)) \qquad (8)$$

where $K(N,\beta) = (1 + 1/N)^{1/2}\, t(N\text{-}1,(1 - \beta)/2)$ and $t(N\text{-}1,(1 - \beta)/2)$ is a Student variable with N -1 degrees of freedom. As it is clear from tables for tolerance factors [60], β-expectation tolerance intervals defined by (8) coincide with β-content tolerance intervals for residuals of linear regression described with (2) if N is large (> 1000).

For very large sample sizes, the null distribution could be approximated by a normal approximation $N(0,s^2(I))$:

$$\pm (s(I)\, \lambda\, (q)) \qquad (9)$$

where $s(I)$ is intensity-dependent scale estimator and $\lambda(q)$ is a two-sided quantile (100%q is a portion of the population to be covered) of the $N(0,1)$ defined by $\Phi(\lambda) = (1 - \beta/2)^{1/N}$ ($\Phi(\lambda)$ is a cumulative distribution function for the standard normal distribution). The goal is to quantify $s(I)$ in the presence of strong outliers. Our solution is the use of robust scatter plot smoothers for absolute residuals which has been described previously [25] (see details of our implementation below).

Relationships between STIs and simultaneous prediction intervals

In cDNA microarrays, the total number of predictions to be made is unknown or subject to chance, or the number of prediction intervals to be estimated simultaneously is large. Given such conditions, STIs are preferred over simultaneous prediction intervals (SPIs) [30]. When N is large, as in our case, the corresponding STIs and SPIs are indistinguishable on the scatter plots. If the majority of residuals are normally distributed, then using Bonferrroni procedure [30] we have this expression for the residuals:

$$\pm \sqrt{\{t(N-2,1-\alpha/2k)\}^2 [1 + \frac{1}{N} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}] s^2} \qquad (10)$$

where k is the number of 'future observations' or the number of the null hypotheses tested (we test k null hypotheses about k outliers). When N is large then the second term under the square root is about 1 and the following approximation holds:

$$\pm \sqrt{\{t(N-2,1-\alpha/2k)\}^2 s^2} \qquad (11)$$

It is a horizontal band based on Bonferroni-corrected t-value and one can expect that (1 - α/2)100% of the residuals will lie in the interval in repeat runs under the same experimental conditions. Thus, observations with the residuals situated far from the horizontal band can be identified as outliers. Equations 10 and 11 assume an outside estimate for k. We note that Equation 11 coincides with Equation 8 (within a degree of freedom) if one takes Bonferroni correction into account. In previous work, we used SPIs rather than STIs to identify candidates for differential expression [61].

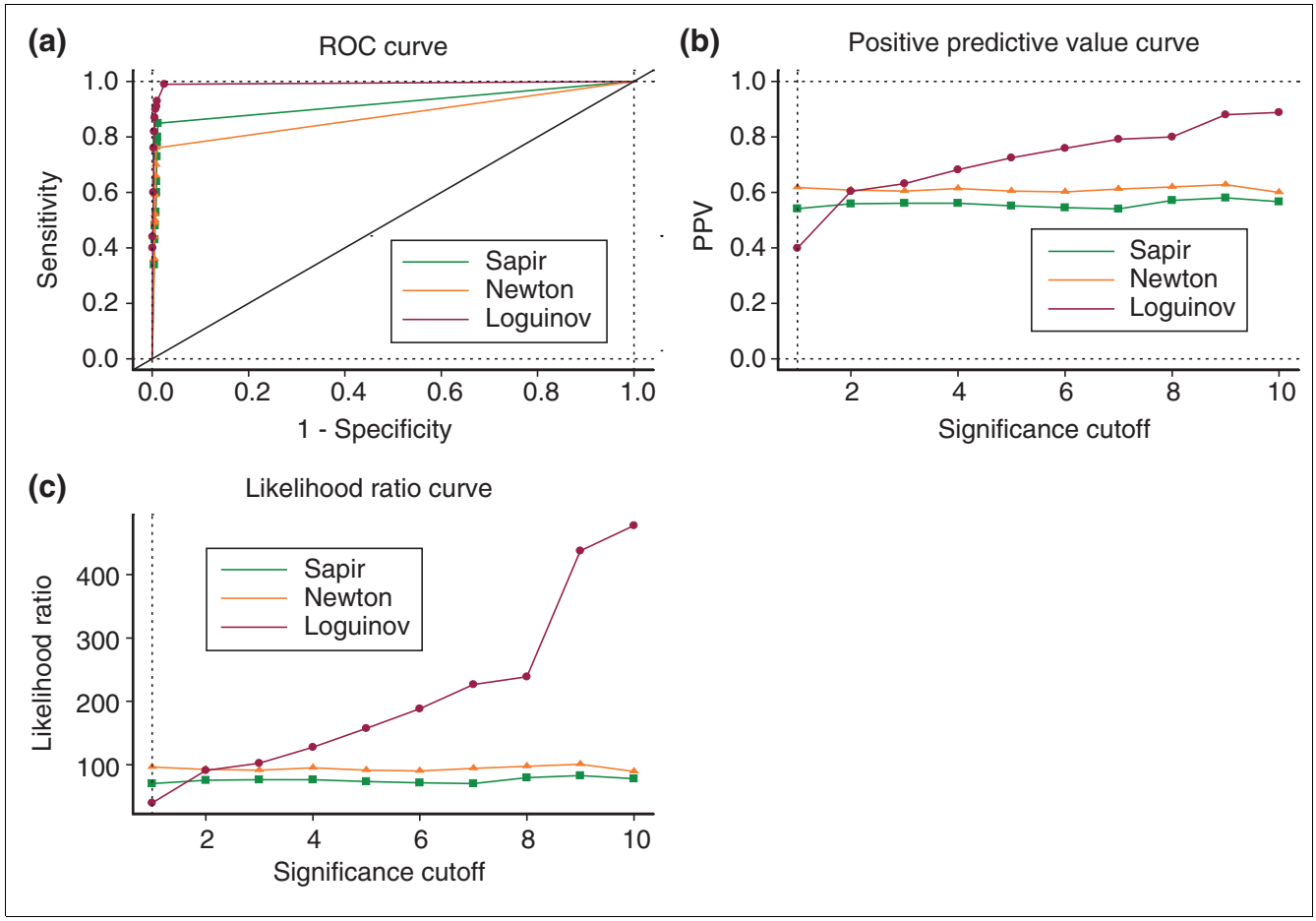

Figure 23
Comparison of performance of three single-slide methods on simulated data. **(a)** ROC curves for simulated data for each method. Each point represents one of the ten cut-offs used in Table 4 for each method. Area under curve is one in the case of an ideal method. ROC curves do not take prevalence into account and upper curves have better accuracy than lower ones (see text for details). **(b)** PPV curves which represent the probabilities that a gene identified as differentially expressed represents a true positive. **(c)** Likelihood ratio (or Bayes' factor) curves are calculated as Sensitivity/(1-Specificity) and can also be defined as the ratio of posterior odds to prior odds. PPV values takes into account prevalence of being differentially expressed in the simulated population.

If we have a replication (for example, we consider K points X_i with n_i replicates in each point, $i = 1,..., K$) then this information also may be incorporated while building SPIs [58,59]:

$$\pm\sqrt{\{t(N-2,1-\alpha/2k)\}^2[1/q+\frac{1}{\sum_{i=1}^{K}n_i}+\frac{(X_i-\bar{X})^2}{\sum_{i=1}^{K}n_i(X_i-\bar{X})^2}]s^2} \quad (12)$$

where: q is the number of future observations to be averaged at point X_i, n_i is the number of replicates at point X_i, residual variance

$$s^2=\sum_{i=1}^{K}\sum_{j=1}^{n_i}Y_{ij}^2/(\sum_{i=1}^{K}n_i-2), \quad \bar{X}=\sum_{i=1}^{K}n_iX_i/\sum_{i=1}^{K}n_i,$$

and the total number of observations is

$$N=\sum_{i=1}^{K}n_i.$$

If both $n_i \equiv 1$ and $q \equiv 1$ then Equation 12 coincides with Equation 10 ($K \equiv N$). Equation 12 can be simplified if there is the same number of replicates in each point $n_i \equiv M$ and $q \equiv M$ (that is, we'll predict mean values of future M points for each X_i):

$$\pm\sqrt{\{t(N-2,1-\alpha/2k)\}^2[\{1+\frac{1}{K}+\frac{(X_i-\bar{X})^2}{\sum_{i=1}^{K}(X_i-\bar{X})^2}]s^2/M} \quad (13)$$

If N is large and using approximation similar to Equation 11:

$$\pm\sqrt{\{t(N-2,1-\alpha/2k)^2s^2/M}=\pm t(N-2,1-\alpha/2k)(s/\sqrt{M}) \quad (14)$$

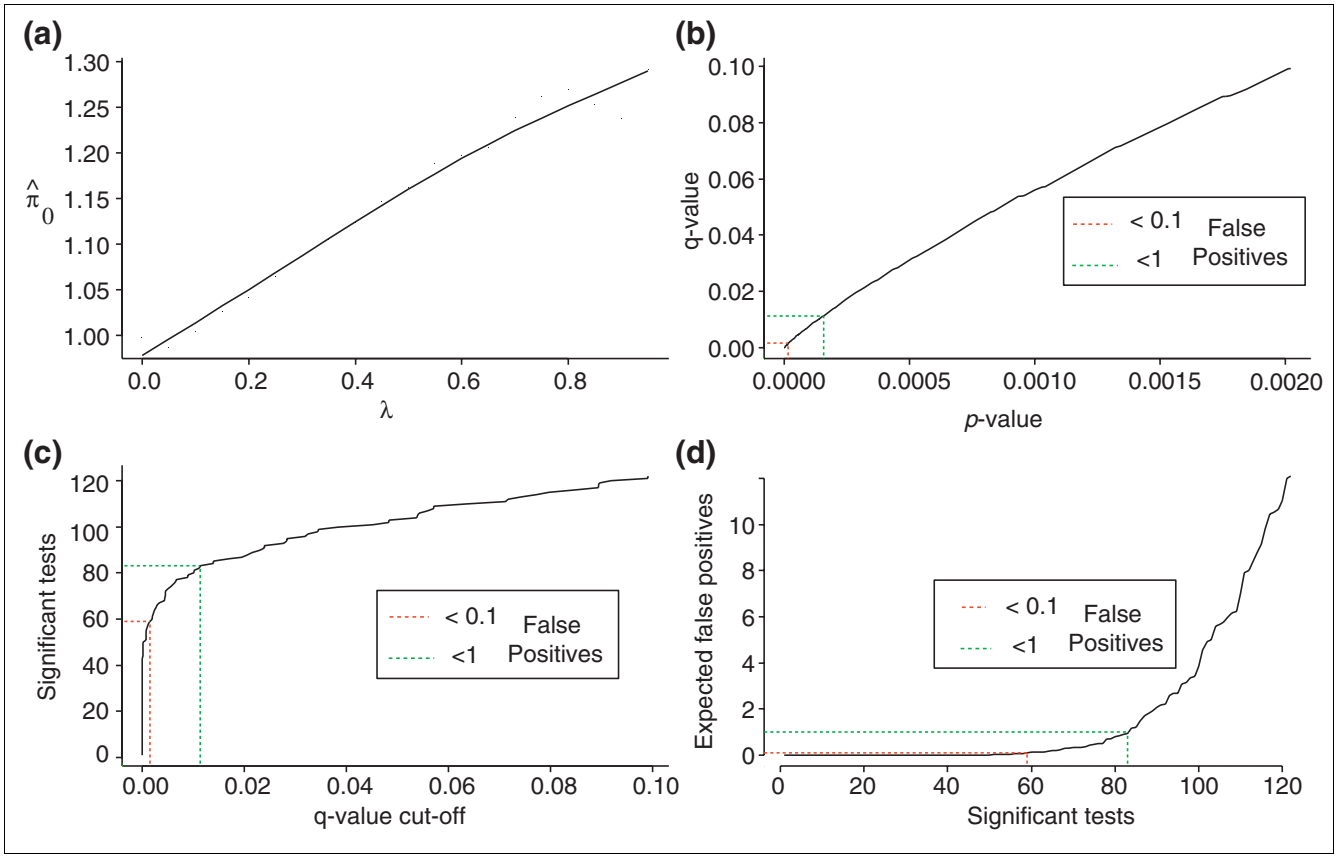

Figure 24

Interpreting *q*-values and calibrating *q*-value cut-offs. Four plots to facilitate *q*-value interpretation and calibrate the *q*-value cut-off [45,46] using the function *qplot()*. **(a)** The estimated portion of the true null hypotheses (π_0) versus the tuning parameter λ ('bootstrap' method is used for automatically choosing λ by the software and π_0 estimate is 0.978). **(b)** The expected proportion of false positives (*q*-value) for different *p*-value cut-offs. **(c)** The number of significant candidates for differential expression for each *q*-value. **(d)** The expected portion of false positives as a function of the number of candidates for differential expression called significant. The dotted black line in (a) is π_0 approximation using bootstrap method; the dotted color lines in (b) (green for expected false positives (FP) on average < 1 and red for < 0.1) are used to match *q*- and *p*-value levels (0.011 and 0.0016 for expected FT < 1, 0.0015 and 0.000015 for expected FT < 0.1, correspondingly) for the expected FP cut-offs; the dotted color lines in (c) are used to match *q*-value cut-offs (0.011 and 0.0015) and the number of significant tests on average (83 and 59); the dotted color lines in (d) are used to match expected FP cut-offs (< 1 and < 0.1) and the number of significant tests on average (83 and 59, correspondingly).

Therefore, for example, if we have spot duplicates on a slide and assume $M \equiv 2$ for each spot then the corresponding SPIs would be narrower than ones for the case without the duplication by a factor $1/\sqrt{2} \approx 0.71$ if predicting averages of future spot duplicates.

Smoothed STIs

STIs computed using Equations 2 and 5 (or SPIs using Equations 10 and 12) assume residual homoscedasticity: s^2 is independent of X_i and thus constant for all values $X_1,\ldots, X_N$. We refer to these STIs as ordinary STIs. For microarray data used, however, residual variance versus predictor variable plots usually provided evidence for heteroscedasticity: s^2 is dependent on X_i and thus not constant. To account for this relationship between X and s^2, we smoothed the data $(X_i, |Y_i|)$, $i = 1,\ldots, N$. The scatter plot smoothers *supsmu()* and *lowess()* perform locally linear (symmetric or containing the *k*-nearest neighbors) OLS-fit to point X_i [51,52,62,63]. The span parameter is the fraction of data points used for smoothing - larger values result in smoother fits. We used *supsmu()* with the span parameter *bass* = 2-4 and *lowess()* with *f* = 0.2-0.4. These values provide a compromise between sensitivity to local variation and smoothness (compare [8]).

The function *lowess()* computes a robust locally-weighted linear fit [63]. An extended version of the function, *loess()*, includes an option for locally-quadratic fitting [52]. A window dependent on *f* is positioned around X_i. Data points inside the window are weighted and a robust weighted regression is used to compute $\hat{Y}_i$, the predicted value of Y_i at X_i. No assumptions are made about the X_i values being evenly spaced and the span parameter *f* is constant across the entire range of the predictor variable *X*. A fixed span parameter, however, is problematic if the curvature of the underlying function varies; an increase in curvature would necessitate a decrease in the span, for example. The function *supsmu()*

avoids this difficulty by automatically choosing the variable span with cross-validation [62] of residuals in the neighborhood of X_i. The function *supsmu()* is faster than *lowess()* and whilst it is less robust for small sample sizes ($N < 40$), cDNA data are large (here $N = 6{,}068$) and so it is sufficiently robust. The choice between *supsmu()* and *lowess()* is a choice between more or less sensitivity to underlying curvature; *loess()* with locally-quadratic fitting also performs well recovering curvature in empirical data.

Adjusting the smoothed STIs

The scale estimate for the smoothed STIs, however, may be slightly different than the scale estimate for the ordinary STIs (Figure 19a) suggesting that smoothed STIs for real data require adjustments to improve their accuracy. Such adjusted STIs are determined by first generating simulated datasets assuming bivariate normality with the same parameters as in real data. Adjustments for STI scale estimator are calculated using the scale factor. The scale factor is defined as ratio in the simulated data of the Huber τ-estimate for scale to the average scale estimate (based on the *supsmu* (or *lowess*) scale estimator). The scale factor is then applied as an adjustment to smoothed STIs to derive adjusted smoothed STIs for the real *mac1* (or any other) data (Figure 19a,b). For example, in Figure 19a the scale factors to adjust *supsmu*- and *lowess*-based scale estimates are 1.25 and 1.35, respectively. Those estimates are very stable in repeat simulations: for example, their standard error based on ten repeat simulations for *mac1* dataset was 0.0008.

Computing *p*- and *q*-values

For every gene, we can compute the value of $\log_2(Cy5/Cy3)/s(I)$ for residuals, where $s(I)$ is a robust scale estimator that depends on intensity level. Then one can calculate *p*-values as a measure of statistical significance for every gene, using any appropriate formulas mentioned above if N large. An easy and accurate way to do it is the use, for example, formula 11. Then, having a list of *p*-values, one can calculate the corresponding *q*-values using R software developed by John Storey [46].

Calibrating *q*-values

Calibrating plots help to choose a cut-off for *q*-values (Figure 24). For example, if we select the number of false positives less than one, on average, then the number of significant tests = 83, *q*-value cut-off = 0.011 that corresponds to *p*-value cut-off = 0.0016.

Simulating differential expression

We took sample parameters and the 100 most prominent residual outliers from *mac1* dataset. We then considered the non-regular log-transformed ratios ($k = 100$) as location estimates, adding random component to simulate intensity-dependent variation. Heteroscedasticity for the main body of data (N-k = 6,068-100 = 5,968 regular observations) was also simulated using the same intensity dependence. Specifically, we define that heteroscedasticity takes place for $\log_2(Cy3)$ intensity values ≤ -3 or ≥ +2.

For that intensity range, the scale estimator was considered as a linear function of $\log_2(Cy3)$. For example, for regular observations:

$$simulated\ \log_2(Cy5/Cy3) = initial\ \log_2(Cy5/Cy3) + e(I) \quad (15)$$

where $e(I)$ is intensity-dependent random variation generated using *rnorm()* function, $e(I) = rnorm(k, mean{=}0, sigma{=}aI{+}b)$, where $I{=}\log_2(Cy3)$.

We define $a = -0.5$, $b = -1.5$ for the lower intensity area ($\log_2(Cy3) < -3$) and $a = 0.5$, $b = -1$ for the upper intensity area ($\log_2(Cy3) > +2$).

Heteroscedasticity for outliers was simulated in a similar way. R code and datasets used for simulations are available from the authors upon request.

Software implementation

All data processing, analysis and visualization were performed using S-plus 2000® [64]. Routines were written that used standard S-plus procedures and functions wherever possible and which generated HTML output. The R version of the code (DIGEX.R) is available from [65]. R [66] is a language and environment for statistical computing and graphics similar to S-plus. The R equivalents of S-plus functions are given in the text.

Additional data files

A comparative summary table (in html format; Additional data file 1) coupled with an auxiliary legend file (Additional data file 2) shows candidates for differential gene expression in all ten experiments used in the paper. An Excel table (Additional data file 3) gives the test accuracy definitions that were used for simulations to evaluate method performance.

Acknowledgements

We would like to acknowledge Ruben Zamar, Victor Yohai and Ricardo Maronna for critical reading of the manuscript. Our thanks go to Rus Yukhananov as well for helpful comments. We thank Cathy White for technical assistance. Work by Alex Loguinov and Chris Vulpe was supported by the Life Sciences Informatics Program of the University of California and the International Copper Association. Work by I.S.M. was supported the Director, Office of Energy Research, Office of Health and Environmental Research, Division of the US Department of Energy under Contract number DE-AC03-76F00098.

References

1. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M, Johnston R: **An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression.** *Nucleic Acids Res* 2001, **29:**E41.

2. Lee ML, Kuo FC, Whitmore GA, Sklar J: **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations.** *Proc Natl Acad Sci USA* 2000, **97:**9834-9839.
3. Kerr MK, Churchill GA: **Experimental design for gene expression microarrays.** *Biostatistics* 2001, **2:**183-201.
4. Kerr MK, Churchill GA: **Statistical design and the analysis of gene expression microarray data.** *Genet Res* 2001, **77:**123-128.
5. Kerr MK, Martin M, Churchill GA: **Analysis of variances for gene expression microarray.** *J Comput Biol* 2000, **7:**819-837.
6. Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM: **Microarray expression profiling identifies genes with altered expression in HDL-deficient mice.** *Genome Res* 2000, **10:**2022-2029.
7. Tsodikov A, Szabo A, Jones D: **Adjustments and measures of differential expression for microarray data.** *Bioinformatics* 2002, **18:**251-260.
8. Dudoit S, Yang LH, Callow MJ, Speed TP: **Statistical methods for identifying differentially expressed genes in restricted cDNA microarray experiments.** *Statistica Sinica* 2002, **12:**111-139.
9. Pan W, Lin J, Le CT: **How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach.** *Genome Biol* 2002, **3:**research0022.1-research0022.10.
10. Kohane IS, Kho AT, Butte AJ: *Microarrays for integrative genomics* Cambridge: The MIT Press; 2002.
11. Yang IV, Chen E, Hassenman J, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, *et al.*: **Within the fold: assessing differential expression measures and reproducibility in microarray assays.** *Genome Biol* 2002, **3:**research0062.1-0062.12.
12. Baggerly KA, Coombes KR, Hess KR, Stivers DN, Abruzzo LV, Zhang W: **Identifying differentially expressed genes in cDNA microarray experiments.** *J Comput Biology* 2001, **8:**639-659.
13. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278:**680-686.
14. Gross C, Kelleher M, Iyer VR, Brown PO, Winge DR: **Identification of the copper regulon in *Saccharomyces cerevisiae* by DNA microarrays.** *J Biol Chem* 2000, **275:**32310-32316.
15. Davies L, Gather U: **The identification of multiple outliers.** *J Amer Statist Assoc* 1993, **88:**782-792.
16. Gather U, Becker C: **Outlier identification and robust methods.** In *Handbook of statistics* Edited by: Maddala GS, Rao CR. Amsterdam: Elsevier Sciences; 1997:123-143.
17. Chen Y, Dougherty ER, Bittner ML: **Ratio-based decisions and the quantitative analysis of cDNA microarray images.** *J Biomed Optics* 1997, **2:**364-374.
18. Hughes TR, Marton Mj, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, *et al.*: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.
19. Newton MA, Kendziorsky CM, Richmond CS, Blattner FR, Tsui KW: **On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.** *J Comput Biol* 2001, **8:**37-52.
20. Sapir M, Churchill G: **Estimating the posterior probability of differential gene expression from microarray data.** *Poster* Jackson Laboratory, Bar Harbor, ME; 2000.
21. Rocke DM, Durbin B: **A model for measurement error for gene expression arrays.** *J Comput Biol* 2001, **8:**557-569.
22. Durbin B, Hardin J, Hawkins D, Rocke DM: **A variance-stabilizing transformation for gene-expression microarray data.** *Bioinformatics* 2002, **18 Suppl 1:**S105-S110.
23. Huber W, von Heydebreck A, Sultman H, Potuska A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18 Suppl11:**S96-S104.
24. Draghici S: **Statistical intelligence: effective analysis of high-density microarray data.** *Drug Discov Today* 2002, **7 (11 Suppl):**S55-S63.
25. Carroll RJ, Ruppert D: *Transformation and weighting in regression* New York: Chapman and Hall; 1988.
26. Cook RD, Weisberg S: *Residuals and influence in regression* New York: Chapman and Hall; 1982.
27. Hawkins DM: *Identification of outliers* New York: Chapman and Hall; 1980.
28. Barnett V, Lewis T: *Outliers in statistical data* New York: Wiley; 1994.
29. Lieberman GJ, Miller RG: **Simultaneous tolerance intervals in regression.** *Biometrika* 1963, **50:**155-168.
30. Miller RG Jr: *Simultaneous Statistical Inference* New York: Springer; 1981.
31. Rousseeuw PJ, Leroy AM: *Robust regression and outlier detection* Wiley series in probability and mathematical statistics. New York: Wiley; 1987.
32. Altman DG: *Practical statistics for medical research* Boca Raton: Chapman & Hall/CRC; 1999.
33. Jungmann J, Reins HA, Lee J, Romeo A, Hassett R, Kosman D, Jentsch S: **MAC1, a nuclear regulatory protein related to Cu-dependent transcription factors is involved in CU/Fe utilization and stress resistance in yeast.** *EMBO J* 1993, **12:**5051-5056.
34. Labbe S, Zhu Z, Thiele DJ: **Copper-specific transcriptional repression of yeast genes encoding critical components in the copper transport pathway.** *J Biol Chem* 1997, **272:**15951-15958.
35. Yamaguchi-Iwai Y, Serpe M, Haile D, Yang W, Kosman DJ, Klausner RD, Dancis A: **Homeostatic regulation of copper uptake in yeast via direct binding of MAC1 protein to upstream regulatory sequences of FRE1 and CTRL.** *J Biol Chem* 1997, **272:**17711-17718.
36. Zhu Z, Labbé S, Peña MM, Thiele DJ: **Copper differentially regulates the activity and degradation of yeast Mac1 transcription factor.** *J Biol Chem* 1998, **273:**1277-1280.
37. Dancis A: **Genetic analysis of iron uptake in the yeast *Saccharomyces cerevisiae*.** *J Pediatr* 1998, **132:**S24-S29.
38. De Freitas J, Wintz H, Kim JH, Poynton H, Fox T, Vulpe C: **Yeast, a model organism for iron and copper metabolism studies.** *Biometals* 2003, **16:**185-197.
39. Rutherford JC, Jaron S, Winge DR: **Aft1p and Aft2p mediate iron-responsive gene expression in yeast through related promoter elements.** *J Biol Chem* 2003, **278:**27636-27643.
40. Gross C, Kelleher M, Iyer VR, Brown PO, Winge DR: **Identification of the copper regulon in *Saccharomyces cerevisiae* by DNA microarrays.** *J Biol Chem* 2000, **275:**32310-32316.
41. Wiens BL: **When log-normal and gamma models give different results: a case study.** *The Am Stat* 1999, **53:**89-93.
42. Zamar R: **Robust estimation in the error in variables model.** *Biometrika* 1989, **76:**149-60.
43. Maronna RA, Yohai VJ: **Robust estimation of multivariate location and scale.** In *Encyclopedia of Statistical Sciences* Edited by: Kotz S, Read C, Banks D. New York: Wiley; 1998:589-596.
44. Cui X, Kerr MK, Churchill G: **Data transformations for cDNA microarray data.** *Technical Report* Jackson Laboratory, Bar Harbor, ME; 2002.
45. Storey JD: **A direct approach to false discovery rates.** *J R Statist Soc* 2002, **B64:**479-498.
46. Storey JD, Tibshirani R: **Statistical significance for genome-wide studies.** *Proc Natl Acad Sci USA* 2003, **100:**9440-99445.
47. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Statist Soc* 1995, **B57:**289-300.
48. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res* 2002, **30:**e15.
49. Mardia KV: **Tests of univariate and multivariate normality.** In *Handbook of Statistics* Edited by: Krishnaiah PR. North-Holland: Elsevier; 1980:279-320.
50. D'Agostino RB: **Tests for the normal distribution.** In *Goodness-of-fit techniques* Edited by: D'Agostino RB, Stephens MA. New York: Marcel Dekker; 1986:367-419.
51. Venable WN, Riply BD: *Modern applied statistics with S-plus* 3rd edition. New York: Springer; 1999.
52. MathSoft: *S-plus 2000: Guide to Statistics* Seattle-Washington: MathSoft; 1999.
53. Chambers JM, Cleveland WS, Kleiner B, Tukey PA: *Graphical Methods for Data Analysis* Belmont, California: Wadsworth; 1983.
54. Goodall C: **Examining residuals.** In *Understanding Robust and Exploratory Data Analysis* Edited by: Hoaglin DC, Mosteller F, Tukey JW. New York: Wiley; 1983:211-246.
55. Ripley BD: *Spatial statistics* Wiley: New York; 1981.
56. Atkinson AC: *Plots, transformations, and regression* Oxford: Oxford University Press; 1985.
57. Oleksiak MF, Churchill GA, Crawford DL: **Variation in gene expression within and among natural populations.** *Nat Genet* 2002, **32:**261-266.
58. Draper NR, Smith H: *Applied regression analysis* New York: Wiley; 1998.

59. Brownlee KA: *Statistical theory and methodology in science and engineering* New York: Wiley; 1965.
60. Guttman I: *Statistical tolerance regions: classical and Bayesian* London: Griffin; 1970.
61. Loguinov AV, Anderson LM, Crosby GJ, Yukhananov RY: **Gene expression following acute morphine administration.** *Physiol Genomics* 2001, **6:**169-181.
62. Friedman JH: **A variable span smoother. Laboratory for computational statistics. Technical Report 5.** Department of Statistics, Stanford: Stanford University; 1984.
63. Cleveland WS: **Robust locally weighted regression and smoothing scatterplots.** *J Amer Statist Assoc* 1979, **74:**829-836.
64. **Insightful** [http://www.insightful.com]
65. **DIGEX.R** [http://nature.berkeley.edu/~loguinov]
66. **The R Project for Statistical Computing** [http://www.r-project.org/]
67. Yamaguchi-Iwai Y, Stearman R, Dancis A, Klausner RD: **Iron-regulated DNA binding by the AFT1 protein controls the iron regulon in yeast.** *EMBO J* 1996, **15:**3377-3384.
68. De Freitas JM, Kim JH, Poynton HC, Su T, Wintz H, Fox TC, Holman PS, Loguinov AV, Keles S, Van Der Laan M, Vulpe C: **Exploratory and confirmatory gene expression profiling of mac1.** *J Biol Chem* 2004, **279:**4450-4458.
69. Rutherford JC, Jaron S, Winge DR: **Aft1p and Aft2p mediate iron-responsive gene expression in yeast through related promoter elements.** *J Biol Chem* 2003, **278:**27636-27643.
70. Yun C-W, Ferea T, Rashford J, Ardon O, Brown PO, Botstein D, Kaplan J, Philpott CC: **Desferrioxamine-mediated iron uptake in *Saccharomyces cerevisiae:* evidence for two pathways of iron uptake.** *J Biol Chem* 2000, **275:**10709-10715.
71. Spizzo T, Byersdorfer C, Duesterhoeft S, Eide D: **The yeast FET5 gene encodes a FET3-related multicopper oxidase implicated in iron transport.** *Mol Gen Genet* 1997, **256:**547-556.
72. Gross C, Kelleher M, Iyer VR, Brown PO, Winge DR: **Identification of the copper regulon in *Saccharomyces cerevisiae* by DNA microarrays.** *J Biol Chem* 2000, **275:**32310-32316.
73. Foury F, Roganti T: **Deletion of the mitochondrial carrier genes MRS3 and MRS4 suppresses mitochondrial iron accumulation in a yeast frataxin-deficient strain.** *J Biol Chem* 2002, **277:**24475-24483.
74. Martins LJ, Jensen LT, Simon JR, Keller GL, Winge DR, Simons JR: **Metalloregulation of FRE1 and FRE2 homologs in *Saccharomyces cerevisiae*.** *J Biol Chem* 1998, **273:**23716-23721.
75. Portnoy ME, Liu XF, Culotta VC: ***Saccharomyces cerevisiae* expresses three functionally distinct homologues of the nramp family of metal transporters.** *Mol Cell Biol* 2000, **20:**7893-7902.
76. Foury F, Talibi D: **Mitochondrial control of iron homeostasis. A genome wide analysis of gene expression in a yeast frataxin-deficient strain.** *J Biol Chem* 2001, **276:**7762-7768.
77. Protchenko O, Philpott CC: **Regulation of intracellular heme levels by HMX1, a homologue of heme oxygenase, in *Saccharomyces cerevisiae*.** *J Biol Chem* 2003, **278:**36582-36587.
78. Lin S-J, Pufahl RA, Dancis A, O'Halloran TV, Culotta VC: **role for the *Saccharomyces cerevisiae* A ATX1 gene in copper trafficking and iron transport.** *J Biol Chem* 1997, **272:**9215-9220.
79. Protchenko O, Ferea T, Rashford J, Tiedeman J, Brown PO, Botstein D, Philpott CC: **Three cell wall mannoproteins facilitate the uptake of iron in *Saccharomyces cerevisiae*.** *J Biol Chem* 2001, **276:**49244-49250.