Title: OpWise: Operons aid the identification of differentially expressed genes in bacterial microarray experiments

Authors: Morgan N. Price, Adam P. Arkin, and Eric J. Alm

Author affiliation: Lawrence Berkeley Lab, Berkeley CA, USA. A.P.A. is also affiliated with the Howard Hughes Medical Institute (Berkeley, CA) and the UC Berkeley Dept. of Bioengineering. All authors are also affiliated with the Virtual Institute for Microbial Stress and Survival.

Corresponding author: Eric J. Alm, ejalm@lbl.gov, phone 510-486-6899, fax 510-486-6219, address Lawrence Berkeley National Lab, 1 Cyclotron Road, Mailstop 977-152, Berkeley, CA 94720

Abstract

Motivation: Differentially expressed genes are typically identified by analyzing the variation between replicate measurements. These procedures implicitly assume that there are no systematic errors in the data, but several sources of systematic error are known. To estimate the amount of systematic error in bacterial microarray data, we assume that genes in the same operon have matching expression patterns.

Results: We describe "OpWise," an empirical Bayes analysis of a linear model that uses this assumption to estimate significance. In simulations, OpWise corrects for systematic error and is robust to deviations from its assumptions. In several bacterial data sets, significant amounts of systematic error are present, and replicate-based approaches overstate the confidence of the changers dramatically, while OpWise does not. Finally, OpWise assigns genes higher confidence if they are consistent with other genes in the same operon. This allows more changers can be identified at any given level of significance.

Availability: OpWise is available at http://gtlweb1.lbl.gov/OpWise, including source code in R and data sets analyzed in this paper.

Contact: ejalm@lbl.gov

Introduction

Microarray measurements of gene expression have become a popular tool for studying bacterial physiology, and hundreds of such studies are being conducted each year. Generally, these studies compare a treatment, either environmental or genetic, to a control condition. After obtaining raw hybridization intensities by scanning the slides or chips, the next steps are to normalize the data to remove experimental artefacts and then to identify differentially expressed genes. The resulting list of differentially expressed genes is typically the input to more biologically motivated analyses, such as a detailed examination of the differences

in expression patterns between conditions. The genes in this list are also candidates for confirmatory experiments to verify the differential expression or to test the role of individual genes in the stress response.

Because microarray experiments test many genes at once, and because the measurements contain significant noise, the question of how to define statistical significance is complex. At one extreme, one can argue that the ranking of genes within the list, with the most significant changers at the beginning of the list, is what matters, and that the length of the list can be determined by practical considerations, such as how many genes the analyst has the patience to examine or the amount of resources that are available for confirmatory experiments (Lonnstedt and Speed 2001; Smyth *et al.* 2003). Although we feel that this view contains some truth, we also note that one would like an estimate of the list's reliability. Furthermore, biologists often make statements describing the number of differentially expressed genes in an experiment. Because the length of these lists reflects arbitrary statistical cutoffs and also technical issues such as the amount of noise in the measurements, such statements should be avoided. Nevertheless, they reflect a reasonable desire to know how many changers have been reliably identified.

Another important reason to assign significance quantitatively, instead of merely ranking genes, is that it allows the analyst to test specific hypotheses. Given prior knowledge about the stress, one might surmise that a specific gene or pathway should be up- or down-regulated. A ranked list of genes may not test this hypothesis in a meaningful way. Ideally, the analyst would be given a confidence interval for the fold-change of each gene in the pathway.

To assess the reliability of the microarray measurements and to distinguish significant changers from other genes, statisticians have analyzed the variation between replicate experiments (Kerr *et al.* 2000; Ideker *et al.* 2000; Baldi and Long 2001; Tusher *et al.* 2001; Lonnstedt and Speed 2001; Dudoit *et al.* 2002; Storey and Tibshirani 2003; Smyth 2004). Implicitly, assessing significance by testing replication error assumes that replication captures all of the error in the data, and that there are no systematic biases. However, systematic errors have been observed due to many factors, including cross-hybridization, non-specific hybridization, dye incorporation bias, intensity-dependent effects, and spatial artefacts (Kerr *et al.* 2000; Jin *et al.* 2001; Kuo *et al.* 2002; Yang *et al.* 2002). Although normalization methods attempt to correct for these, normalization may not be entirely successful. More importantly, most normalization methods do not attempt to correct for all of these sources of error (e.g., most methods do not correct for cross-hybridization or for non-specific hybridization). If significant amounts of systematic bias remain and are not accounted for then the reported significance will be overstated. To determine if systematic errors are present, additional information besides the replicates is required.

For bacterial microarray experiments, we use operons to assess the amount of systematic

error in the data. Bacterial genes are often co-transcribed in multi-gene operons, and genes in the same operon should, in principle, have the same expression pattern. Although genes in the same operon are often expressed at different levels due to the varying stability of different segments of the mRNA, in steady-state situations, this will not affect the ratio in expression levels between conditions. Because most mRNA half-lives are short (under 10 minutes: Bernstein *et al.* 2002; Selinger *et al.* 2003), the steady state approximation should generally hold, and expression ratios should be consistent across an operon. Another reason why expression patterns can vary within an operon is that some operons have internal promoters or differential regulation of mRNA stability that can lead to differences in expression patterns (Adhya 2003). In practice, however, genes known to be in the same operon usually have very similar expression patterns, and expression patterns can be used to predict operons (Sabatti *et al.* 2002).

We assume that genes in the same operon have identical expression patterns, and infer that differences between the expression patterns of genes in the same operon are due to errors, which may be systematic or not. This assumption is somewhat conservative, because any true differences in expression patterns between genes in the same operon will be mistaken for errors, leading to overestimation of the amount of systematic error and conservative assessments of significance. In practice, however, this effect appears to be slight. Because the operon structure of most genes has not been experimentally determined, we rely on operon predictions, which are available for all prokaryotes, along with estimates of their reliability (Price *et al.* 2005; Ermolaeva *et al.* 2001; Moreno-Hagelsieb and Collado-Vides 2002).

Given this assumption about operons, we wish to estimate the amount of systematic bias in the data. One simple test is to ask how often two genes that are in the same operon have the same direction of change. However, even if one of the genes is a confident changer, and even if the operon prediction is highly confident, the measurement for the other gene in the operon may be noisy. In this case, the second gene will often report a change in the opposite direction from the first gene because of variation between the replicate measurements, and not because of systematic bias. Thus, interpreting the external information from operons requires us to have a model of the replication error.

We extend linear models for microarray data with replicates (Baldi and Long 2001; Lonnstedt and Speed 2001; Smyth 2004) to include systematic errors, and present an empirical Bayes analysis of the overall amount of systematic error and of the significance of each gene. Because we have observed that even low-confidence changers show a significant amount of agreement with operons, we do not assume that a minority of genes are changers and that the rest of the genes do not change (Lonnstedt and Speed 2001; Smyth 2004). Instead, we will assume that all genes are changing, even if, for most of them, the magnitude of change is small and the direction of change cannot be determined with confidence. Consequently,

rather than trying to distinguish the changers from the rest of the genes, we estimate for each gene the posterior distribution for the gene's fold-change given the data and the model. This can be summarized by a confidence interval or by the posterior probability that the gene's expression level went up (or down) in response to the treatment. As an example of a more stringent test, the method also can report the probability that a particular gene increased by 1.5-fold or more.

To test our method, we conducted simulations and also analyzed several experimental data sets. In simulations, the method correctly estimates the amount of systematic bias in the data and gives reasonable p-values even when some of the assumptions of the method are violated. On real data, we tested the agreement with operons of genes having varying levels of significance. For both two-color cDNA data and Affymetrix oligonucleotide data, our method finds significant amounts of systematic error, and gives p-values that are plausible, with a gradual reduction in agreement with operons as significance decreases. In contrast, approaches based on replication error, including non-parametric approaches (Tusher *et al.* 2001; Dudoit *et al.* 2002; Storey and Tibshirani 2003), often show low agreement with operons for confident changers (genes with $> 99\%$ probability of being true changers). Thus, replication-based approaches that ignore systematic bias are dramatically overstating significance.

We can also take advantage of operon structure to identify more changers. Intuitively, if two or three genes in the same operon all change in the same direction then they are unlikely to be false positives, but a changer that disagrees with the other genes in the same operon is suspect. Such reasoning is often used by biologists when examining microarray data. We derive a statistically sound "operon-wise" p-value, and show that these operon-wise p-values allow the identification of more changers at any specified level of significance than do single-gene p-values.

Methods

We present "OpWise," an empirical Bayes method for estimating the significance of the changes reported for each gene. The key elements of the method are (i) a linear error model that includes systematic errors, (ii) an approach for estimating the parameters of the error model (the hyperparameters), and in particular, a method for inferring the amount of systematic error from the agreement within operons, (iii) a mathematical solution for the posterior distribution of a gene's change in expression given the data for the gene and the parametrized error model, and (iv) an extension to the method to take other genes in the same operon into account when estimating the significance of each gene.

To describe the expression of each gene, we use normalized expression ratios, as these should be consistent within each operon. In practice, we use log-ratios (base 2) rather than raw ratios because the log-ratios have a better fit to the normal distribution. Instead of assuming that only a small fraction of genes are changing, we assume that every gene is changing (but only a small fraction of them might be measured with high confidence). Furthermore, we assume that there is some unknown amount of systematic error in the measurement for each gene, so that errors will remain no matter the number of replicates. Then, given the data for a gene i, we estimate the posterior distribution for the true log-ratio μ_i. This distribution can be summarized with a confidence interval or with the probability $P(\mu_i > 0)$ that a gene's expression level went up in the treatment condition. In contrast to p-values from testing a null hypothesis, where values near zero indicate strong significance, this probability will be near zero for highly confident down-changers and near one for highly confident up-changers. For genes whose direction of change cannot be confidently assessed, $P(\mu_i > 0)$ will be near 0.5. Because $P(\mu_i > 0)$ is a posterior p-value and does not test a null hypothesis, it is not affected by the number of genes being analyzed.

A Linear Model with Systematic Errors

First consider a simple experimental design with direct comparisons, where the samples from the conditions being compared are hybridized to the same chip. Each gene i has an unknown true response μ_i, systematic error ϵ_i, and variance between replicates σ_i^2. The measurements $\vec{x_i}$ for gene i is assumed to be normally distributed around $\mu_i + \epsilon_i$, and can be summarized by the observed mean $m_i = \sum_j x_{ij}/n_i$, where n_i is the number of measurements for gene i, and the total squared deviance $s_i^2 = \sum_j (x_{ij} - m_i)^2$, so that the likelihood of the data for each gene i is given by

$$f(\vec{x_i}) = \prod_{j=1}^{n_i} f(x_{ij}|\mu_i, \sigma_i, \epsilon_i) \propto \sigma_i^{-n_i} \exp(-\frac{\sum_j (x_{ij} - \mu_i - \epsilon_i)^2}{2\sigma_i^2}) = \sigma_i^{-n_i} \exp(-\frac{n_i(\mu_i + \epsilon_i - m_i)^2 + s_i^2}{2\sigma_i^2}) \quad \text{(Eq. 1)}$$

Another popular experimental design is to compare two types of samples separately to an external standard, such as genomic DNA or pooled mRNA samples. In these types of experiments, there are two sets of measured log levels for each gene, and the difference between them gives the log ratio. We refer to these log levels as $\vec{x_{1i}}$ and $\vec{x_{2i}}$, and summarize them with counts n_{1i} and n_{2i}, sample means m_{1i} and m_{2i}, and total squared deviances s_{1i}^2 and s_{2i}^2. We assume that the true variance in measurements $\vec{x_{1i}}$ and $\vec{x_{2i}}$ is identical, and that the unknown systematic bias ϵ_i affects the difference. We wish to estimate the distribution

of $\mu_i \equiv \mu_{1i} - \mu_{2i}$. Using the summary statistics

$$\begin{aligned} n_i &\equiv n_{1i} + n_{2i} - 1 \\ N_i &\equiv (n_{1i}^{-1} + n_{2i}^{-1})^{-1} \\ m_i &\equiv m_{1i} - m_{2i} \\ s_i^2 &\equiv s_{1i}^2 + s_{2i}^2 \end{aligned} \qquad \text{(Eq. 2)}$$

it is straightforward to show that

$$f(m_i, s_i^2 | \mu_i, \sigma_i, \epsilon_i) \propto \sigma_i^{-n_i} \exp(-\frac{N_i(\mu_i + \epsilon_i - m_i)^2 + s_i^2}{2\sigma_i^2}) \qquad \text{(Eq. 3)}$$

which is the same as the form for the direct comparison case except that N_i has replaced n_i in the exponential. Intuitively, the loss of a degree of freedom in n_i represents the fact that we do not care about the mean of x_{1i} or x_{2i} but only their difference, and the harmonic mean in N_i is used because in likelihood functions, precision (the inverse of variance) sums.

For both types of experiments, we follow Lonnstedt and Speed (2001) and Smyth (2004) and use an analytically tractable conjugate prior so that we can solve for the posterior distribution of μ_i given the observations and the estimates for the hyperparameters. (The hyperparameters control the distribution of gene-specific parameters such as μ_i and σ_i.) Specifically, we assume that the distribution of σ_i is given by an inverse chi-squared or inverse gamma distribution:

$$\begin{aligned} \sigma_i^2 &\equiv 1/\theta_i \\ \theta_i/\alpha &\sim \chi^2(\nu) \\ f(\theta_i) &= \frac{\theta_i^{\frac{\nu-1}{2}} e^{-\frac{\alpha\theta_i}{2}} (\frac{\alpha}{2})^{\frac{\nu+1}{2}}}{\Gamma(\frac{\nu+1}{2})} \end{aligned} \qquad \text{(Eq. 4)}$$

where α and ν are the hyperparameters. α is the scale of the chi-squared, and ν is the degrees of freedom.

We assume that the true mean μ_i and systematic error ϵ_i are normally distributed with variance proportionate to $1/\theta_i$:

$$\mu_i \sim N(0, \frac{1}{\theta_i \beta}) \qquad \text{(Eq. 5)}$$

$$\epsilon_i \sim N(0, \frac{1}{\theta_i \gamma}) \qquad \text{(Eq. 6)}$$

where β and γ are additional hyperparameters corresponding to the inverse of the amount of true variation and systematic error in the data, respectively.

We assume that the true means for the genes are independent, except that genes in the same operon have the same θ_i and μ_i (but independent bias ϵ_i). As discussed in the introduction, genes in the same operon are generally co-regulated, so μ_i should be similar. The assumption that θ_i is identical is required because in our model μ_i depends on θ_i; the empirical justifiability of this assumption will be discussed in the Results. Because operon predictions are only 80-90% accurate, we use a method that estimates the probability $P(Operon_{ij})$ that two adjacent genes are co-transcribed (Price *et al.* 2005), and treat the actual state of each potential operon pair as an unknown random variable. For example, the prediction method might estimate that two genes have a 90% probability of being in the same operon; in our model, we use this estimate as the true probability. We use only the likely operon pairs (those with $P(Operon_{ij}) \geq 0.5$).

Solving a Simplified Model without Systematic Errors

We first describe parameter estimation and significance testing in a simplified version of the above model that lacks systematic errors (that is, all $\epsilon_i = 0$ and $\gamma = \infty$). In this model, we need to estimate the prior distribution for θ_i (or σ_i^2), which is determined by the scale α and degrees of freedom ν, and the scale of variation for the true log-ratio μ_i given the variance σ_i^2, which is given by $1/\beta$.

Estimating the Hyperparameters

Although we assume that μ_i is normally distributed for all genes, instead of being allowed to vary for a minority of genes, the variation between replicates in our model is the same as in Smyth (2004). As discussed by Smyth (2004), the distribution of $\log s_i^2$ (the log of the squared deviances) is more normally distributed than that of s_i^2 and hence more suitable for fitting, and the relationship between the mean and variance of $\log s_i^2$ and the data is given by

$$e_i \equiv \log s_i^2 - \psi(\frac{n_i - 1}{2}) + \log(\frac{n_i - 1}{2})$$

$$\psi'(\frac{\nu+1}{2}) = \text{mean}\{(e_i - \bar{e})^2 \cdot \frac{N_{genes}}{N_{genes}-1} - \psi'(\frac{n_i-1}{2})\}$$
$$\frac{\alpha}{\nu+1} = \exp\{\bar{e} + \psi(\frac{\nu+1}{2}) - \log(\frac{\nu+1}{2})\} \quad \text{(Eq. 7)}$$

where $\psi()$ is the digamma function, $\psi'()$ is the trigamma function, $\bar{e}$ is the mean of the e_i, and the inverse of the trigamma requried to solve for ν can be obtained numerically by Newton iteration. This gives us estimates for α and ν, which describe the distribution of the true variances σ_i^2 for each gene (see Eq. 4).

We then find the maximum likelihood estimate of β, which describes the distribution of the true means μ_i^2 for each gene (see Eq. 5). The likelihood of the data is given by

$$\prod_i f(m_i, s_i^2) = \prod_i \int_0^\infty d\theta_i f(\theta_i) \int_{-\infty}^\infty d\mu_i f(\mu_i) f(m_i, s_i^2|\mu_i, \theta_i)$$
$$\propto \prod_i \sqrt{\frac{\beta}{\beta+N_i}} \cdot \left(\alpha + s_i^2 + m_i^2 \cdot \frac{N_i \cdot \beta}{\beta+N_i}\right)^{-\frac{\nu+n_i+1}{2}} \quad \text{(Eq. 8)}$$

where for direct comparison experiments, $N_i \equiv n_i$. We choose β to maximize the (logarithm of) this likelihood, using a Newton iteration method (*nlm* in the R statistics package: http://www.r-project.org/).

Significance of Individual Genes

Given estimates for the hyperparameters and the observed mean m_i and total squared deviance s_i^2 for a gene i, the posterior probability distribution for μ_i is given by

$$f(\mu_i|m_i, s_i^2) \propto \int_0^\infty f(\theta_i) f(\mu_i|\theta_i) f(m_i, s_i^2|\theta_i, \mu_i) d\theta_i \propto \left(\alpha + \beta\mu_i^2 + N_i(\mu_i - m_i)^2 + s_i^2\right)^{-\frac{\nu+n_i}{2}-1} \quad \text{(Eq. 9)}$$

which is a t distribution with $\nu + n_i + 1$ degrees of freedom, and

$$t = \frac{\mu_i - m_i'}{\sqrt{V_i}} \quad \text{(Eq. 10)}$$

where m_i' and V_i are "shrunk" estimates of the mean and of the uncertainty:

$$m_i' = m_i \frac{N_i}{\beta + N_i}$$

$$V_i = \frac{\alpha + s_i^2 + m_i^2 \frac{N_i \beta}{\beta + N_i}}{(\beta + N_i)(\nu + n_i + 1)} \qquad \text{(Eq. 11)}$$

Intuitively, m_i^2 appears in the estimate of the variance because m_i^2 contains information about the variance (in our model the expectation of μ_i^2 is σ_i^2/β). Given this posterior distribution for μ_i, the probability of any hypothesis can be estimated by using the standard t test. We use the probability $P(\mu_i > 0)$ as a measure of significance.

Accounting for Systematic Errors

The key advantage of our approach is to use biological knowledge (i.e., operon predictions) to take systematic errors into account. These systematic errors will not be eliminated by increasing the number of replicate measurements, but can be estimated by their effect on the agreement of changes in expression for genes in the same operons. In this section, we add systematic errors to the above model ($\epsilon_i > 0$, $\gamma < \infty$) and describe how to account for such bias.

Estimating the Parameters

If we ignore the distinction between systematic error ϵ_i and true variation μ_i, then we can replace μ_i with $\mu_i' \equiv \mu_i + \epsilon_i$. The distribution of μ_i' is given by

$$\mu_i' \sim N\left(0, \frac{1}{\theta_i \beta}\right) + N\left(0, \frac{1}{\theta_i \gamma}\right) = N\left(0, \frac{1}{\theta_i} \cdot (\frac{1}{\beta} + \frac{1}{\gamma})\right) = N\left(0, \frac{1}{\theta_i \beta'}\right) \qquad \text{(Eq. 12)}$$

where

$$\frac{1}{\beta'} \equiv \frac{1}{\beta} + \frac{1}{\gamma} \qquad \text{(Eq. 13)}$$

so that the distribution of m_i for a model with systematic errors is the same as that for a model without systematic errors, except that we replace β with β'. The distribution of s_i^2

is not affected by systematic errors. Thus, we can estimate α, ν and β' using the bias-free method described above.

We then find the maximum likelihood estimate of γ, which controls the amount of bias, from the operons, using our assumption that genes in the same operon will have the same values of μ_i and of $\theta_i = 1/\sigma_i^2$. The overall likelihood can be decomposed into terms for individual genes and pairwise terms for operon pairs:

$$f(\vec{x_1}...\vec{x_N}) = \prod_i f(\vec{x_i}) \prod_{ij} \frac{f(\vec{x_i}, \vec{x_j})}{f(\vec{x_i}) \cdot f(\vec{x_j})} \qquad \text{(Eq. 14)}$$

We have already taken into account the effect of γ on the single-gene likelihoods $f(\vec{x_i})$ by introducing β', which is now being held constant, so these terms do not need to be considered. The pairwise terms are given by

$$\prod_{ij} \frac{f(\vec{x_i}, \vec{x_j})}{f(\vec{x_i}) \cdot f(\vec{x_j})} = \prod_{ij} \left(1 - P(Operon_{ij}) + P(Operon_{ij}) \frac{f(\vec{x_i}, \vec{x_j} | Operon_{ij})}{f(\vec{x_i}) \cdot f(\vec{x_j})}\right) \qquad \text{(Eq. 15)}$$

This expression takes into account the possiblity of errors in the operon predictions. The inner term can be derived from

$$f(\vec{x_i}, \vec{x_j} | Operon_{ij}) = \int_0^\infty d\theta_{ij} f(\theta_{ij}) \int_{-\infty}^\infty d\mu_{ij} f(\mu_{ij}) \cdot f(m_i, s_i^2 | \mu_{ij}, \theta_{ij}) \cdot f(m_j, s_j^2 | \mu_{ij}, \theta_{ij})$$

$$f(\vec{x_i}) = \int_0^\infty d\theta_i f(\theta_i) \int_{-\infty}^\infty d\mu_i f(\mu_i) \cdot f(m_i, s_i^2 | \mu_i, \theta_i)$$

$$f(m_i, s_i^2 | \mu_i, \theta_i) = \int_{-\infty}^\infty d\epsilon_i f(\epsilon_i) \cdot f(m_i, s_i^2 | \mu_i, \theta_i, \epsilon_i) \qquad \text{(Eq. 16)}$$

giving

$$\frac{f(\vec{x_i}, \vec{x_j} | Operon_{ij})}{f(\vec{x_i}) \cdot f(\vec{x_j})} = \left(\frac{\alpha}{2}\right)^{-\frac{\nu+1}{2}} \sqrt{\frac{\beta}{\beta + N_i' + N_j'}} \cdot \frac{\sqrt{(\beta' + N_i)(\beta' + N_j)}}{\beta'} \cdot \frac{\Gamma(\frac{\nu+n_i+n_j+1}{2})\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu+n_i+1}{2})\Gamma(\frac{\nu+n_j+1}{2})}$$

$$\cdot \left(\frac{\alpha + s_i^2 + m_i^2 N_i \frac{\beta'}{\beta' + N_i}}{2} \right)^{\frac{\nu + n_i + 1}{2}} \cdot \left(\frac{\alpha + s_j^2 + m_j^2 N_j \frac{\beta'}{\beta' + N_j}}{2} \right)^{\frac{\nu + n_j + 1}{2}}$$

$$\cdot \left(\frac{\alpha + s_i^2 + s_j^2 + N_1' m_1^2 + N_2' m_2^2 - \frac{(m_1 N_1' + m_2 N_2')^2}{\beta + N_1' + N_2'}}{2} \right)^{-\frac{\nu + n_1 + n_2 + 1}{2}} \quad \text{(Eq. 17)}$$

with

$$N_i' \equiv (N_i^{-1} + \gamma^{-1})^{-1}$$
$$N_j' \equiv (N_j^{-1} + \gamma^{-1})^{-1} \quad \text{(Eq. 18)}$$

Besides the constant factors, this is a t distribution form for $f(\vec{x_i}, \vec{x_j} | Operon_{ij})$ divided by t distribution forms for $f(\vec{x_i})$ and $f(\vec{x_j})$. We use a Newton iteration method to find the value of γ that maximizes this product of the pairwise likelihood ratios.

Significance of Individual Genes

If we ignore the information from other genes, then the posterior distribution of μ_i is given by a t distribution with $\nu + n_i + 1$ degrees of freedom, where

$$t = \frac{\mu_i - m_i'}{\sqrt{V_i}}$$
$$m_i' = m_i \frac{N_i'}{\beta + N_i'}$$
$$V_i = \frac{\alpha + s_i^2 + m_i^2 \frac{N_i' \beta}{N_i' + \beta}}{(\beta + N_i')(\nu + n_i + 1)} \quad \text{(Eq. 19)}$$

which is the same as the formula for the case without systematic bias except that N_i has been replaced by N_i'.

Significance taking Operons into Account

Although the method as described so far uses operon predictions to estimate the hyper-parameters, it uses only the information for each gene when computing p-values. We will

refer to these as "single-gene" p-values. Here we use information from other genes in the same operon to improve our estimates of the significance of each gene, giving "operon-wise" p-values. As we will show in the Results, using this additional information often allows increased confidence in the measurements.

First, assume that we have two genes i and j that are known to be in the same operon, with the same (unknown) μ_{ij} and θ_{ij} but with differing biases ϵ_i, ϵ_j. Given measurements for the two genes, the posterior distribution for μ_{ij} is a t distribution with $\nu + n_i + n_j + 1$ degrees of freedom, and

$$t = \frac{\mu_{ij} - m'_{ij}}{\sqrt{V_{ij}}}$$

$$m'_{ij} = \frac{N'_i m_i + N'_j m_j}{\beta + N'_i + N'_j}$$

$$V_{ij} = \frac{\alpha + s_i^2 + s_j^2 + N'_i m_i^2 + N'_j m_j^2 - \frac{(N_i m_i + N_j m_j)^2}{\beta + N'_i + N'_j}}{(\nu + n_i + n_j + 1)(\beta + N'_i + N'_j)} \qquad \text{(Eq. 20)}$$

It is straightforward to extend this formula to three or more genes.

In practice, operon predictions are uncertain, and we need to take this uncertainty into account in estimating confidence. We use only the adjacent pairs that are predicted to be in the same operon (those with $P(Operon_{ij}) \geq 0.5$), as more distant pairs are less reliable. In the most complicated case, we have genes i and k on either side of our target gene j and four possible cases: singleton transcript j, two-gene operon ij, two-gene operon jk, or three-gene operon ijk. The posterior distribution of μ_j is then a mixture of the corresponding four posterior distributions. However, rather than using the input probabilities $P(Operon_{ij})$ and $P(Operon_{jk})$ we use the posterior operon probabilities given the data. That is, we use the microarray data to help estimate the likelihood that a pair of genes are co-transcribed. Using Bayes' law, these probabilities are given by

$$\frac{P(Operon_{ij}|\vec{x_i}, \vec{x_j})}{P(\neg Operon_{ij}|\vec{x_i}, \vec{x_j})} = \frac{P(Operon_{ij})}{P(\neg Operon_{ij})} \cdot \frac{f(\vec{x_i}, \vec{x_j}|Operon_{ij})}{f(\vec{x_i}) \cdot f(\vec{x_j})} \qquad \text{(Eq. 21)}$$

where the formula for the ratio on the right was given in Eq. 17. Using the posterior operon probabilities gives the rigorously correct posterior distribution for μ_j (derivation not shown). Using the posterior operon probabilities also prevents the method from giving a low posterior distribution to a gene that went up but is in an operon with genes that went down, because in this situation the posterior $P(Operon)$ will be low.

Results

We tested on our method on four data sets collected with a variety of measurement platforms and from several different bacteria. In brief, we first fit our model to each data set, and found statistically significant systematic bias in each. We used simulations to test how well our model fit each data set and whether our method was robust to modest deviations from the underlying assumptions. We then compared the significance estimates from our method to those from several other methods. Finally, we tested whether operon-wise significance values were more sensitive than single-gene values.

Data Sets

The four data sets we tested were:

dvSalt30 – *Desulfovibrio vulgaris* salt shock at 30 minutes (Z. He and J. Zhou, personal communication). This data was collecting using two-color glass slides with 70-mer probes. The experiment was an indirect comparison through a genomic control. There were three biological replications for each condition, measured with 1 slide each, and 2 spots per gene per slide, for a total of six replicate measurements for each gene and condition.

ecox – A comparison of aerobic and anaerobic log-phase growth in *Escherichia coli* (Covert *et al.* 2004). This data was from Affymetrix oligonucleotide chips with 3 or 4 replicate hybridizations for each of the two conditions and is available as GEO accession GDS680.

shCold5 – *Shewanella oneidensis* cold shock at 5 minutes (Z. He and J. Zhou, submitted). This data was a direct comparison of two-color glass slides using cDNA probes. There were 5 biological replicates with 1 slide each and 2 spots per gene per slide (10 measurements per gene total), but no dye swap (the same dyes were used for the control and treatment samples throughout).

shHeat5 – *Shewanella oneidensis* heat shock at 5 minutes (Gao *et al.* 2004). This data was also a direct comparison of two-color cDNA probes. There were three biological replicates, with two replicate slides each and 2 spots per gene per slide (12 total measurements per gene), and with dye swap (Cy3 dye was used for the treatment in half of the slides and for the control in the other half of the slides).

Normalization

For the two-color direct comparison data sets (shCold5 and shHeat5), we performed intensity-dependent and then spatial normalization on each slide. Specifically, we first used a locally smooth estimator to remove intensity-dependent effects and then subtracted the median from each sector, similar to recommendations of Dudoit *et al.* (2002). For the indirect comparison data set (dvSalt30), we treated the ratio of intensities between the channels corresponding to cDNA and to genomic DNA as a raw expression level. We first performed a global normalization for each slide so that the total expression level was the same for each slide, and then computed the average of the log-expression levels across slides from the two conditions. This gave us log-ratios between the two conditions to which we could apply the intensity-dependent and spatial normalization approaches. For all three of these data sets, we then merged the data from the two spots for each gene, considering them as two independent sets of replicates. There was little difference between within-slide and between-slide variance (data not shown). For the Affymetrix data set (ecox), the data we downloaded had already been normalized with dChip (Li and Wong 2001), so we used the normalized expression levels provided; to prevent small values of expression level from giving extreme outliers for log ratios, we added a small constant (5) to the expression levels before taking a logarithm.

Fit of Model to Data

To see how well the model fit the data, we inferred the hyperparameters for each data set, used these parameters to create simulated data, and compared the simulated data to the original data sets. We ran 50 simulations for each of the 4 original data sets. Each simulation had the same proportion of missing data as the corresponding data set. For operons, we randomly assigned adjacent genes on the same strand to be in the same operon or not with the probabilities given by the prediction method, but only if the probability was 0.5 or greater. With these "model" simulations, we were able to test our assumptions about the distribution of means and variances. To emulate the heavy tails in ecox (see below), we performed 50 "mixture" simulations where 10% of the genes had much higher variation in the mean (a much lower β) than the other genes. Finally, to test our assumptions that (i) the true mean and true variance are correlated and (ii) the true variance is correlated within each operon, for each data set we performed 50 "uncoupled" simulations where the mean was independent of variance (the mean was normal with a fixed width) and genes in the same operon had independent variances.

As shown in Figure 1, the inverse gamma distribution provided an excellent fit to the observed distribution of squared deviance s_i^2. Furthermore, the simulated distribution of observed

means had heavier tails than a normal distribution, due to the wide spread of deviances (compare "Model" means in Figure 1 to the "Uncoupled" means, which follow the normal distribution). The distribution of means fit the data fairly well for three of the data sets, but for the ecox data set, the true distribution had even heavier tails.

To test our assumptions that the variation in the true means depends on the true variances, we compared the correlations of observed means and squared deviances in the real data to both the coupled and uncoupled simulations. As shown in Table 1, the observed mean and squared deviance were much more correlated than in the uncoupled model, except in the shCold5 data set. Similarly, within each operon the squared deviances were significantly correlated. However, the correlations were generally weaker than in the simulations, indicating deviations from the assumptions.

Finally, our method identified large amounts of systematic bias, similar in magnitude to the true changes in gene levels and the replication error, in all four data sets (Table 2). Furthermore, the bias was statistically highly significant in all four data sets, as determined by a maximum likelihood ratio test (see Table 2). This confirms that the problem of systematic bias is real. In a later section, we will show that ignoring this bias can lead to large overstatements of the reliability of measurements for individual genes.

Robustness of Method in Simulations

Because the method uses operons to estimate the overall reliability of the measurements, we hypothesized that the method would be robust to the modest deviations from its assumptions, such as the heavy tails in the distribution of the means in the ecox data set or the weaker than expected correlation between the means and the variances. We also wanted to verify that the estimated hyperparameters were accurate enough to give reasonable p-values. To test these hypotheses, we examined the single-gene estimates of $P(\mu_i > 0)$ for the simulated data (μ_i is the true log-change for gene i). For the simulations that followed our model, we compared these p-values computed with estimated hyperparameters to "ideal" p-values computed with the true hyperparameters. For the uncoupled simulations, we compared the p-values to the actual sign of μ_i for each gene.

When comparing the log odds of the estimated p-values to the log odds of the ideal p-values, we consistently observed a strongly linear relationship, with correlation coefficients above 0.9999 (see Figure 2A; $\text{logodds(p)} \equiv \log \frac{\text{p}}{1-\text{p}}$). In other words, the ordering and shape of the significance values was not affected, but the overall scale of significance could be. To summarize this linear relationship between the two sets of significance estimates, we used the slope of the ideal log odds as a function of the estimated log odds. Slopes less than one indicate that the significance values with the estimated hyperparameters are

overly conservative, and slopes greater than one indicate that errors in the estimates of the hyperparameters led to overly aggressive significance values. As shown in Figure 2B, most simulations had slopes very close to 1.0. In a total of 200 simulations across 4 data sets, the most extreme aggressive slope was 1.12 (for shHeat5). This corresponds to reporting $P(\mu > 0) = 0.964$ for a gene with a true p-value of 0.95.

For the uncoupled and mixture simulations, which violated the assumptions of our model, we did not have ideal p-values to compare to, so we instead used logistic regression (glm in R, http://r-project.org) to determine the slope. Logistic regression identifies the multiplier for the estimated log odds that best fits the observed pattern of whether $\mu > 0$ or not – see Figure 2C. As shown in Figure 2D, the accuracy of the method was not dramatically affected by uncoupling the mean from the variance. However, the mixture simulation, which emulated the heavy tails in the ecox data set, produced slopes around 1.2, with a maximum of 1.58. This outlier simulation included extreme and biologically unrealistic outliers in true μ (a log-ratio of -44!) that led to a very high estimate of the true variance of the mean and to β' being an order of magnitude too low. Such outliers are not present in our genuine data sets and need to be removed before using our method. A slope of 1.2, which corresponds to reporting a p-value of 0.97 when the true p-value is 0.95, is not ideal, but as we will show, methods that do not account for systematic bias, including non-parametric methods, can perform dramatically worse.

For all simulations, we also compared the operon-wise p-values to either the ideal or true significance. These gave similar slopes as the single-gene p-values, but with consistently smaller deviations from 1.0 (data not shown).

Quality of Significance Estimates

To test the quality of the significance estimates on real data, we compared the confidence assigned by our method to the extent of agreement with operons. For each data set, we sorted genes by confidence into eight groups. Although our p-values are single-tailed – they test only the hypothesis that $\mu_i > 0$ – we wanted a two-tailed notion of confidence, because this is more comparable to other methods. We defined the two-tailed confidence as $C = 2 \cdot |p - 1/2|$. For each gene in each group, we identified other genes predicted to be in the same operon, and asked whether the two genes changed in the same direction. We used only adjacent genes, as operon predictions for more distant genes are less confident. Intuitively, if a group of genes are 99% confident, then 99% of the time, the measurement for that gene is correct, and it will always have the same sign as other genes in the operon; the other 1% of the time, there is no information about the gene, and the genes will have the same sign, by chance, 50% of the time. That is, $P(Agree) = C + (1 - C)/2 = (1 + C)/2$, or $2 \cdot P(Agree) - 1 = C$. We also needed to correct for the possibility that the operon prediction

is incorrect, which gives $2 \cdot P(Agree) - 1 = C \cdot P(Operon)$. Thus, we defined an adjusted measure of agreement, whose expectation ranges from 0 for data that is all noise to 1 for perfect data, as $Adjusted = (2 \cdot Agree - 1)/P(Operon)$, where $Agree$ is 1 if true and 0 if false. This measure corrects for variations in the confidence of operon predictions between groups of genes – in some data sets, the most confident changers were, on average, in more confidently predicted operons (data not shown). Finally, even if the measurement for the first gene in the operon is highly confident and correct, the measurement for the other gene in the operon may be noisy, and the two genes may not agree. As there is no simple way to correct for this, we used the simulations described above, and compared the relationship between confidence and agreement in the real data to that in the simulations. The relationship between confidence and adjusted agreement with operons was approximately linear in all data sets (Figure 3) and was largely consistent with simulations (the simulations according to our model are shown in Figure 3; the other simulations are similar and are shown in Supplementary Figure 1).

Furthermore, for most groups of genes, including those with modest confidence values, the adjusted agreement with operons was much larger than zero. This suggests that the expression levels of all genes in these experiments were in fact changing, even if many individual genes could not be measured with confidence. In all four data sets, when we grouped genes into eight sets with similar confidence values, the top six of eight confidence groups had significantly more operon pairs that agreed with microarray data than not (all $p < 0.05$, binomial test). This confirmed our assumption that all genes are changers.

Significance Estimates Are Unreasonable Unless Bias is Accounted For

Figure 3 also shows the relationship between confidence and operons for our model with the bias removed (using $\gamma = \infty$). Naturally, the confidence estimates from the model without bias were higher. In the shHeat5 and shCold5 data sets, the bias-free estimates of confidence were much too high: the highest and second-highest confidence groups both had confidence levels very near one, but the second-highest group had a much lower level of agreement with operons than the highest group. This also rules out one alternative explanation for why we detected significant bias in these data sets, which is that microarray data lacks bias but the operon predictions were flawed or systematically overconfident. In the latter case, the agreement with operons should have been lower for changers at every level of confidence, including the most confident changers. For dvSalt30, the bias-free confidence estimates appear to be more modestly over-confident, while for ecox, the difference between models with and without bias is small.

Finally, we compared the confidence estimates from our model to those from a popular non-parametric method, namely significance analysis of microarrays, or SAM (Tusher *et al.* 2001).

For each gene, SAM tests the null hypothesis that the gene's expression level is identical in the two conditions. To do this, SAM uses a modified t statistic with a pseudovariance term in the denominator. However, rather than using a t test, SAM estimates the null distribution for the modified t statistic by performing random permutations of the data. For a direct comparison experiment, this involves randomly flipping the signs of the replicates, while for an indirect comparison, it involves randomly reassigning replicates between conditions. Given p-values from each gene from these permutation tests, SAM then uses the proportion of genes with high p-values to estimate the proportion of genes that are non-changers, and hence the proportion of genes that are true changers (similar to Storey and Tibshirani (2003)). Finally, given this proportion, it corrects for multiple testing and estimates the false discovery rate (FDR). For each gene, the FDR is an estimate of the proportion of false positives among genes that are at that genes' significance level or more significant. To compare these significance values to the confidence levels from our method in Figure 3, we needed the proportion of false positives within each group, also known as the local false discovery rate – the confidence is 1 minus the local FDR. For the most significant group, the local FDR is simply the FDR for the least significant member of the group. For the less significant groups, the number of false positives can be estimated from the FDR by subtracting the false positives expected for the more significant groups (similar to Aubert *et al.* (2004)).

As shown in Figure 3, for the shHeat5 and shCold5 data sets, SAM is far too confident, and similar to the parametric model without bias. Indeed, for the shHeat5 data set, SAM assigned an FDR of under 10^{-4} to 78% of all genes! (This is why there are only two groups for SAM shown in the plot – most of the values were identical.) This behavior was far worse an overstatement of p-values that we ever observed in the simulations that violated our distributional assumptions (Figure 2D). For the dvSalt30 data set, which has a moderate amount of bias, SAM was also more confident than our model, at least for the more significant changers (the three right-most groups containing the top 1,300 genes). The SAM curve was also noticeably below the simulation curve, suggesting that it was (moderately) over-confident. Finally, for ecox, which has little bias and a heavy-tailed distribution, SAM performed well (see top right of curve), while our method was perhaps slightly over-confident. (The odd behavior of the left side of the SAM curve reflects noise inherent in our method of calculating the local FDR.) Overall, we concluded that the bias we inferred in these data sets was not somehow due to violations of our distributional assumptions, and that the bias can be a much larger source of over-confidence than violations of distributional assumptions.

Increased Sensitivity of Operon-wise Significance Estimates

We hypothesized that when genes in operons have consistent measurements, higher confidence can be assigned to those measurements. As described in the Methods, we calculated "operon-wise" p-values that, for each gene, take into account the data for other genes in

the same operon (if such genes exist; otherwise the operon-wise and single-gene p-values are identical). To test whether operon-wise p-values were more powerful than single-gene p-values, we compared the distributions of the operon-wise significance values to that of the single-gene significance values. Significance was defined as $1 - C$. As shown in Figure 4, the operon-wise significance estimates are much more confident in each of the data sets, and at a significance cutoff of 0.01, 2-10 times more genes can be identified.

Discussion

We have described how operons can be used to detect systematic errors in measurements of prokaryotic gene expression patterns, to account for the bias when estimating significance, and to increase the confidence of measurements that are consistent within an operon. The method relies on the assumption that genes in the same operon have matching expression profiles. Although this assumption is only approximately correct, it is effective in practice, and is strongly preferable to ignoring the presence of systematic errors in the data. The method also relies on assumptions about the distributions of the true means and variances of the data. These assumptions are not entirely accurate, but without such assumptions, it would not be possible to distinguish low agreement within operons due to replication noise from that due to systematic bias. In simulations, the method was robust to the observed deviations from the assumptions.

In four data sets, the method identified significant and sometimes large amounts of systematic error. If this bias is not taken into account, as is generally the case with current approaches, then the statistical analysis will be far too aggressive. Our results indicate that the bias is not an artefact arising from errors in operon predictions or from our distributional assumptions. When using methods that did not correct for bias, for the three data sets that had large amounts of bias, we observed excellent agreement with operons for the most significant genes, and much less agreement for genes that were only slightly less significant if bias was not taken into account (e.g., 99% confident instead of 99.9% confident). This strongly suggests the methods that did not correct for bias were over-confident and that the bias is genuine. For the fourth data set, the bias was small, and it is possible that the rare cases where genes within the same operon truly have different expression patterns (Adhya 2003) might be responsible for the apparent bias. It is also conceivable that the method could underestimate bias if the systematic errors were somehow correlated for genes in the same operon. For example, as genes in the same operon are in the same mRNA molecules, any bias due to the RNA extraction step might not be detected.

Where does the bias come from? Many potential sources of biases are known, including intensity-dependent effects, spatial biases, dye incorporation biases, and cross-hybridization or non-specific hybridization. The normalization methods we used should correct for the

intensity and spatial effects, and we would also expect that the effects remaining after normalization might vary between replicates and hence might not be systematic. The shCold5 data set included a dye swap, which should correct for dye biases, but this set still had very high bias. Because the shCold5 and shHeat5 data sets, which showed the most bias, were measured only 5 minutes after the stress was applied, we considered the possibility that the mRNA levels were far from steady-state and that some operons would have poor agreement because of differential mRNA decay. However, later time points from these same experiments showed similar amounts of bias (data not shown).

Another source of apparent bias might be correlation between the replicates. That is, if the replicate measurements are not truly independent and some of the replicates are correlated then the noise of the averaged replicate will be larger than expected. For example, the shHeat5 data set had a total of 12 measurements per gene (3 biological samples times two slides per sample with dyes reversed times two spots per gene on each slide). In this data set, the replicate measurements with the same dye assignment were more correlated than those with reversed dyes. To test the pattern of bias with fully independent replicates, we created two subsets of the data. First, we used only the first spot for each gene on the slides and a single biological replicate, leaving 2 replicates with different dye assignments. Second, we used only a single dye assignment and only the 1st spot per slide, leaving 3 replicates from different samples. In both cases, we still observed large amounts of bias (data not shown). We also verified that our method was not sensitive to correlations between replicates. We created an exact duplicate of each replicate, and this "doubled" data set gave significance values very similar to the original data set (results not shown).

The data set without major bias (ecox) was collected using Affymetrix gene chips, which use multiple probes per gene, and was normalized with a method that attempts to identify "bad" probes and remove them from the data (Li and Wong 2001). We speculate that cross-hybridization or non-specific hybridization can create problems for some probes, and that when multiple probes are used, this bias can be removed. Such biases have been observed directly. For example, after hybridizing identical samples to Affymetrix and two-color microarrays, Kuo *et al.* (2002) inferred both kinds of bias from discrepancies between the measurements. A correlation of bias with GC content was attributed to non-specific hybridization, while a correlation of bias with the existence of paralogs was attributed to cross-hybridization. We tested for these sources of bias in the shHeat5 data set, using agreement within operons as a crude measure of bias. However, we did not see any relationship between paralogs and agreement. We did see a significant (albeit weak) tendency for probes with lower GC content to show stronger agreement, but as the genes with low-GC probes also showed larger changes in expression, this need not reflect bias. Additional experimental data will be required to clarify the source of bias and to estimate its magnitude more directly.

Irrespective of bias and for all four data sets, the operon-wise method identified many more

genes at any desired level of significance than the single-gene method. Although we only tested the operon-wise approach with one method for assessing significance, in principle, operon-wise p-values can be computed using single-gene p-values from any method. However, operon-wise p-values cannot necessarily be used in the same way as single-gene p-values. For example, it is common practice to rank genes by their significance. Lonnstedt and Speed (2001) go so far as to provide a ranking without an absolute significance scale. The operon-wise p-values may not be suitable for ranking, as genes in consistent operons that did not show large changers can be ranked highly, and these could well be indirect effects that are of less biological interest. To produce a ranked list of genes, we recommend either (i) setting a confidence threshold and then ranking the genes above that confidence level by their fold-change or (ii) testing a more stringent hypothesis than whether the gene changed in the measured direction, such as whether the gene's expression changed by 1.5-fold or more. Another attractive possibility is to rank operons rather than individual genes, but the confidence values will still be higher for larger operons. In any case, the main benefit of the present work is not for ranking or other broad exploratory analyses but in the ability to obtain plausible p-values for specific hypotheses of the form "was gene X or operon Y up-regulated in this experiment?"

As microarray technology becomes less expensive, experiment designs with high amounts of replication are becoming common. We observed that the systematic error can be comparable to or even larger than the variation betwen replicates ($\gamma < 1$, as in shHeat5 and shCold5). If systematic error is large relative to replication error, then performing many replicate measurements may not be cost-effective. If these systematic errors are due to probe-specific effects, as seems likely, then using several different probes for each gene might be preferable. Alternatively, control experiments to measure non-specific hybridization and cross-hybridization might be possible.

Finally, although the method we describe here requires operons and is only applicable to prokaryotic data, a similar approach might be useful for eukaryotes as well. The key requirement is for prior knowledge of pairs of genes that have matching expression patterns. These could perhaps be identified among pairs that are known to be both functionally related and co-expressed (in data other than that being analyzed). For example, in yeast, many stable complexes are known, large amounts of expression data are available, and genes that code for stable complexes are often strongly co-expressed (Jansen *et al.* 2002). And the worm *C. elegans*, which is a model system for studying development, has "operons" of a sort, but the extent of co-expression is less clear than for prokaryotic operons (Lercher *et al.* 2003).

Acknowledgements

We thank Zhili He and Jizhong Zhou for pre-publication access to data and Pat Flaherty for suggesting that we examine the correlation between replicates. This work was supported by a grant from the DOE GTL program (DE-AC03-76SF00098).

References

Adhya, S. (2003) Suboperonic regulatory signals. *Sci. STKE*, **2003**, pe22.

Aubert, J., Bar-Hen, A., Daudin, J.-J. and Robin, S. (2004) Determination of the differentially expressed genes in the microarray experiments using local FDR. *BMC Bioinformatics*, **5**.

Baldi, P. and Long, A. D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–19.

Bernstein, J. A., Khodursky, A. B., Lin, P. H., Lin-Chao, S. and Cohen, S. N. (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 9697–702.

Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J. and Palsson, B. O. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, **429**, 92–6.

Dudoit, S., Yan, Y. H., Speed, T. P. and Callow, M. J. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Ermolaeva, M. D., White, O. and Salzberg, S. L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–21.

Gao, H., Wang, Y., Liu, X., Yan, T., Wu, L., Alm, E., Arkin, A., Thompson, D. K. and Zhou, J. (2004) Global transcriptome analysis of the heat shock response of Shewanella oneidensis. *J. Bacteriol.*, **186**, 7796–803.

Ideker, T., Thorsson, V., Siegel, A. F. and Hood, L. E. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comp. Bio.*, **7**, 805–17.

Jansen, R., Greenbaum, D. and Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.*, **12**, 37–46.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G. and Gibson, G. (2001) The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nat. Genet.*, **29**, 389–95.

Kerr, M. K., Martin, M. and Churchill, G. A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–37.

Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L. and Kohane, I. S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–12.

Lercher, M. J., Blumenthal, T. and Hurst, L. D. (2003) Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. *Genome Res.*, **13**, 238–43.

Li, C. and Wong, W. H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA*, **98**, 31–6.

Lonnstedt, I. and Speed, T. (2001) Replicated microarray data. *Statistica Sinica*, **12**, 31–46.

Moreno-Hagelsieb, G. and Collado-Vides, J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18 Suppl. 1**, S329–36.

Price, M. N., Huang, K. H., Alm, E. J. and Arkin, A. P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, **33**, 880–92.

Sabatti, C., Rohlin, L., Oh, M. K. and Liao, J. C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.*, **30**, 2886–93.

Self, S. G. and Liang, K.-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.*, **82**, 605–610.

Selinger, D. W., Saxena, R. M., Cheung, K. J., Church, G. M. and Rosenow, C. (2003) Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. *Genome Res.*, **13**, 216–23.

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

Smyth, G. K., Yang, Y. H. and Speed, T. (2003) Statistical issues in cdna microarray data analysis. *Methods Mol. Biol.*, **224**, 111–36.

Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, **100**, 9440–5.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, **98**, 5116–21.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Table 1: Relationship between means and variances in the data and in simulations. r_S indicates a Spearman (ranked) correlation coefficient, and subscripts 1 and 2 indicate pairs of adjacent genes that are predicted to be in the same operon ($P(Operon_{ij}) > 0.5$).

Data Set	$r_S(\|m\|, s^2)$	$r_S(m_1, m_2)$	$r_S(s_1^2, s_2^2)$
dvSalt30			
Actual	0.209	0.405	0.137
Model	0.268	0.434	0.536
Uncoupled	0.123	0.469	-0.000
ecox			
Actual	0.284	0.655	0.412
Model	0.339	0.708	0.554
Uncoupled	0.052	0.785	0.003
Mixture	0.313	0.545	0.547
shHeat5			
Actual	0.235	0.727	0.169
Model	0.262	0.684	0.499
Uncoupled	0.065	0.684	0.002
shCold5			
Actual	0.156	0.403	0.200
Model	0.328	0.465	0.573
Uncoupled	0.175	0.423	0.002

Table 2: Systematic bias in four biological data sets. The typical size of the bias in the apparent $\log_2$-ratio is the square root of its variance, or $\sqrt{E(\frac{1}{\theta_i \cdot \gamma})}$, where $E(1/\theta_i) = \frac{\alpha}{\nu - 1}$. The bias over the signal is the square root of the ratio of variances ($\sqrt{\beta/\gamma}$). The bias over the replication error is also the square root of the ratio of variances ($\sqrt{1/\gamma}$), and considers a single measurement (is not divided by the number of replicates). To show that the bias is statistically significant, we compared the likelihood ratio of the best-fitting model given systematic error to that without (with $\gamma = \infty$; see Eq. 15). Because we are testing whether γ lies at a boundary, in the absence of bias the distribution of $2 \cdot \log(ratio)$ approximates a 50:50 mixture of two chi-squared distributions with 0 and 1 degrees of freedom (Self and Liang 1987).

	dvSalt30	ecox	shHeat5	shCold5
Typical bias	0.25	0.12	0.37	0.88
Bias / signal (%)	70.4%	19.6%	49.9%	86.9%
Bias / replication error (%)	72.7%	35.8%	143.1%	199.1%
Significance of bias				
Likelihood ratio	1.74e+02	9.38e+00	1.48e+03	1.81e+03
p-value	$< 10^{-77}$	$< 10^{-5}$	$< 10^{-646}$	$< 10^{-786}$

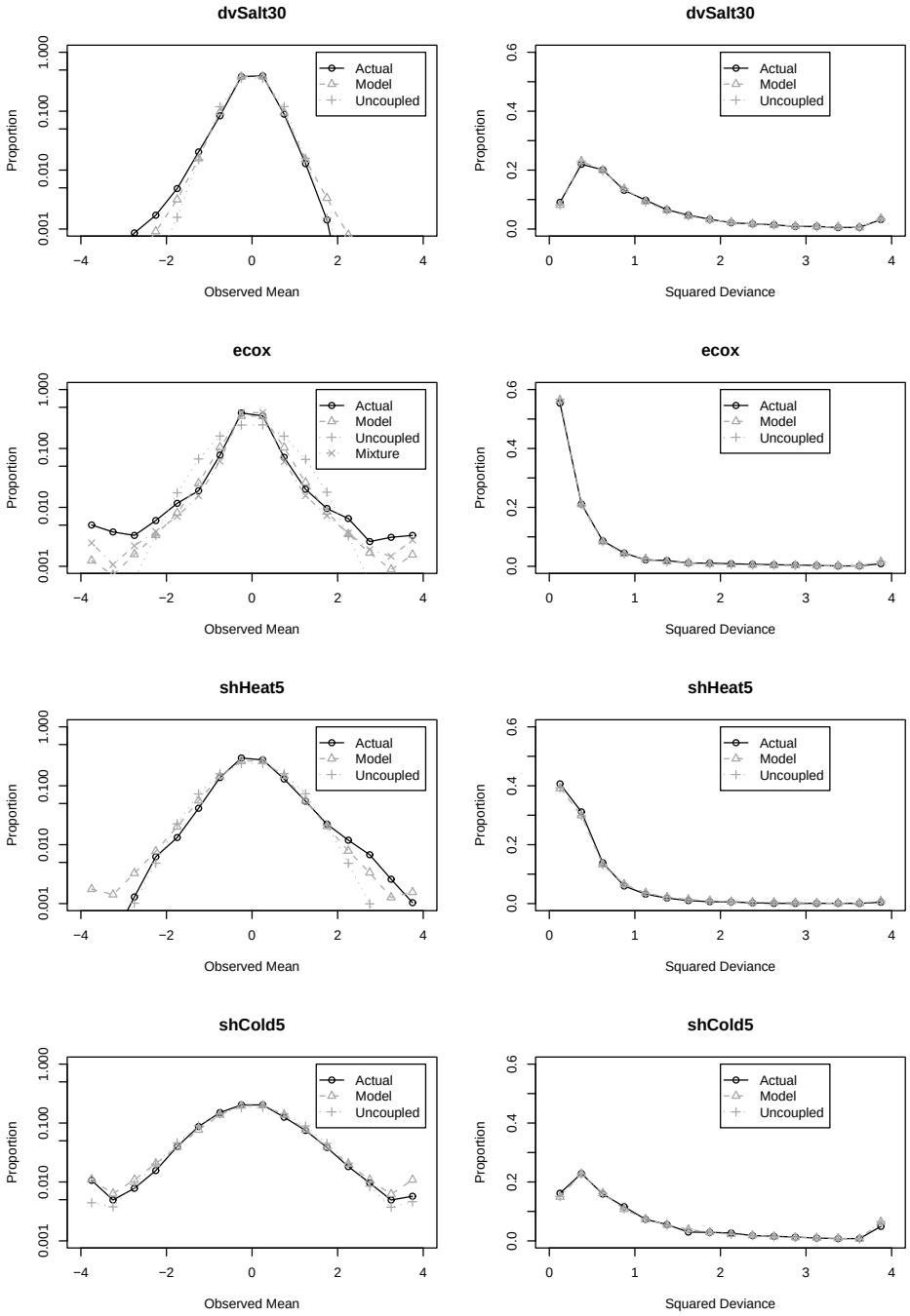


Figure 1: Distributions, in actual and simulated data, for observed means (left) and squared total deviances (right). The leftmost and rightmost bins include all more extreme values. Note the log y-axis for the means.

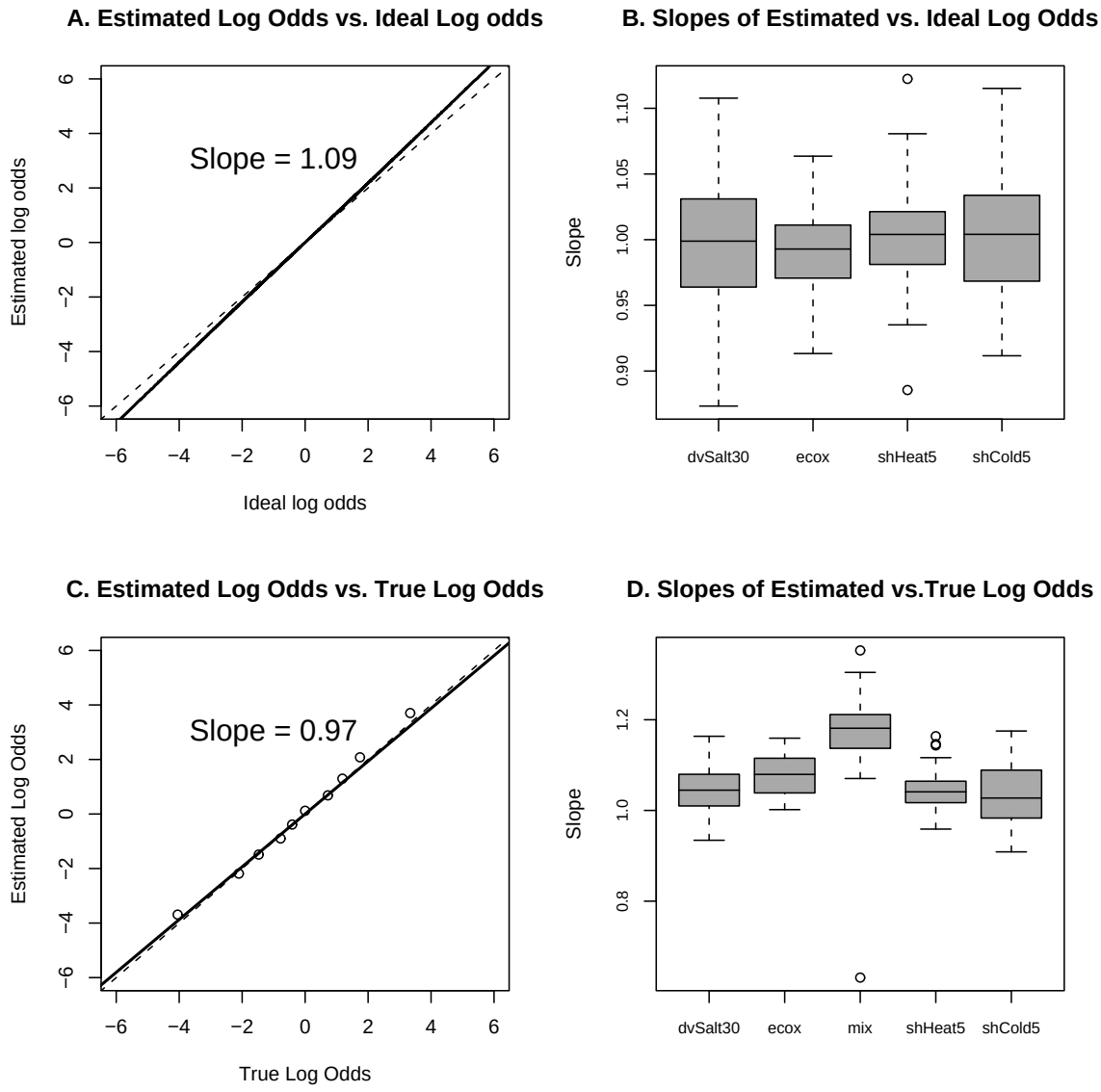


Figure 2: Accuracy of p-values in simulations. (A) For a simulation matching the error model, the solid line shows the estimated log odds for each gene ($\log \frac{P(\mu_i>0)}{1-P(\mu_i>0)}$) as a function of the "ideal" log odds based on the true values of the hyperparameters. The dashed line shows $x = y$. A linear regression with the intercept fixed at zero gives a slope of 1.09. (B) The distribution of these slopes across 50 simulations for each data set's hyperparameters. The boxes show the first and third quartiles and the medians, the whiskers show the most extreme point within 1.5 times the inter-quartile range of the box, and the points indicate outliers. (C) For an "uncoupled" simulation where means and variances were independent, we sorted the genes by their estimated log odds into 10 bins of equal size. For each bin, a point shows the true log odds (from the number of genes with $\mu_i > 0$ and $\mu_i < 0$) and the average of the estimated log odds. Logistic regression gave a slope of 0.97 (shown with a solid line). (D) The distribution of these slopes from 50 uncoupled simulations for each data set and from the "mixture" simulation for the ecox data set.

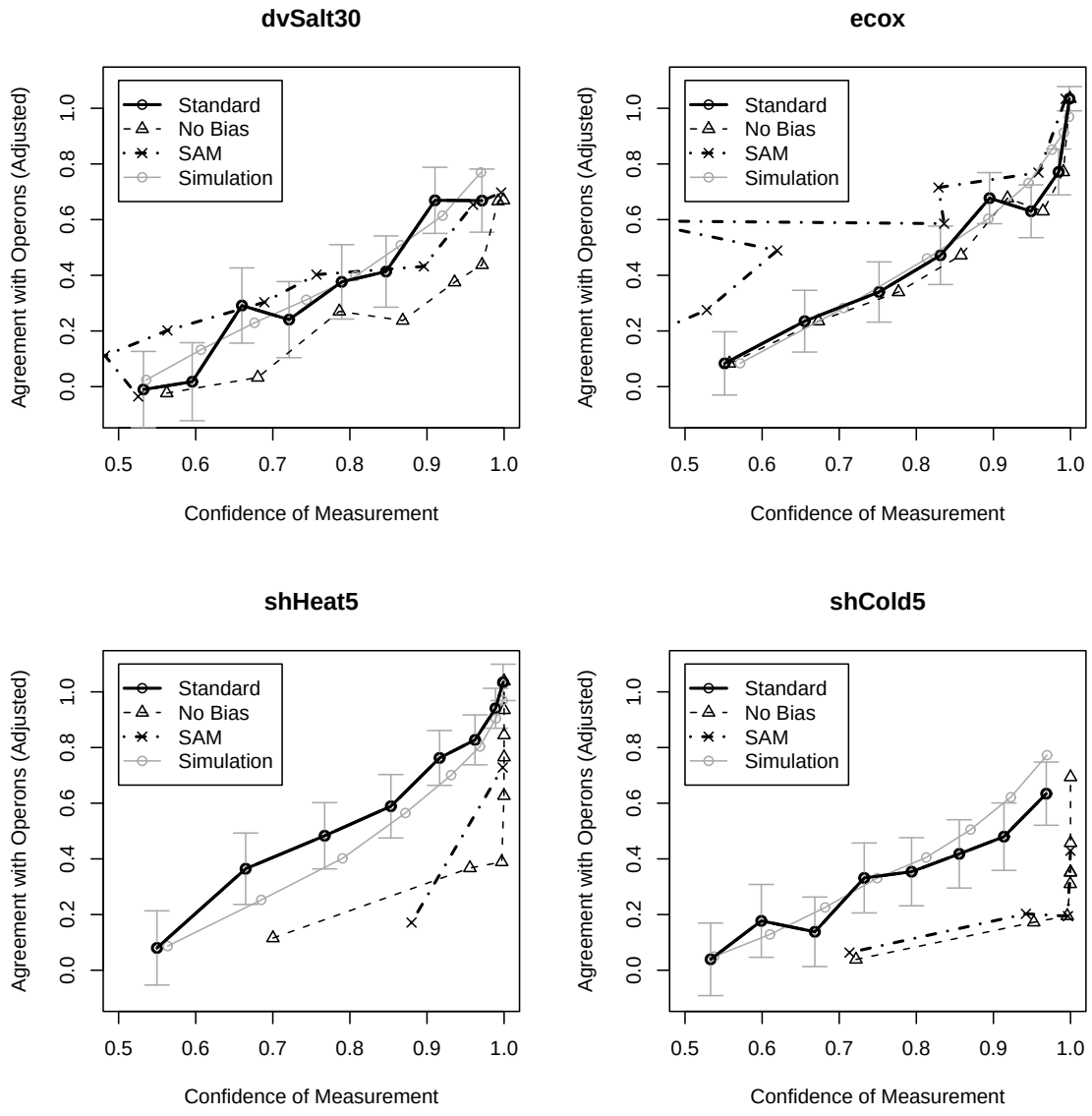


Figure 3: Single-gene significance and agreement with operons. For each data set and for three methods of assessing significance (our standard model, our model without bias, and significance analysis of microarrays), we divided the changers into eight groups of genes with different levels of confidence. The x axis shows the average confidence within each group of genes. For each group, the y axis shows the adjusted agreement with operon pairs (the adjusted proportion of pairs which have the same sign of log-ratio), which ranges from 0 for random pairs to 1 for perfect measurements. We also show average results from simulations for each data set (simulated and analyzed with our standard model). The error bars give the 95% confidence interval (from a t test) for the mean agreement for each group from the "standard" (bias-adjusted) significance values.

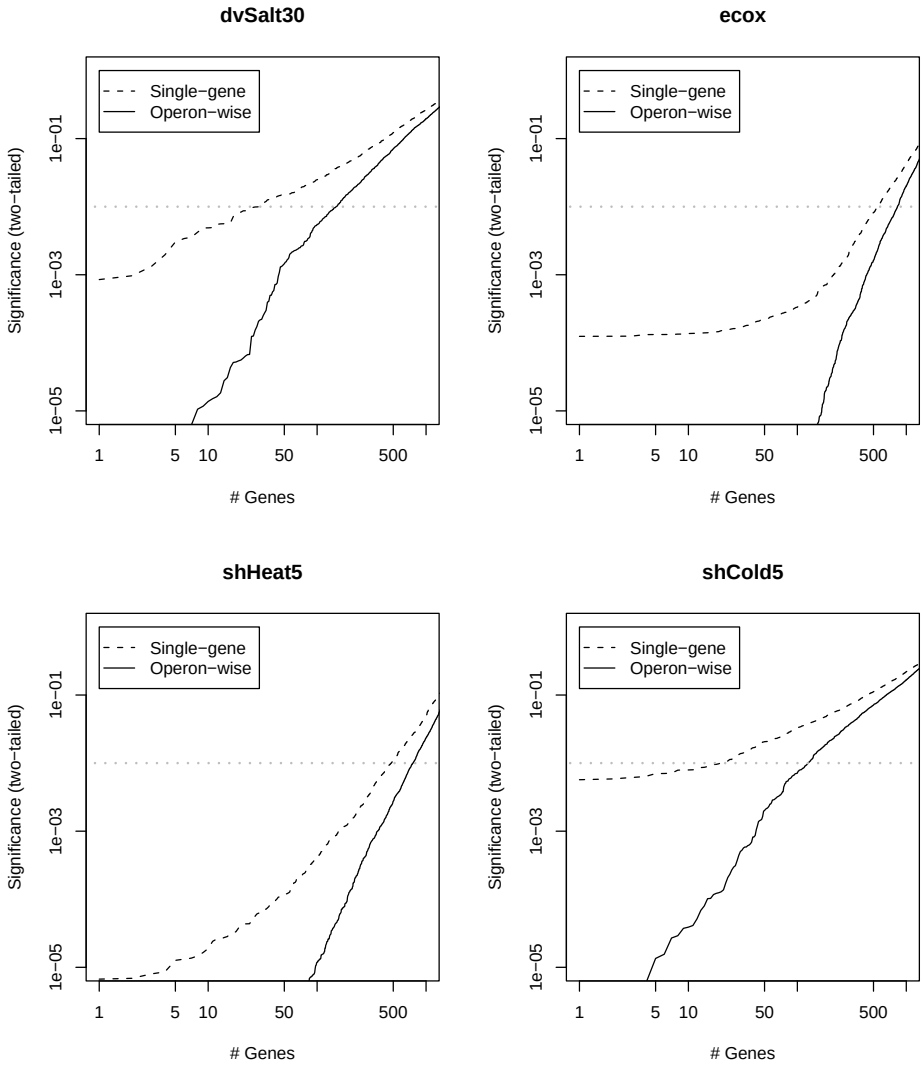
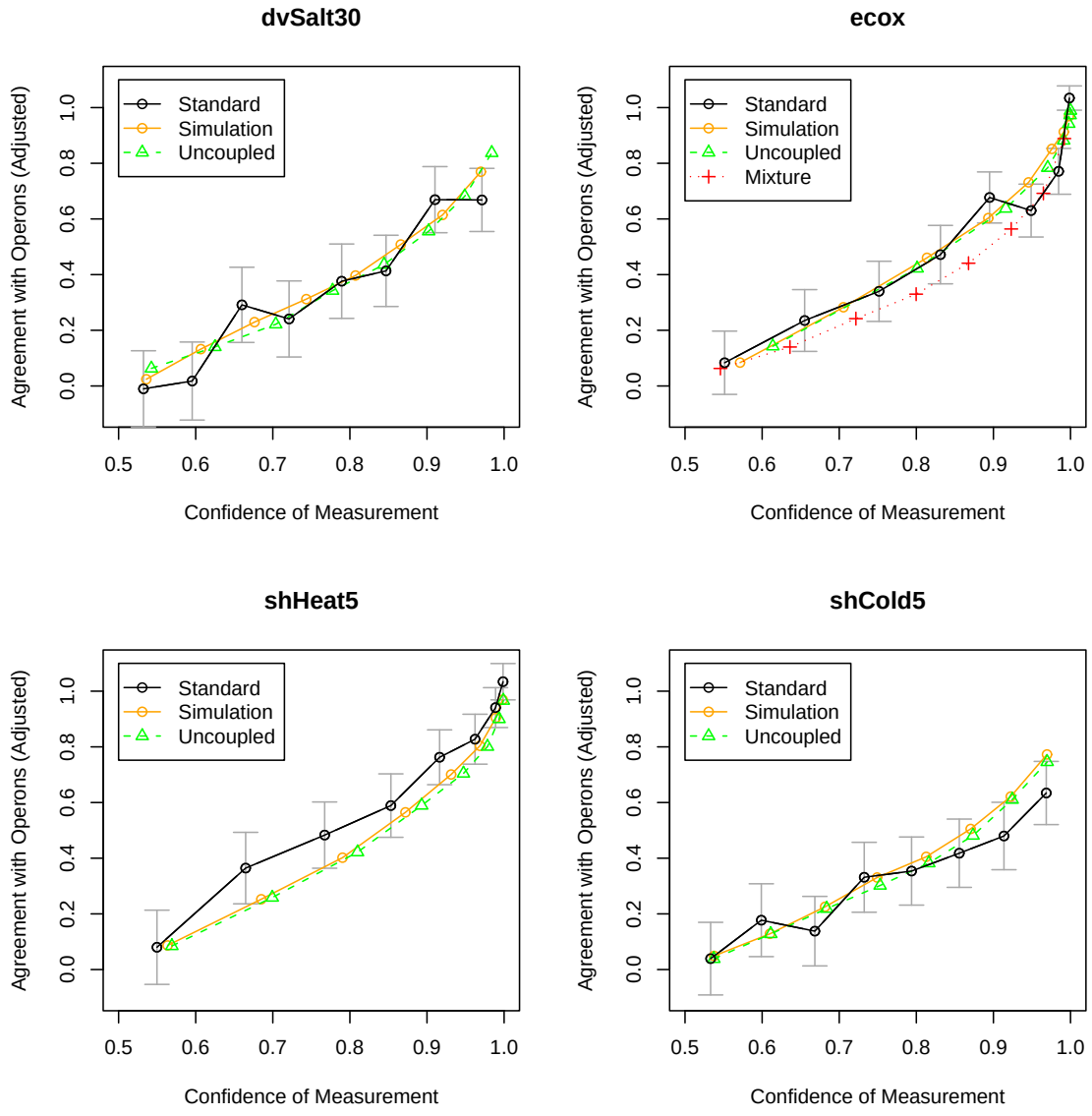


Figure 4: Sensitivity of single-gene and operon-wise methods. For each data set, we show the cumulative number of changers identified at varing levels of significance. Note the log scales. The horizontal line is at 0.01. Genes that are not in operons are included in the operon-wise results.



Supplementary Figure 1: Single-gene significance and agreement with operons. As in Figure 3, we divided the changers into eight groups of genes with different levels of confidence, and compare the average confidence within each group to the adjusted agreement with operon pairs. Here we compare the actual data set to additional "uncoupled" and (for ecox) "mixture" simulations.