4-16-2004

Screening for Differentially Expressed Genes: Are Multilevel Models Helpful?

Dongmei Liu
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University, gp@jimmy.harvard.edu

Brian Caffo
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Screening for differentially expressed genes: are multilevel models helpful?

Dongmei Liu, Giovanni Parmigiani and Brian Caffo
Johns Hopkins University

March 22, 2004

Acknowledgment. Data from Dudley, Aach, Steffen and Church (2002) was downloaded from `http://arep.med.harvard.edu/masliner/supplement.htm`. Data from Tusher, Tibshirani and Chu (2001) was downloaded from `http://www-stat-class.stanford.edu/SAM/SAMServlet`.

Work of Giovanni Parmigiani supported by grants NCI 5P30-A06973-39, NCI P50CA88843, NCI P50CA62924-05, DK-58757.

Abstract

Screening for changes in gene expression across biological conditions using microarrays is now a common tool in biology. Efficient use of these data for identifying important biological hypotheses is inherently a statistical problem. In this paper we present a broad Bayesian multilevel framework for developing computationally fast shrinkage-based screening tools for this purpose. Our scheme makes it easy to adapt the choice of statistics to the goals of the analysis and to the genomic distributions of signal and noise. We empirically investigate the extent to which these shrinkage-based statistics improve performance, and the conditions under which such improvements takes place. Our evaluation uses both extensive simulations and controlled biological experiments. The experimental data include a so-called spike-in experiment, in which the target biological signal is known, and a two-sample experiment, which illustrates the typical conditions in which the methods studied are applied.

Our results emphasize two important practical concerns that are not receiving sufficient attention in applied work in this area. First, while shrinkage strategies based on multilevel models are able to improve selection performance, they require careful verification of the assumptions on the relationship between signal and noise. Incorrect specification of this relationship can negatively affect a selection procedure. Because this inter-gene relationship is generally identifiable in genomic experiments, we suggest a simple diagnostic plot to assist model checking. Secondly, no statistic performs optimally across two common categories of experimental goals: selecting genes with large changes, and selecting genes with reliably measured changes. Therefore, careful consideration of analysis goals is critical in the choice of the approach taken.

1 Background

Many genomics investigations using expression arrays take the form of searching for genes whose expression level is different across experimental conditions or phenotypes. The list of gene transcripts produced by a microarray analysis is usually the starting point for extensive additional biological work, including independent validation, and both in-silico and laboratory work on sequences and proteins related to the transcripts selected. In this context, microarray experiments are screening, not testing, experiments. Because of the wide range of important questions that can be explored using these arrays, and the costs involved, comparisons across conditions are often made using a limited number of replications. Efficient use of data is critical in improving a laboratory's ability to correctly identify important biological hypotheses and proceed to test them by appropriate further experimentation.

Specific screening goals vary with the study. Two simple but representative situations are the selection of genes that are changed by a large amount, and the selection of genes that are changed by a reliably measured amount. In either case, the comparison of gene expression across two conditions based on replicated experiments requires a trade-off of signal, the variation of expression across the two conditions, versus noise, the variation of expression within each condition. Therefore the problem is statistical in nature (Kohane et al. 2002, Speed 2003, Parmigiani et al. 2003). In this paper we discuss a Bayesian multilevel framework for developing screening tools that adapt to the goals of the analysis and to the genomic distributions of signal and noise. We evaluate a representative set of these tools using both extensive simulations and controlled biological experiments in which the set of altered genes is known.

A variety of approaches for selecting differentially expressed genes have been proposed (see Pan (2002) for a review). The simplest and still the most widely used is to set a threshold on a measure of signal alone, for example an estimated fold–change. This can be motivated by the desire to identify large changes, although often it is used by simple analogy with other gene expression essays that have much less noise. Upper and lower thresholds of two and one half are often seen in applications. One limitation of this approach is that it does not consider how reliably gene-specific changes are measured. That is, it implicitly assumes that all genes are subject to the same level of noise. This may not be the case because even after appropriate preprocessing of the data, the within–gene variation in expression can be highly gene–dependent.

A straightforward way to account for both signal and noise is to select genes based on statistics motivated by two–sample testing, such as the T-ratio or the Wilcoxon statistic. For each gene, the T–ratio is an estimate of the signal–to–noise ratio. Because it requires

estimating two or three parameters instead of one, when the number of replicates is small, the T–ratio does not necessarily perform better than fold–change, even when the goal is point-null-like. Gains in efficiency over both fold–change and T–ratios can be obtained by considering the ensemble of gene expression measures at once, rather than each gene in isolation. This occurs for at least two reasons. First, genes measured on the same array type in the same laboratory are all affected by a number of common sources of noise. Secondly, many changes in expression are part of common biological mechanisms.

A widely used approach that uses genome–wide information is "significance analysis of microarrays" or SAM (Tusher et al. 2001). SAM involves transforming the signal–to–noise ratios so that they are approximately independent of noise across genes. The type of transformation used by SAM is designed to protect against false discoveries generated by very small denominators in the t–ratios. The denominators that are really small are highly likely to be so by chance, because genes share many sources of variability, and a certain amount of variation is to be expected from all of them.

More broadly, joint estimation of many related quantities is often approached by multilevel modeling, and the associated Empirical Bayesian (Robbins 1956, Efron and Morris 1973) and Hierarchical Bayesian (Lindley and Smith 1972) estimation techniques. See Carlin and Louis (2000) for a detailed discussion. In genomics, these may represent variation in two stages. The first stage defines summaries at the gene level, for example test statistics, or estimates of fold change and noise. These describe variability of samples within each gene. The second stage posits a "genomic" distribution for these gene–level summaries. Such multilevel modeling provides tools for borrowing strength from other genes when making inference on each gene. Some examples of implementations in microarrays are provided by Baldi and Long (2001), Newton, Kendziorski, Richmond, Blattner and Tsui (2001), Efron, Tibshirani, Storey and Tusher (2001), Lönnstedt and Speed (2002), Ibrahim, Chen and Gray (2002), Parmigiani, Garrett, Anbazhagan and Gabrielson (2002), among others.

In practice, a question often raised by genomics practitioners is the extent to which simple, real-time, shrinkage statistics motivated by multilevel models would outperform single-gene-at-a-time analysis or SAM. In this paper we set out to systematically address this question. To this end, we found it necessary to develop a general framework for developing and evaluating these fast shrinkage statistics. Our answer will turn out to be that shrinkage can furnish substantial improvement over single-gene-at-a-time analysis or SAM, provided that the statistics chosen will a) take into account the goals of the screening experiments and b) will be chosen based on examination of the properties of the genomic distribution of signal and noise. Even though we considered an extensive collection of statistics, the goal was not that of providing an exhaustive comparison of all approaches that have been proposed, but

rather that of highlighting the critical role of the signal–to–noise trade–off and of providing tools to choose among alternative approaches based on the genome–wide behavior of signal and noise. We compared the performance of these statistics using both extensive simulations and real data sets in which fold changes were known.

2 Methods

2.1 Multilevel Models for Two-group Comparisons

We consider a design in which two biological types are compared on a microarray that probes G genes. Each type is measured on n arrays using either technical or biological replicates. Here technical replicates refer to experiments that have multiple aliquots of the same RNA, while biological replicates refer to experiments that have multiple subjects from a population. Each situation requires a different interpretation of the array-to-array variability, but the formal structure is the same. We do not consider both levels of replication at the same time here. We denote by X_{1gj} the expression for gene g in sample j in the first group, and by X_{2gj} the expression for gene g in sample j in the second group. Expression levels are assumed to be centered around an overall experiment-wise mean.

Recall our interest lies in studying approaches for selecting genes that are differentially expressed between groups. We begin by describing an additive group effect and independent Gaussian errors. That is we assume that the observed expressions are conditionally independent draws from

$$\begin{aligned} X_{1gj}|\mu_g, \sigma_g^2, \delta_g &\sim N\left(\mu_g - \frac{1}{2}\delta_g, \sigma_g^2\right) \\ X_{2gj}|\mu_g, \sigma_g^2, \delta_g &\sim N\left(\mu_g + \frac{1}{2}\delta_g, \sigma_g^2\right). \end{aligned}$$

Here, δ_g is the difference in expression level for gene g across groups, μ_g is an overall expression level for gene g, also referred to as abundance, or intensity, and σ_g^2 is the variance of expression level for gene g in both groups. We refer to δ_g as true signal, and to σ_g as true noise. in a multilevel setting, our parameterization is different from that assuming $E\{X_{1gj}\} = \mu_g$ $E\{X_{2gj}\} = \mu_g + \delta_g$, which would lead to two different marginal variances in the two groups. The fit of the normal distribution can often be improved by a suitable transformation of the data. Departures from normality and unequal variance across groups are not considered in this manuscript.

Multilevel models postulate a distribution for the abundance, signal, and noise parameters across genes. A common assumption to many of the multilevel models used in microarray analysis is that of conjugate distributions for the second stage of the statistical model, in

Data Level

$$X_{1gj}|\mu_g, \sigma_g^2, \delta_g \sim N\left(\mu_g - \frac{1}{2}\delta_g, \sigma_g^2\right)$$
$$X_{2gj}|\mu_g, \sigma_g^2, \delta_g \sim N\left(\mu_g + \frac{1}{2}\delta_g, \sigma_g^2\right).$$

II. Independence	CI. Independence of Signal and Noise
$\mu_g\|\tau^2 \sim N(0,\tau^2)$ $\delta_g\|\lambda^2 \sim N(0,\lambda^2)$ $\sigma_g^{-2}\|\nu,\beta \sim Ga(\nu,\beta)$	$\mu_g\|\tau^2,\sigma_g^2 \sim N(0,\sigma_g^2\tau^2)$ $\delta_g\|\lambda^2 \sim N(0,\lambda^2)$ $\sigma_g^{-2}\|\nu,\beta \sim Ga(\nu,\beta)$
IC. Independence of Abundance and Noise	**CC. Complete Conjugacy**
$\mu_g\|\tau^2 \sim N(0,\tau^2)$ $\delta_g\|\lambda^2,\sigma_g^2 \sim N(0,\sigma_g^2\lambda^2)$ $\sigma_g^{-2}\|\nu,\beta \sim Ga(\nu,\beta)$	$\mu_g\|\tau^2,\sigma_g^2 \sim N(0,\sigma_g^2\tau^2)$ $\delta_g\|\lambda^2,\sigma_g^2 \sim N(0,\sigma_g^2\lambda^2)$ $\sigma_g^{-2}\|\nu,\beta \sim Ga(\nu,\beta)$

Table 1: The four classes of multilevel models investigated. The array-to-array variation is modeled in the same way in all four cases. In all cases, $j = 1, 2, \ldots, n$ and $g = 1, 2, \ldots, G$. All quantities denoted by Greek letters are unknown. A further set of prior distributions for the hyperparameters is described in the text.

short a "conjugate model". In the case of Gaussian data, the conjugate model implies that the gene–specific signal–to–noise ratios and abundance-to-noise ratios are independent of the corresponding gene–specific noise (Raiffa and Schleifer 1961, Ando and Kaufman 1965). This assumption leads to convenient mathematical representations for many of the steps required by the data analysis, and is sometimes adopted solely for this reason. In practice, however, some microarray experiments follow this independence pattern closely, while others depart from it substantially. The loss of efficiency of screening based on the conjugate model in the latter case can be large.

Here we broaden the conjugate scheme and we investigate four model varieties, that result from the combination of two factors: (i) whether the gene-specific signal is independent of the gene-specific noise, (ii) whether the gene-specific abundance is independent of the gene-specific noise. Formally, for (i) the independence models assumes that δ_g and σ_g are independent, while the conjugate model assumes that δ_g/σ_g and σ_g are independent. The remainder of our distributional assumptions are standard for normal multilevel models (Lindley and Smith 1972, Gelman et al. 1995). The models are summarized in Table 1.

We use the notation d_g for the mean difference of expression across two groups, a_g for the overall mean expression, and s_g for the pooled estimate of the standard deviation. Notationally:

$$\begin{aligned} a_g &= \frac{1}{2}(\bar{X}_{1g} + \bar{X}_{2g}) \\ d_g &= \bar{X}_{2g} - \bar{X}_{1g} \\ s_g^2 &= \frac{\text{RSS}}{n-1} = \frac{1}{n-1}\sum_{j=1}^{n}(X_{1gj} - \bar{X}_{1g})^2 + \frac{1}{n-1}\sum_{j=1}^{n}(X_{2gj} - \bar{X}_{2g})^2. \end{aligned}$$

In all four models of Table 1, these statistics are independent conditional on gene-specific parameters and have distributions

$$\begin{aligned} a_g &\sim N\left(\mu_g, \frac{1}{2n}\sigma_g^2\right) \\ d_g &\sim N\left(\delta_g, \frac{2}{n}\sigma_g^2\right) \\ \frac{n-1}{\sigma^2}s_g^2 &\sim \chi^2_{2(n-1)}. \end{aligned}$$

conditional on gene-specific parameters. All four combinations of Table 1 occur commonly in practice. For example, our two experimental data sets show two markedly different relationships between signal and noise.

The vector of unknown parameters will be denoted by $\xi = (\nu, \beta, \lambda, \tau)$. There are several estimation approaches available for models of this kind. State-of-the art, computationally intensive approaches are usually based on MCMC (Gilks et al. 1996). Instead we focus on a faster and simpler empirical Bayesian approach based on estimating ξ by method of moments from the empirical distributions of s_g^{-2}, d_g's. The resulting estimators are computationally cheap and may include shrinkage of the signal, of the noise, or both. Several method of moments alternatives are available, and results can be strongly affected by this choice. For example, in our experience, the method of moments applied to the distribution of s_g^2, which is inverse gamma, performs poorly, while the same applied to s_g^{-2} performs well.

We approach the task of generating a list of candidate genes by ranking genes according to a one-dimensional statistic, and then selecting all genes whose statistic is above a certain cutoff. This is the norm in practice. While more general decision theoretic approaches evaluating the trade-off between false and missed discoveries are available (Müller et al. 2003, Lin et al. 2003), these are complex, and would have been prohibitive in our vast simulation study. The cutoff is often determined by the ability of a laboratory to perform validatory analyses, or, more inferentially, by false discovery rates (Benjamini and Hochberg 1995, Genovese and Wasserman 2001, Storey 2002, Storey and Tibshirani 2003). In our

MOTIVATING MODEL	ANALYSIS GOAL	
	Large change	Reliably measured change
Independence/Normality of Genes	Difference in Expression (F)	T–statistic (T)
Exchangeability of Genes	Significance Analysis of Microarray (SAM)	
Complete Conjugacy	Signal (CC.F)	Standardized Signal (CC.T)
	Tail probability (CC.TP)	Bayes factor (CC.BF)
Independence of Abundance and Noise	Signal (CI.F)	Standardized Signal (CI.T)
	Tail probability (CI.TP)	Bayes factor (CI.BF)
Independence of Signal and Noise	Signal (IC.F)	Standardized Signal (IC.T)
	Tail probability (IC.TP)	Bayes factor (IC.BF)
Independence	Signal (II.F)	Standardized Signal (II.T)
	Tail probability (II.TP)	Bayes factor (II.BF)

Table 2: Summary of statistics examined, by goal and motivating model structure.

presentation, to simplify the comparison of approaches, we focus on the ranking of genes implied by the statistics, and the ability of each statistic of identifying the top g genes.

Throughout, we draw a distinction between the selection of genes that are changed by a large amount, and genes that are changed by a reliably measured amount. Accordingly we consider two broad families of statistics, ones that estimate the signal, δ_g, and ones that estimate the signal to noise ratio, δ_g/σ_g^2. Because we use statistics as ranking devices and compare them based on ROC curves, we only need to define statistics up to constants that are not gene-specific. A proportionality sign will indicate omission of such constants. Table 2 summarizes the statistics we examined, organizing them by goal and motivating model structure. Table 3 summarizes the expressions of statistics motivated by multilevel models.

2.2 Statistics

In this subsection we enumerate and briefly comment on each of the statistics we considered. The remainder of this section is provided as a reference for future sections. Details of the derivation are given in the Appendix.

Difference in Expression (F). This is the observed average difference d_g. Usually expression data are analyzed in the logarithmic scale, in which case **F** corresponds to an estimate of the log fold change across conditions.

T–statistic (T). This is the common statistics $T \propto d_g/s_g$ used for testing the null hypothesis of $\delta_g = 0$ one gene at the time.

Significance Analysis of Microarrays (SAM). This was proposed by Tusher, Tibshirani and Chu (2001) and is based on the change of gene expression relative to an adjusted

Model	Statistics			
	F	T	BF	TP
CC.m	d_g	$\frac{d_g}{\sqrt{\hat{\sigma}_g^2}}$	$\left(\frac{1+\frac{1}{\hat{\beta}}\left(\frac{(n-1)}{2}s_g^2+\frac{n}{4}d_g^2+\frac{\frac{a_g^2}{2}}{\frac{1}{2n}+\hat{\tau}^2}\right)}{1+\frac{1}{\hat{\beta}}\left(\frac{(n-1)}{2}s_g^2+\frac{\frac{d_g^2}{2}}{\frac{2}{n}+\hat{\lambda}^2}+\frac{\frac{a_g^2}{2}}{\frac{1}{2n}+\hat{\tau}^2}\right)}\right)^{-(n+\hat{\nu})}$	$Pr(\delta_g > D\|d_g, a_g, s_g^2, \hat{\xi})$ $\delta_g\|d_g, a_g, s_g^2, \hat{\xi} \sim St\left(\frac{\frac{n}{2}d_g}{\frac{n}{2}+\frac{1}{\hat{\lambda}^2}}, \left(\frac{n}{2}+\frac{1}{\hat{\lambda}^2}\right)\frac{\nu_n}{\beta_n}, 2(\hat{\nu}+n)\right)$
CC.c	d_g	$\frac{d_g}{\sqrt{\hat{\sigma}_g^2}}$	$e^{\frac{-\frac{n}{2}d_g^2+\frac{d_g^2}{\frac{2}{n}+\hat{\lambda}^2}}{2\hat{\sigma}_g^2}}$	$Pr(\delta_g > D\|d_g, \hat{\sigma}_g^2, \hat{\xi})$ $\delta_g\|d_g, \hat{\sigma}_g^2, \hat{\xi} \sim N\left(\frac{\frac{n}{2}d_g}{\frac{n}{2}+\frac{1}{\hat{\lambda}^2}}, \frac{\hat{\sigma}_g^2}{\frac{n}{2}+\frac{1}{\hat{\lambda}^2}}\right)$
IC	d_g	$\frac{d_g}{\sqrt{\hat{\sigma}_g^2}}$	$e^{\frac{-\frac{n}{2}d_g^2+\frac{d_g^2}{\frac{2}{n}+\hat{\lambda}^2}}{2\hat{\sigma}_g^2}}$	$Pr(\delta_g > D\|d_g, \hat{\sigma}_g^2, \hat{\xi})$ $\delta_g\|d_g, \hat{\sigma}_g^2, \hat{\xi} \sim N\left(\frac{\frac{n}{2}d_g}{\frac{n}{2}+\frac{1}{\hat{\lambda}^2}}, \frac{\hat{\sigma}_g^2}{\frac{n}{2}+\frac{1}{\hat{\lambda}^2}}\right)$
CI	$\frac{\frac{nd_g}{2\hat{\sigma}_g^2}}{\frac{n}{2\hat{\sigma}_g^2}+\frac{1}{\hat{\lambda}^2}}$	$\frac{\frac{nd_g}{2\hat{\sigma}_g^2}}{\sqrt{\frac{n}{2\hat{\sigma}_g^2}+\frac{1}{\hat{\lambda}^2}}}$	$\sqrt{\frac{\left(\frac{2}{n}\hat{\sigma}_g^2+\hat{\lambda}^2\right)}{\hat{\sigma}_g^2}}e^{-\frac{\frac{n}{2}d_g^2}{2\hat{\sigma}_g^2}+\frac{d_g^2}{2\left(\frac{2}{n}\hat{\sigma}_g^2+\hat{\lambda}^2\right)}}$	$Pr(\delta_g > D\|d_g, \hat{\sigma}_g^2, \hat{\xi})$ $\delta_g\|d_g, \hat{\sigma}_g^2, \hat{\xi} \sim N\left(\frac{\frac{nd_g}{2\hat{\sigma}_g^2}}{\frac{n}{2\hat{\sigma}_g^2}+\frac{1}{\hat{\lambda}^2}}, \frac{1}{\frac{n}{2\hat{\sigma}_g^2}+\frac{1}{\hat{\lambda}^2}}\right)$
II	$\frac{\frac{nd_g}{2\hat{\sigma}_g^2}}{\frac{n}{2\hat{\sigma}_g^2}+\frac{1}{\hat{\lambda}^2}}$	$\frac{\frac{nd_g}{2\hat{\sigma}_g^2}}{\sqrt{\frac{n}{2\hat{\sigma}_g^2}+\frac{1}{\hat{\lambda}^2}}}$	$\sqrt{\frac{\left(\frac{2}{n}\hat{\sigma}_g^2+\hat{\lambda}^2\right)}{\hat{\sigma}_g^2}}e^{-\frac{\frac{n}{2}d_g^2}{2\hat{\sigma}_g^2}+\frac{d_g^2}{2\left(\frac{2}{n}\hat{\sigma}_g^2+\hat{\lambda}^2\right)}}$	$Pr(\delta_g > D\|d_g, \hat{\sigma}_g^2, \hat{\xi})$ $\delta_g\|d_g, \hat{\sigma}_g^2, \hat{\xi} \sim N\left(\frac{\frac{nd_g}{2\hat{\sigma}_g^2}}{\frac{n}{2\hat{\sigma}_g^2}+\frac{1}{\hat{\lambda}^2}}, \frac{1}{\frac{n}{2\hat{\sigma}_g^2}+\frac{1}{\hat{\lambda}^2}}\right)$

Table 3: Summary of functional forms of all the statistics motivated by multilevel models. For the complete conjugacy case (CC) we consider both statistics that use analytic integration with respect to σ_g (labeled CC.m) and statistics that use plug-in estimates of σ_g (labeled CC.c).

standard deviation. For the two group case considered here, the SAM statistic for gene g is

$$\text{SAM} = \frac{d_g}{s_g + s_0}$$

where s_0 is the so-called "exchangeability factor". This factor is estimated using information from the entire set of genes to transform the values of SAM so that noise and SAM are approximately independent.

Statistics for Complete Conjugacy (CC). In the Complete Conjugacy model the conditional posterior distribution of δ_g given hyperparameters ξ can be written as

$$\delta_g|d_g, \sigma_g^2, \xi \sim N\left(\frac{\frac{n}{2}d_g}{\frac{n}{2}+\frac{1}{\lambda^2}}, \frac{\sigma_g^2}{\frac{n}{2}+\frac{1}{\lambda^2}}\right) \tag{1}$$

A set of computationally cheap statistics is derived by considering empirical Bayes estimates of this distribution, obtained by replacing hyperparameters ξ with an estimate $\hat{\xi}$. A regularized estimate of signal δ_g is

$$\text{CC.F} = \frac{\frac{n}{2}d_g}{\frac{1}{\hat{\lambda}^2}+\frac{n}{2}} \propto d_g.$$

Regularization is independent of the gene, so for any given experiment CC.F will be proportional to F. For this reason we only consider F, although we keep this correspondence in mind when interpreting the results.

A standardized estimate of signal is derived as the ratio of the conditional posterior mean and standard deviation of δ_g from expression (1),

$$\text{CC.T} \quad \propto \quad \frac{d_g}{\sqrt{\hat{\sigma}_g^2}}.$$

The denominator incorporates a linear shrinkage estimate of the gene-specific variance with gene-varying coefficients, penalizing more heavily genes whose signal or abundance are outlying. For this reason, it is critical that the conjugacy assumption be checked, or very valuable information may be lost. On the other hand, when the assumption is met, an increase in efficiency is gained from estimating the denominator.

An empirical Bayes estimate of the Bayes factor (Kass and Raftery 1995) for the null hypothesis of no gene-specific differential expression, $\delta_g = 0$, is:

$$\text{CC.BF} \propto \left(\frac{1 + \frac{1}{\hat{\beta}} \left(\frac{(n-1)}{2} s_g^2 + \frac{n}{4} d_g^2 + \frac{\frac{a_g^2}{2}}{\frac{1}{2n} + \hat{\tau}^2} \right)}{1 + \frac{1}{\hat{\beta}} \left(\frac{(n-1)}{2} s_g^2 + \frac{\frac{d_g^2}{2}}{\frac{2}{n} + \hat{\lambda}^2} + \frac{\frac{a_g^2}{2}}{\frac{1}{2n} + \hat{\tau}^2} \right)} \right)^{-(n+\hat{\nu})}.$$

Finally, we consider the empirical Bayes approximation,

$$\text{CC.TP} = Pr(\delta_g > D | d_g, a_g, s_g^2, \hat{\xi}),$$

of the probability that the true change δ_g exceeds D. Here, D represents a target change across conditions. This tail probability reflects the observed change, its variability and the likely magnitude of biologically significant changes.

Statistics for Independence of Abundance and Noise (IC). In this model the posterior distribution of δ_g given σ_g and hyperparameters ξ can be written as

$$\delta_g | d_g, \sigma_g^2, \xi \sim N \left(\frac{\frac{n}{2} d_g}{\frac{n}{2} + \frac{1}{\lambda^2}}, \frac{\sigma_g^2}{\frac{n}{2} + \frac{1}{\lambda^2}} \right) \tag{2}$$

Unlike in the complete conjugate case, a closed form marginalization with respect to σ is not possible. Therefore we derive results assuming σ_g^2 is known. In the actual calculations, to obtain a real-time statistics, σ_g^2 is estimated by the posterior mode of the distribution of $\sigma_g^2 | s_g^2, \xi$. Then our estimate of the normalized signal is

$$\text{IC.F} = \frac{\frac{n}{2} d_g}{\frac{n}{2} + \frac{1}{\lambda^2}} \propto d_g.$$

As with CC.F, IC.F is proportional to F, so we only consider F in our results section.

A standardized estimate of signal based on regularized estimates of signal is the the ratio of the marginal posterior mean and standard deviation of δ_g from expression (2), that is

$$\text{IC.T} \quad \propto \quad \frac{d_g}{\sqrt{\hat{\sigma}_g^2}}.$$

The Empirical Bayes estimate of the Bayes factor, conditional on gene specific variance is:

$$\text{IC.BF} \quad \propto \quad e^{\frac{-\frac{n}{2}d_g^2+\frac{d_g^2}{\frac{2}{n}+\hat{\lambda}^2}}{2\hat{\sigma}_g^2}}.$$

Finally, we consider the empirical Bayes tail probability

$$\text{IC.TP} = Pr(\delta_g > D | d_g, \hat{\sigma}_g^2, \hat{\xi}).$$

Statistics for Independence of Signal and Noise (CI). In this model the posterior distribution of δ_g given σ_g and hyperparameters ξ can be written as

$$\delta_g | d_g, \sigma_g^2, \xi \sim N\left(\frac{\frac{nd_g}{2\sigma_g^2}}{\frac{n}{2\sigma_g^2}+\frac{1}{\lambda^2}}, \frac{1}{\frac{n}{2\sigma_g^2}+\frac{1}{\lambda^2}} \right) \tag{3}$$

Again, to obtain a real-time statistic we develop results conditional on σ_g^2 and estimate it with its posterior mode in actual calculation. The estimate of the signal is

$$\text{CI.F} = \frac{\frac{nd_g}{2\hat{\sigma}_g^2}}{\frac{n}{2\hat{\sigma}_g^2}+\frac{1}{\hat{\lambda}^2}}.$$

A standardized estimate of signal based on regularized estimates of signal is the the ratio of the marginal posterior mean and standard deviation of δ_g from expression (3), that is

$$\text{CI.T} \quad = \quad \frac{\frac{nd_g}{2\hat{\sigma}_g^2}}{\sqrt{\frac{n}{2\hat{\sigma}_g^2}+\frac{1}{\hat{\lambda}^2}}}.$$

The Empirical Bayes estimate of the Bayes factor is:

$$\text{CI.BF} \quad \propto \quad \sqrt{\frac{\left(\frac{2}{n}\hat{\sigma}_g^2+\hat{\lambda}^2\right)}{\hat{\sigma}_g^2}} e^{-\frac{\frac{n}{2}d_g^2}{2\hat{\sigma}_g^2}+\frac{d_g^2}{2\left(\frac{2}{n}\hat{\sigma}_g^2+\hat{\lambda}^2\right)}},$$

while the tail probability approximation is

$$\text{CI.TP} = Pr(\delta_g > D | d_g, \hat{\sigma}_g^2, \hat{\xi}).$$

Statistics for Complete Independence (II). In this model the posterior distribution of δ_g given σ_g and hyperparameters ξ can be written again as

$$\delta_g | d_g, \sigma_g^2, \xi \sim N\left(\frac{\frac{nd_g}{2\sigma_g^2}}{\frac{n}{2\sigma_g^2} + \frac{1}{\lambda^2}}, \frac{1}{\frac{n}{2\sigma_g^2} + \frac{1}{\lambda^2}} \right) \tag{4}$$

A regularized estimate of δ_g, motivated by the independence model is obtained by replacing ξ with $\hat{\xi}$ and σ_g^2 with its conditional posterior mode evaluated at $\hat{\xi}$, and approximating the posterior mean by

$$\text{II.F} = \frac{\frac{nd_g}{2\hat{\sigma}_g^2}}{\frac{n}{2\hat{\sigma}_g^2} + \frac{1}{\hat{\lambda}^2}}.$$

Unlike IC.F and CC.F, both II.F and CI.F imply a linear shrinkage which depends on the genomic variability of the signal. Dividing II.F by the square root of the variance of δ_g, and approximating as before, we obtain

$$\text{II.T} = \frac{\frac{nd_g}{2\hat{\sigma}_g^2}}{\sqrt{\frac{n}{2\hat{\sigma}_g^2} + \frac{1}{\hat{\lambda}^2}}}.$$

The Empirical Bayes estimate of the Bayes factor is:

$$\text{II.BF} \quad \propto \quad \sqrt{\frac{\left(\frac{2}{n}\hat{\sigma}_g^2 + \hat{\lambda}^2\right)}{\hat{\sigma}_g^2}} e^{-\frac{\frac{n}{2}d_g^2}{2\hat{\sigma}_g^2} + \frac{d_g^2}{2\left(\frac{2}{n}\hat{\sigma}_g^2 + \hat{\lambda}^2\right)}},$$

while the empirical Bayes approximation to the tail probability is

$$\text{II.TP} = Pr(\delta_g > D | d_g, \hat{\sigma}_g^2, \hat{\xi}).$$

Notice that the definitions of statistics for the CI and II cases would be the same if σ_g^2 was known. Hence the only difference in practice is the posterior mode for σ_g^2. It should not then be a surprise that the performance of these two are very close. For the same reason, this is also true for statistics in the CC and IC settings conditional on σ_g^2.

The empirical Bayes estimators, both standardized and not, have functional similarities to the SAM score, although shrinkage of the noise in the denominators are determined differently. In empirical Bayes analyses, the shrinkage is driven by the parameters of the genomic distributions of signal and noise, in a form that depends on whether or not conjugacy is assumed. In SAM one applies linear shrinkage to the standard deviation rather than the variance, and the shrinkage intercept s_0 is chosen to approximate independence of SAM ratios from noise.

3 Simulation Results

3.1 Design of simulation study

We study simulated data sets from each of the four models of Table 1. We considered three sample sizes: 3, 10 and 100 per group. Use of samples as small as 3 is a common scenario in the gene screening experiments taking place during the routine activities of many laboratories, while 10 per group is a common scenario in comparisons across conditions for systematic genomic studies. Sample size as large as 100 per group are rare and considered here mostly as a check.

For each combination of scenario and sample size, we simulated data from 2009 models, resulting in a total of 24108 datasets. The 2009 combinations of hyperparameters are based on the grid:

$$\begin{aligned} E[\sigma_g^{-2}] &\in \left(\frac{1}{100}, \frac{1}{25}, \frac{1}{5}, 1, 5, 25, 100\right) \\ var[\sigma_g^{-2}] &\in \left(\frac{1}{100}, \frac{1}{25}, \frac{1}{5}, 1, 5, 25, 100\right) E[\sigma_g^{-2}] \\ \lambda^2 &\in \left(\frac{1}{100}, \frac{1}{25}, \frac{1}{5}, 1, 5, 25, 100\right) E[\sigma_g^{-2}] \\ \tau^2 &\in \left(\frac{1}{100}, \frac{1}{25}, \frac{1}{5}, 1, 5, 25, 100\right) E[\sigma_g^{-2}]. \end{aligned}$$

For each combination we derive ν and β from $E[\sigma_g^{-2}]$ and $var[\sigma_g^{-2}]$. Here the total number of combinations is 2009 rather than 2401 because some expectation/variance combinations lead to unrealistic settings for ν yielding potentially numerically unstable results.

We applied simple real-time shrinkage statistics in the analysis to avoid time consuming numerical integration. Whenever there is no close form analytically, we use the conditional estimation by plugging in the posterior mode of σ_g^2. To verify how reasonable these real-time statistics are, we considered the Complete Conjugacy model where closed forms are available for all the statistics. We performed another set of simulation with exactly the same hyperparameters comparing the results based on CC.TP and CC.BF in two cases: one based on the conditional posterior of δ_g with posterior mode of σ_g^2 plugged in, the other based on the marginal posterior of δ_g integrating out σ_g^2.

3.2 Summary of Simulation Results

Mining the massive information generated by the 24 thousand scenarios required drastic summarization. Our approach has been to identify a subset of changes "of interest" based on the simulated parameters, and computing, for each statistic, the area under the ROC curve obtained when using the statistic as a diagnostic of change (Egan 1975, National

Research Council; Panel on Discriminant Analysis Classification and Clustering 1988). In the analyses presented here we considered two perspective: in one the genes of interest are the top genes by absolute change δ_g. In the other the genes of interest are the top genes by signal-to noise ratio δ_g/σ_g. We considered the top 1%, 2% and 10%, with 1% being the baseline case. For a given threshold on a statistic, some of the genes declared to be changed are truly changed (true positives) while others are not (false positives). We analyze this correspondence by an ROC curve is a graph of the true positive fraction versus the false positive fraction for varying thresholds. As a summary, we consider the area under the ROC curve (Pepe et al. 2000). We prefer this measure to others incorporating δ_g and σ_g/δ_g explicitly for two reasons: the goal of the microarray experiments we are trying to models is screening rather then estimation; interest usually lies in a relative small fraction of important findings.

Based on these criteria, our simulations suggest two general conclusions about the alternative approaches for identifying differential genes: i) simple, real-time, shrinkage statistics motivated by multilevel modes can outperform alternatives based on analyzing each gene separately, in some cases by a large margin; ii) the same statistics can perform better than the commonly used SAM (Tusher et al. 2001) statistic, provided careful checking of the multilevel modeling assumptions.

In more detail, Figure 1 gives the signal-to-noise (SN) plots (Tusher et al. 2001, Dudoit et al. 2002) in examples from each of the four model settings, II, IC, CI and CC. We graph boxplots of signal by noise deciles. An SN plot in which the location and dispersion of signal are stable across noise levels suggests the use of an independence model, while one in which the location and dispersion of signal increase with noise level suggests the use of a conjugate model. Thus, a constant box size indicates independence while an increasing box size indicates conjugacy. For these four data sets, the diagnostic plot clearly distinguishes conjugacy with respect to signal and noise as well as conjugacy with respect to abundance and noise.

Figure 2 illustrates the ROC curves for a single simulation from the complete independence model for identifying genes with large signals. The areas under these curves constitute the data used to graph a single point in the subsequent plots. For this particular data set, we see a large separation in the performance of the statistics, both in the statistics motivated by different models and the different statistics within a given model. For this simulation, the tail probability statistics are the clear winners regardless of the motivating model.

Figure 3 summarizes the 2009 simulated data sets from the complete independence model with three replicates. Each point represents the paired areas under the curve for two statistics for a simulated data set. In the plots above the diagonal, we report the ROC curves for

identifying genes with large reliably measured changes, while in plots below the diagonal we report the ROC curves for identifying genes with large signal. Here the the T and SAM statistics are compared to the four statistics motivated by the complete independence model. The II.BF and II.F statistics, have the best performance for detecting reliably measured differentially expressed genes and large signal genes respectively. Note that it is important to compare the summary statistics in this plot as over-plotting points leads to visually misleading results in some cases.

If this figure were repeated on the same simulated data sets using the statistics motivated by the IC model, the IC.BF statistic is preferred for detecting reliably measured differential expression while the IC.F statistic is preferred for detecting raw differential expression (signal). To summarize, for the data simulated from the complete independence model, the best statistics from each of the four motivating model were II.BF, IC.BF, CI.BF and CC.T for detecting large signal to noise ratios and II.F, IC.F, CI.F and CC.F for detecting large signals. Note that the CC.T statistic invokes the same ranking of genes as the CC.BF statistic as shown in the Appendix. The areas under the ROC curves for these statistics for the complete independence simulation data sets are plotted against one another in

Figure 4. Also plotted are results for the SAM statistic and the T ratio. Two important features are present in this plot. First, the SAM statistic generally performs reasonably well across models, but worse than any of these optimal multilevel modelling statistics within a model. Secondly, the model for the abundance and noise has little effect on the performance of the statistic. For example, the II.F and CI.F and statistics behave similarly for detecting large signals. The best performing statistics assume independence of the signal and the noise. Finally, Figure 5, presents a comparison with best performing statistics from other models as well.

The discussion above concerns only simulation from the complete independence model. Rather than reproducing these plots for each of the simulation scenarios, Figures 6, 7, 8, and 9 summarize the results for the four simulation settings, using "heat maps" that synthesize pairwise comparisons of estimators. Furthermore, the best performing statistics from the heat maps for sample sizes 3, 10 and 100 replicates are summarized in Tables 4, 5 and 6. To see the impact of parameter estimation, the results obtained by plugging in the true parameter values are also given. A wildcard "*" indicates the statistic from either the conjugate or independence model was the best performer. For example, the "*C.F" in Table 4 indicates that the CC.F and IC.F statistics were roughly equivalent best performers for data simulated from the complete independence model with three replicates. These tables indicate that accounting for the experimental goal is important for all sample sizes. Fold changes and tail probabilities appear to be the best statistics for estimating large signals changes.

In contrast the modified t ratios and Bayes factors appear to be optimal for estimating reliably measured differential expression. Moreover, accounting for the appropriate modelling assumptions becomes increasingly important with the number of replicates.

We investigate the differences between the CC.TP and CC.BF statistics calculated exactly by integrating over σ_g^2 (marginal) versus the real time approximation obtained by plugging in the posterior mode of σ_g^2 (conditional). Results are shown in Figure 11. Real-time statistics generally perform as well as the exact statistics in most cases, although there is a minority of cases in which the exact statistics does much better than real-time approximation, especially for tail probabilities.

Finally, we further investigated the seemingly counterintuitive result where the best performing statistic for the data simulated from the Complete Conjugacy model are the IC.TP, IC.BF for genes ranked by both signal alone and signal-to-noise ratio. This behavior persists at larger sample sizes. Figure 10 focuses on the comparison between IC.TP and CC.TP for data simulated from the complete conjugacy model. The reason for the counterintuitive behavior is that some of the hyperparameter combinations lead to simulated dataset that have diagnostic plots consistent with an IC model, in which case IC statistic performs well while the abundance-based shrinkage applied by the CC statistics leads to loss of some of the signal.

4 Experimental Results

4.1 Two-group comparison

The first data set we study was reported by Tusher, Tibshirani and Chu (2001) in the context of comparing radiated and unirradiated cell lines. A subset of the genes' changes, identified based on the SAM statistic, were subsequently validated by independent essays. While the experiment includes some blocking, we analyze it here as though it were a two–class comparison with 4 replicates.

We begin by investigating the relationship between signal and noise. Figure 12 reports the diagnostic plots for the Tusher data. Because results are sensitive to the type of transformation applied to the data, we consider both the original scale and the cube root. Untransformed data show a pattern consistent with the conjugate model, while data transformed using the cube root appear consistent with the independent model.

Figure 14 shows pairwise scatterplots of the statistics CC.TP, II.F, and SAM for the two transformations in Figure 12. CC.TP and II.F are the two best statistics among all shrinkage estimators to identify genes with large signal. We focus on these two statistics and compare them to SAM. The two best statistics based on multilevel modeling for selecting reliably

measured genes, CC.BF and II.T, are also plotted against SAM in Figure 15. In evaluating these results, one must keep in mind that only genes that exceeded a certain SAM threshold were validated independently in the study. Therefore, direct performance comparisons with SAM are not reliable here.

4.2 Spike-in Data Set

The second data set is from an experiment reported by Dudley, Aach, Steffen and Church (2002). They performed a so called "spike-in" experiment to study how to measure absolute expression with a calibrated reference sample and an extended signal intensity range. Their experiment used cDNA microarrays and included a total of 6307 genes. They selected 9 genes with very low expression and "spiked-in" Cy3-labeled gene-specific oligos in increments from 0.5 fold to 200 fold. The experiment had two replicates. We work from ratios of Cy3-to-Cy5 channels, after normalization (Dudoit and Yang 2003). While spike in experiments are useful in that true fold changes are known, both the magnitudes of the changes, and the sparsity of changes in the genome are unrealistic.

Figure 13 investigates the relationship between signal and noise using three transformations. The original scale shows a marked positive relationship between estimated signal and noise, the cube root scale a mild positive relationship, and the logarithm an almost stable relationship, with some indication of larger variation in the signal at lower noise level. These figures do not inform us about absolute intensity, so the larger variation of signal at the low end after the log transformation is not the same as the well known "fishtail" effect observed in MVA plots.

In this data set the cube root transformation is the most effective in helping identify the truly differentially expressed genes, performing better than the commonly used log. In practical application one does not have the advantage of knowing the true changes when choosing a transformation. The important lesson here is, however, that choosing transformations based on convenient statistical properties such as variance stabilization does not necessarily improve, and could prejudice, our ability to detect signal.

These two datasets stress that both the independence of signal and noise and independence of signal-to-noise ratio and noise may need to be tackled in real applications. While transformation of the measured intensities may allow one to achieve independence, it is not clear that such transformations would be optimal in terms of gene screening.

Figures 17 and 18 compare statistics. For the log transformed spike-in data, we would expect a better performance from II.F than CC.TP based on the SN plot. In fact, the II.F statistics shrinks the effects excessively and gives a less efficient ranking. For the cube root transformed spike-in data we would expect and, in fact, see a better performance from

CC.TP than II.F in Figure 17. For the untransformed spike–in data, Figure 17 confirms the intuition from the exploratory plots that the conjugacy model should outperform the independence model. Figure 17 does show this result. In a close view of comparison between CC.TP and SAM on raw data (Figure 19), CC.TP also clearly picks up all the spiked genes, while SAM does not. Spiked genes are genes have large signals, so that poor performance of II.T and CC.BF is no surprise in Figure 18.

This data warns of a danger in the use of multilevel modelling. Assuming that the signals follow a Gaussian law assumes that all genes are differentially expressed to some extent, and the goal is to either detect the largest signals or the largest reliably measured signals. This is realistic in case-control comparisons and in experiments in which the experimental intervention changes a large portion of the expression, as during cell division. In contrast, in this spike-in experiment, the distribution of signals is in fact degenerate at 0 for all but the 8 spiked in genes. Therefore, the statistics motivated by the Gaussian law on the signal are not validated from the data. While extreme, the spike-in situation may be relevant in practice when experimental intervention modifies a small set of genes involved in a very specialized pathway.

Diagnosing empirically whether the signal distribution is a mixture is difficult. Appropriate weight should be given to the biological circumstances of the experiment. For example, here an independence relationship is suggested for the signal and noise for the cubic and log transformed data. However, as the majority of the genes are biologically known to have no signal, these plots do not inform us on the question of interest. Furthermore, Figures 17 and 18 show that the best complete independence statistics for identifying large signal and reliably measured signal changes, II.F and II.T, perform poorly for detecting the spiked-in genes. In summary, aggressively modelling the distribution of signals when the overwhelming majority of genes have no signal can produce poor results.

Finally, we note that an alternative visualization for the SN plot is a scatterplot of signal versus noise. A limitation of this approach is that it can be difficult to establish whether increased variation in signal at different level of noise is due to a true relationship or simply to a higher number of genes at that noise level. Figure 16 shows the scatterplot of signal versus noise for the two data sets with no transformation. It is more difficult to identify a relationship between signal and noise in these graphs than it was in the boxplot. On the other hand, in the spike-in data, it is easier to assess the position of the changed genes using the scatterplot. Therefore it may be useful to use both in practice.

5 Conclusion

In identifying differentially expressed genes using microarrays, improvements in efficiency over simple fold–change have been pursued in two main directions: consideration of noise in addition to signal, and borrowing of strength from the genomic distribution. Systematic approaches for achieving these goals are being developed using multilevel models, fit using either hierarchical Bayes or empirical Bayes methods. In this article we present a framework for interpreting, selecting, and estimating shrinkage-based screening statistics for identifying differentially expressed genes. We also evaluated a representative set of these tools using both extensive simulations and controlled biological experiments in which the set of altered genes is known or partially known.

Our results emphasize two important practical concerns that are not receiving sufficient attention in applied work in this area. First, while shrinkage strategies based on multilevel models are able to improve selection performance, they require careful verification of the assumptions on the relationship between signal and noise. Incorrect specification of this relationship can negatively affect a selection procedure. Because this inter-gene relationship is generally identified in genomic experiments, we suggest a simple diagnostic plot to assist model checking. Secondly, no statistic performs optimally across two common categories of experimental goals: selecting genes with large changes, and selecting genes with reliably measured changes. Therefore, careful consideration of analysis goals is critical in the choice of the approach taken.

The commonly used SAM statistics emerges as a reasonable compromise between the two goals above and is, to some extent, automatically adaptive to different relationships between signal and noise. Improving on SAM is possible but requires careful validation of the assumptions about the upper level distribution. The assumption of conjugacy in the abundance dimension requires careful attention as it is not robust. In particular, estimators based on the CC assumption can be outperformed even on data generated under CC.

Our simulation analysis relies on the assumed normality of data. In practice, two aspects of it are critical. At the first state, in small samples, the functional form of the error distribution across samples is hard to assess. Normality at the second stage can be checked, and transformations may help, although the caveats discussed in Section 4.2 should be considered. Alternative multilevel models have been studied, for example by Newton and Kendziorski (2003) who consider gamma models, and Müller, Parmigiani, Robert and Rousseau (2003), who extend those to mixtures of gamma models. While these alternatives are worth serious consideration, here we focus on Gaussian models and statistics motivated by the Gaussian setting, primarily because in this way we can practically investigate a va-

riety of relevant statistics on a massive number of simulation scenarios. Other interesting multilevel approaches have been proposed to analyze designs that are more complex than the two-group comparisons considered here. We refer the reader to Kerr, Martin and Churchill (2000), Wolfinger, Gibson, Wolfinger, Bennett, Hamadeh, Bushel, Afshari and Paules (2001) and Kooperberg, Sipione, LeBlanc, Strand, Cattaneo and Olson (2002) for further details.

In our analysis we assumed that all genes on the array are potentially changed. This is realistic in case control designs across populations or comparison of cells at different stages of the cell cycle, regulation can be expected in the majority of genes, although differences will vary randomly and many genes will be changed by amounts that are smaller than noise. In tightly controlled experiments, such as a comparison of wildtype versus mutant mice, or treated and untreated cell lines, can result in differential expression in a small number of pathways and genes. While this situation can be reasonably handled in the framework considered here, it would be more accurately modeled by assuming that only a fraction of the genes are differentially expressed across groups A multilevel model could assume assume that a fraction of the δ_g are identically zero, while the rest are normally distributed (Lönnstedt and Speed 2002). In a separate manuscript, we are pursuing approaches that rely on the assumption that the distribution of signals is a mixture of a point mass at 0 and a continuous component.

Our focus here has been on simple and easy-to-compute statistics for gene selection. Multilevel models were used only to provide motivation and a conceptual framework for the derivation of the shrinkage statistics. More generally, multilevel models give raise to potentially more efficient strategies than those considered here, at the price of increased computational expense, for example MCMC of MCEM. Systematic exploration of those in the thousands of data sets considered here would have been impractical. However, our results suggest that that appropriate shrinkage is a critical part of gene selection and this will hopefully encourage practitioners to consider these more computing intensive approaches as well.

References

Ando, A. and Kaufman, G. (1965). Bayesian analysis of the independent multi-normal process —neither mean nor precision known, *Journal of the American Statistical Association* **60**: 347–358.

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(6)**: 509–519.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* **57**: 289–300.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall, Boca Raton, FL.

Dudley, A., Aach, J., Steffen, M. and Church, G. M. (2002). Measuring absolute expression with microarrays using a calibrated reference sample and an extended signal intensity range, *Proc. Natl. Acad. Sci. USA* **99(11)**: 7554–7559.

Dudoit, S. and Yang, J. (2003). Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data, *in* G. Parmigiani, E. Garrett, R. Irizarry and S. Zeger (eds), *The Analysis of Gene Expression Data: Methods and Software*, Springer Verlag, New York.

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

Efron, B. and Morris, C. (1973). Combining possibly related estimation problems (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological* **35**: 379–421.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association* **96**: 1151–1160.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*, Academic Press, New York.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (1995). *Bayesian Data Analysis*, Chapman and Hall, London.

Genovese, C. and Wasserman, L. (2001). Bayesian and frequentist multiple testing, *Discussion paper*, Department of Statistics, Carnegie Mellon University.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

Ibrahim, J. G., Chen, M. H. and Gray, R. J. (2002). Bayesian models for gene expression with DNA microarray data, *Journal of the American Statistical Association* **97**: 88–99.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *Journal of the American Statistical Association* **90**: 773–795.

Kerr, M. K., Martin, M. and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data, *Journal of Computational Biology* **7**: 819–837.

Kohane, I. S., Kho, A. and Butte, A. J. (2002). *Microarrays for an Integrative Genomics*, MIT Press, Cambridge, MA.

Kooperberg, C., Sipione, S., LeBlanc, M., Strand, A. D., Cattaneo, E. and Olson, J. M. (2002). Evaluating test statistics to select interesting genes in microarray experiments, *Human Molecular Genetics* **11**: 2223–2232.

Lin, R., Louis, T. A., Paddock, S. M. and Ridgeway, G. (2003). Loss function based ranking in two-stage, hierarchical models, *Technical report 2003-6*, Johns Hopkins University, Department of Biostatistics, `http://www.bepress.com/jhubiostat/paper6`.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion), *Journal of the Royal Statistical Society, Series B* **34**: 1–41.

Lönnstedt, I. and Speed, T. (2002). Replicated microarray data, *Statistica Sinica* **12**(1): 31–46.

Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2003). Optimal sample size for multiple testing: the case of gene expression microarrays, *Technical report 2003-3*, Johns Hopkins University, Department of Biostatistics, `http://www.bepress.com/jhubiostat/paper3`.

National Research Council; Panel on Discriminant Analysis Classification and Clustering (1988). *Discriminant Analysis and Clustering*, National Academy Press, Washington, D. C.

Newton, M. A. and Kendziorski, C. M. (2003). Parametric empirical bayes methods for micorarrays, *The analysis of gene expression data: methods and software*, Springer, New York.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. and Tsui, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data, *Journal of Computational Biology* **8**: 37–52.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments, *Bioinformatics* **18**: 546–554.

Parmigiani, G., Garrett, E. S., Anbazhagan, R. and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer, *Journal of the Royal Statistical Society, Series B* **64**: 717–736.

Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. L. (eds) (2003). *The analysis of gene expression data: an overview of methods and software*, Springer, New York.

Pepe, M. S., Longton, G., Anderson, G. L. and Schummer, M. (2000). Selecting differentially expressed genes from microarray experiments, *Biometrics* **59**: 133–142.

Raiffa, H. and Schleifer, R. (1961). *Applied Statistical Decision Theory*, Harvard University Press, Boston.

Robbins, H. (1956). An empirical Bayes approach to statistics, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pp. 157–163.

Speed, T. P. (ed.) (2003). *Statistical Analysis of Gene Expression Microarray Data*, Chapman and Hall, London.

Storey, J. D. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society, Series B* **64**: 479–498.

Storey, J. S. and Tibshirani, R. (2003). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays, *The analysis of gene expression data: methods and software*, Springer, New York.

Tusher, V., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Science, USA* **98**: 5116–5121.

Wolfinger, R. D., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models, *Journal of Computational Biology* **8**: 625–637.

Model		II			CI			IC			CC		
top %		1	2	10	1	2	10	1	2	10	1	2	10
MM	S/N	CI.T	CI.T	CI.T	CI.T	CI.T	CI.T	IC.TP CI.TP	IC.TP CI.TP	IC.TP CI.TP	IC.TP CI.TP	IC.TP CI.TP	IC.TP CI.TP
	Signal	CI.F *C.F	CI.F *C.F	CI.TP IC.TP	CI.F *C.F	CI.F *C.F	IC.TP	*C.TP	*C.TP	IC.TP	*C.TP	*C.TP	*C.TP
TT	S/N	CI.T	CI.T T	CI.T T	CI.T T	CI.T T	CI.T T	IC.BF *I.TP	IC.BF	IC.BF CI.TP	IC.BF CI.BF II.TP	IC.BF CI.BF CI.TP	IC.BF CI.BF CI.TP
	Signal	CI.F	CI.F	*I.TP IC.TP	CI.F	*I.F	*I.TP IC.TP	CC.TP	CC.TP	CC.TP	CC.TP	CC.TP	CC.TP

Table 4: Best statistics for each simulation scenario with three replicates. Here "model" corresponds to the true model used for simulation. Statistics were differentiated in their ability to identify the top 1%, 2% and 10% of genes with large signals (labeled Signal) and large signal-to-noise ratios (S/N). The rows labeled MM correspond to parameter estimation using the method of moments. Results using the actual true parameter values (labeled TT) are also given. Instances where the best statistic did not matching the appropriate true model and goal are highlighted in red. Instances where the best statistic did not match the model but was consistent with goal are highlighted in blue. The results highlight the importance of matching the statistic to the goal; in only one case did a T statistic out-perform other statistics to identify high signal changes while in no cases did a fold change out-perform other statistics to identify high SNRs.

Model		II			CI			IC			CC		
top %		1	2	10	1	2	10	1	2	10	1	2	10
MM	S/N	*I.T	*I.T	II.T	CI.T	CI.T	CI.T	IC.BF *I.TP	IC.BF *I.TP	IC.BF CI.TP	*I.TP	IC.BF *I.TP	IC.BF *I.TP
	Signal	*I.TP CI.F	*I.TP CI.F	CI.F CI.TP IC.TP	II.TP IC.TP	CI.F II.TP IC.TP	*I.TP *.F IC.TP	*C.TP	*C.TP	*C.TP	*C.TP	CC.TP	*C.TP
TT	S/N	*I.T	*I.T	II.T	CI.T	CI.T	CI.T	IC.BF *I.TP	IC.BF *I.TP	IC.BF CI.TP	IC.BF *I.TP	IC.BF *I.TP CI.BF	IC.BF CI.TP
	Signal	*I.TP CI.F	*I.TP CI.F	CI.F CI.TP	II.TP IC.TP	*I.TP	*I.TP *.F IC.TP	CC.TP	CC.TP	CC.TP	CC.TP	CC.TP	CC.TP

Table 5: Best statistics for each simulation scenario with ten replicates. Here "model" corresponds to the true model used for simulation. Statistics were differentiated in their ability to identify the top 1%, 2% and 10% of genes with large signals (labeled Signal) and large signal-to-noise ratios (S/N). The rows labeled MM correspond to parameter estimation using the method of moments. Results using the actual true parameter values (labeled TT) are also given. Instances where the best statistic did not matching the appropriate true model and goal are highlighted in red. Instances where the best statistic did not match the model but was consistent with goal are highlighted in blue. The results highlight the importance of matching the statistic to the goal; in no cases did a T statistic out-perform other statistics to identify high signal changes nor did a fold change out-perform other statistics to identify high SNRs. In most cases statistics based on the appropriate conjugacy assumptions outperformed the other competitors.

Model		II			CI			IC			CC		
top %		1	2	10	1	2	10	1	2	10	1	2	10
MM	S/N	II.T	II.T IC.BF	II.T	CI.T IC.BF	CI.T IC.BF	CI.T	IC.BF	IC.BF	IC.BF	IC.BF *I.BF	IC.BF	IC.BF
	Signal	*I.F *I.TP	*I.F *I.TP	*I.F *I.TP	*I.TP	*I.TP *.F	*I.TP *.F	IC.TP	IC.TP	IC.TP	IC.TP	IC.TP	IC.TP
TT	S/N	II.T	II.T IC.BF	II.T	CI.T IC.BF	CI.T IC.BF	CI.T	IC.BF	IC.BF	IC.BF	IC.BF *I.BF	IC.BF *I.BF	IC.BF
	Signal	*I.F *I.TP	*I.F *I.TP	*I.F *I.TP	*I.TP	*I.TP *.F	*I.TP *.F	IC.TP	IC.TP	IC.TP	IC.TP	IC.TP	IC.TP

Table 6: Best statistics for each simulation scenario with one hundred replicates. Here "model" corresponds to the true model used for simulation. Statistics were differentiated in their ability to identify the top 1%, 2% and 10% of genes with large signals (labeled Signal) and large signal-to-noise ratios (S/N). The rows labeled MM correspond to parameter estimation using the method of moments. Results using the actual true parameter values (labeled TT) are also given. Instances where the best statistic did not matching the appropriate true model and goal are highlighted in red. Instances where the best statistic did not match the model but was consistent with goal are highlighted in blue. The results highlight the importance of matching the statistic to the goal; in no cases did a T statistic out-perform other statistics to identify high signal changes nor did a fold change out-perform other statistics to identify high SNRs. In most cases statistics based on the appropriate conjugacy assumptions outperformed the other competitors.

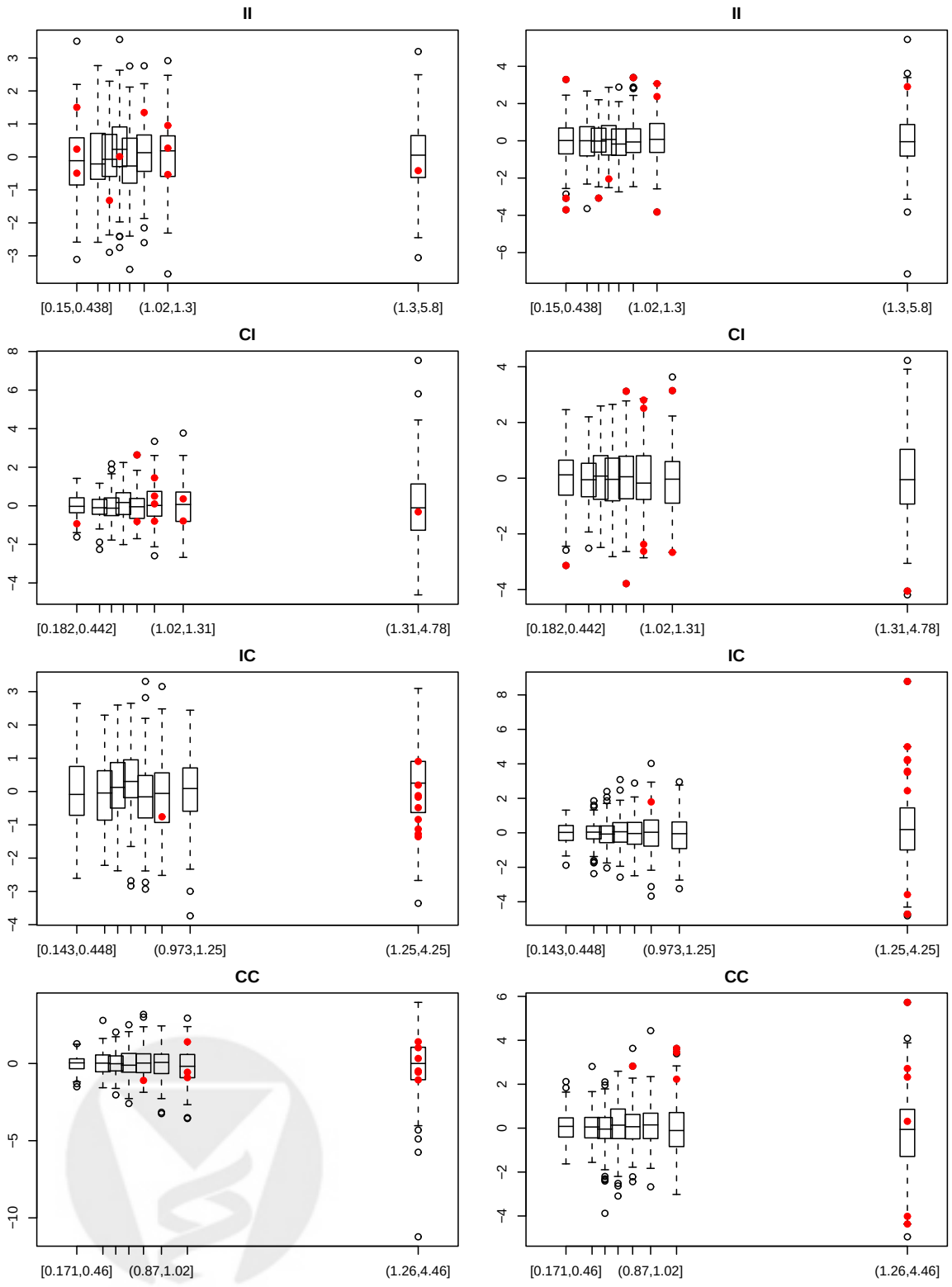

Figure 1: Diagnostic plots for data simulated from the four models. Red points are the genes with the largest fold changes. The hyper parameter values used to simulate the data were $\nu = 2$, $\beta = 1$, $\gamma = 1$ and $\tau = 1$. Each simulation included 5 replicates and 1,000 genes. The plots on the left depict the abundance versus the noise while the plots on the right depict the signal. Here the conjugate versus independence relationship is very apparent in each of the plots.

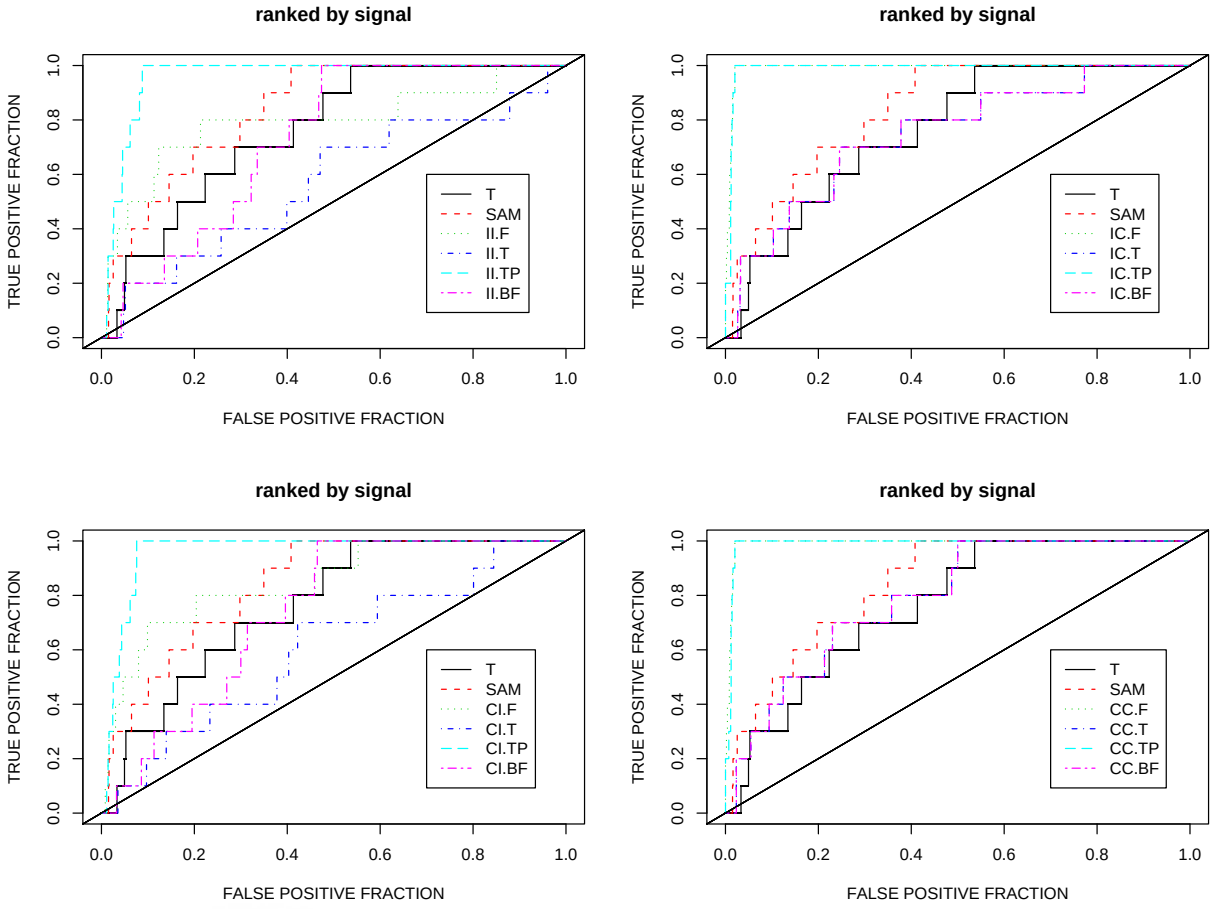

Figure 2: ROC curves for all of the statistics under study for a single simulation with 5 replicates and 1000 genes. The hyperparameter values used to simulate the data were $\nu = 2$, $\beta = 1$, $\gamma = 1$ and $\tau = 1$. The ROC curves are grouped by statistical model. The areas under the curve for each statistic represent the values used in subsequent plots.

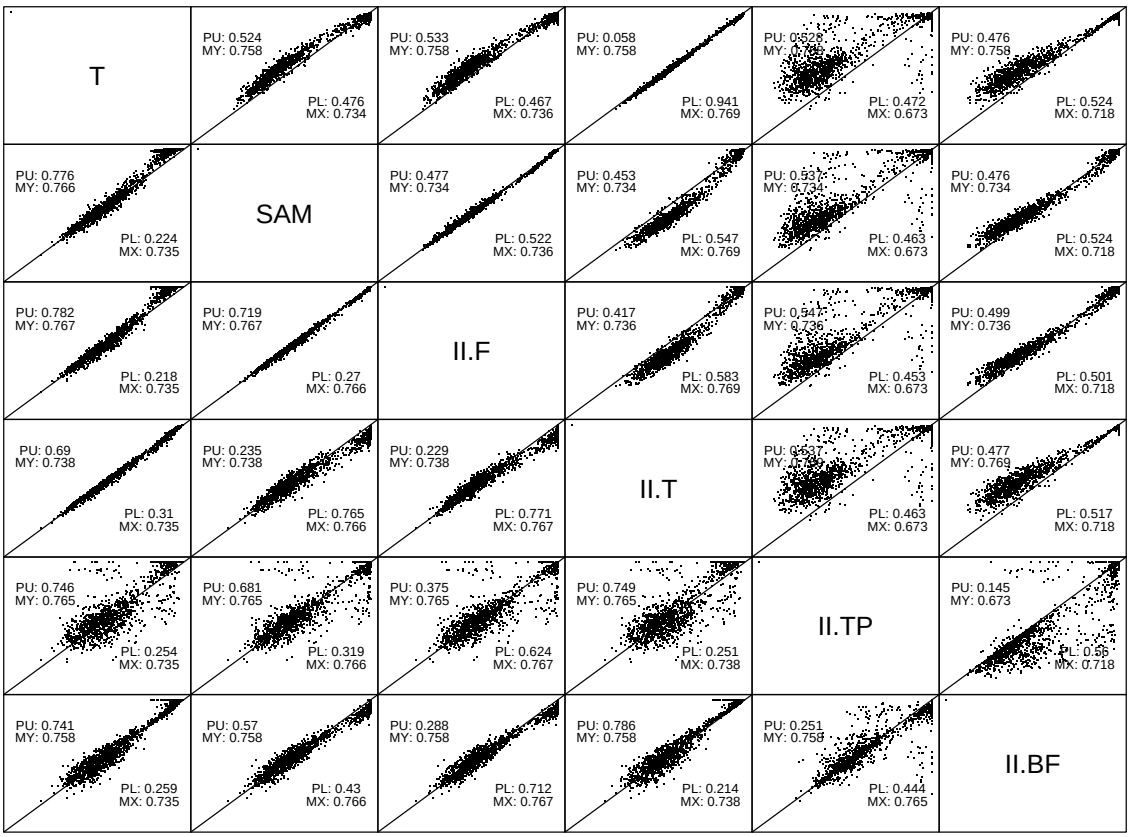

Figure 3: Comparison of performance in II data for T, SAM and the II statistics. Each point represents the paired areas under the curve for two statistics for a simulated data set. Plots above the diagonal base the ROC curves on identifying genes with large reliably measured changes while plots below the diagonal are based on identifying large signal genes. Results for the standard T statistic, SAM and the statistics based on the independence model are given. The summary statistics PU and PL are the percentage of points lying above and below the diagonal respectively while MX and MY are the average of the averages for the horizontal and vertical variables respectively. Here the II.BF statistic is the winner for identifying reliably measured genes while the II.F statistic apparently is preferable for identifying large signal changes.

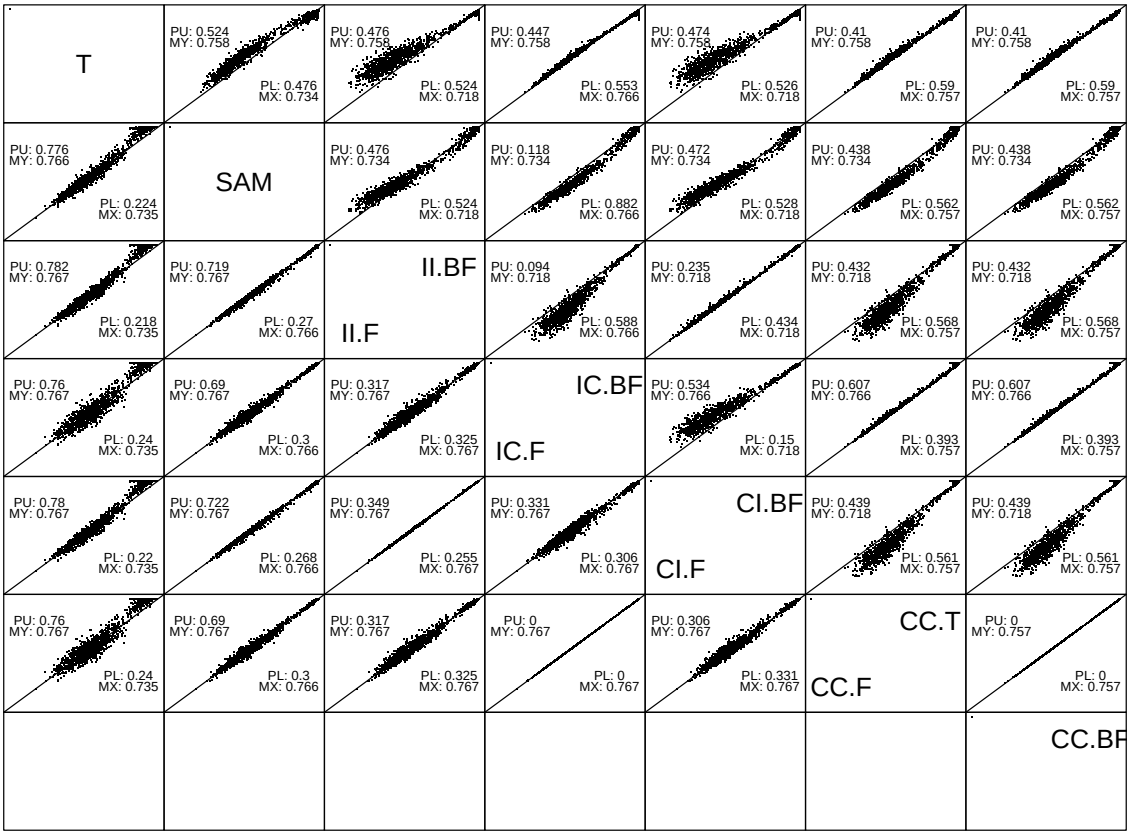

Figure 4: Comparison of performance in II data for T, SAM and the best performing statistics by model class. Each point represents the paired areas under the curve for two statistics for a simulated data set. Plots above the diagonal base the ROC curves on identifying genes with large reliably measured changes while plots below the diagonal are based on identifying large signal genes. Included are results for the T statistic, SAM and overall best performing statistic motivated by any of the multilevel models. Note that as the statistics performed differently across the two goals, some diagonal cells cells are labeled with two statistics. The summary statistics PU and PL are the percentage of points lying above and below the diagonal respectively while MX and MY are the average of the horizontal and vertical variable respectively. Note that the CC.T statistic invokes the same ranking of genes as the CC.BF statistic. ROC results of II and CI statistics are nearly identical and so are results of CC and CI models.

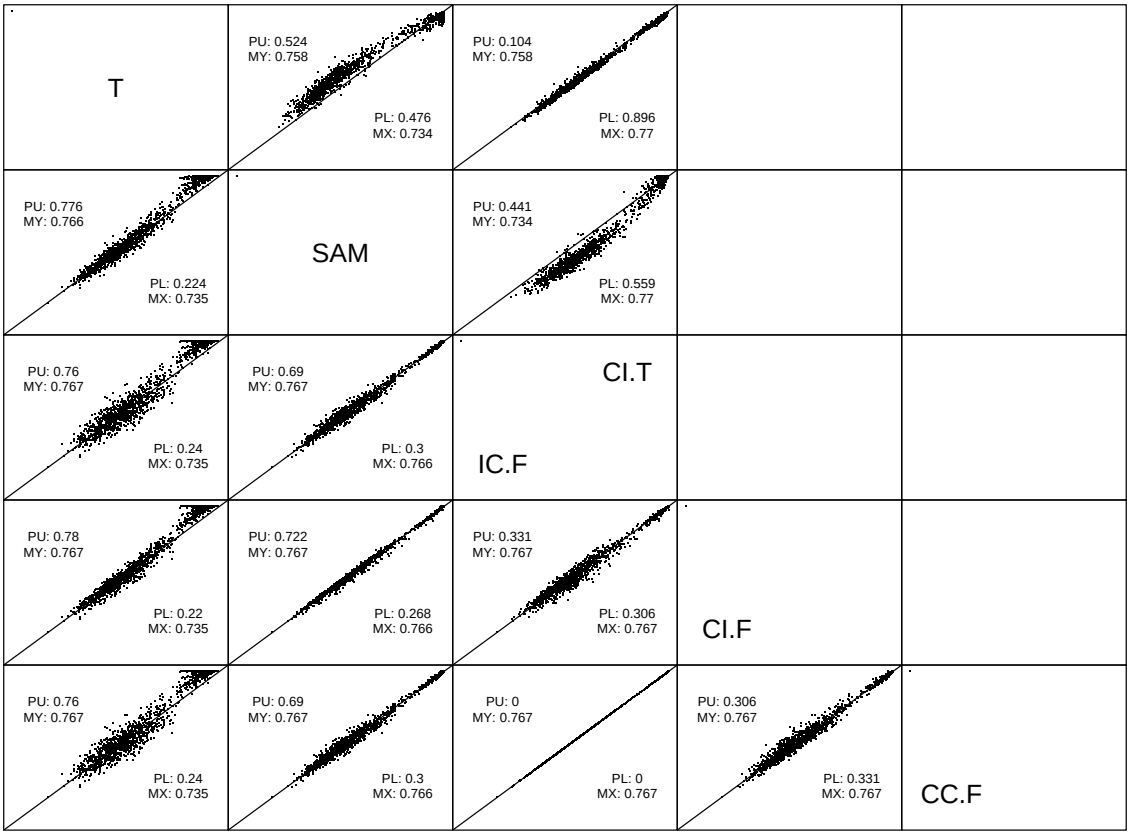

Figure 5: Comparison of performance in II data for T, SAM and the overall best performing statistics. Each point represents the paired areas under the curve for two statistics for a simulated data set. Plots above the diagonal base the ROC curves on identifying genes with large reliably measured changes while plots below the diagonal are based on identifying large signal genes. Included are results for the T statistic, SAM and best performing statistic motivated from the multilevel models. Note that as the statistics performed differently across the two goals, some diagonal cells cells are labeled with two statistics. The summary statistics PU and PL are the percentage of points lying above and below the diagonal respectively while MX and MY are the average of the horizontal and vertical variable respectively.

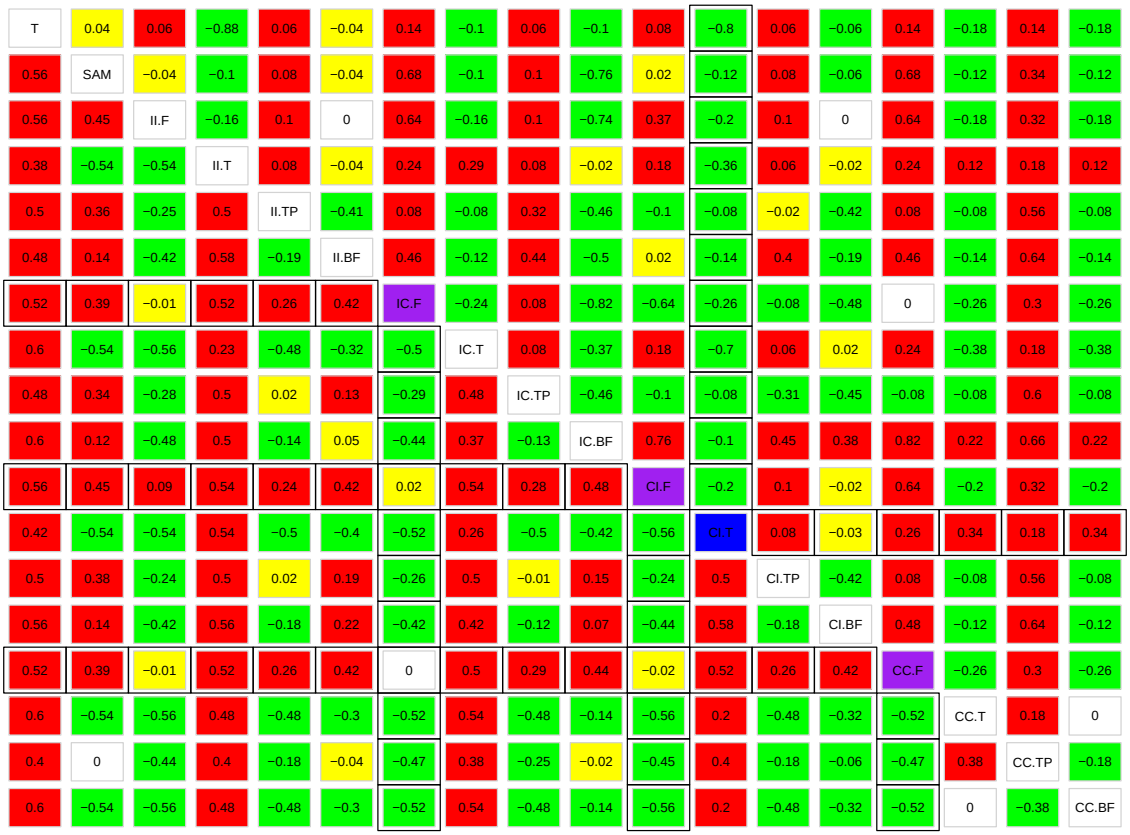

Figure 6: Heat map of the result for data simulated from the Complete Independence model with three replicates. This plot picks the difference between PU and PL for all of the statistics and labeled with colors. Green indicates PU<PL, red indicates PU>PL, while yellow indicates those cases where the difference was negligible, less than 0.05. The best statistics to identify genes with large signal are labeled with purple, while the best statistics to identify reliably measured differentially expressed genes are labeled with blue. The results are summarized in Table 4.

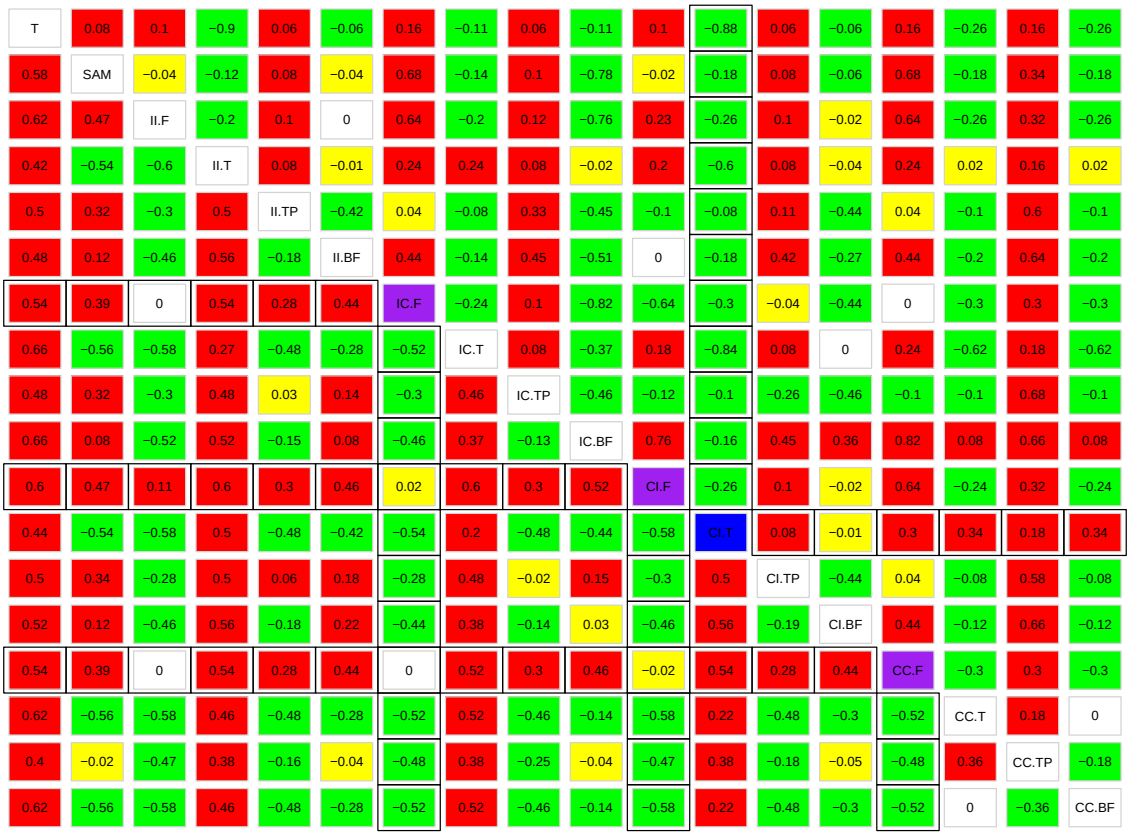

Figure 7: Heat map of the result for data simulated from the Independence of Signal and Noise model with three replicates. This plot picks the difference between PU and PL for all of the statistics and labeled with colors. Green indicates PU<PL, red indicates PU>PL, while yellow indicates those cases where the difference was negligible, less than 0.05. The best statistics to identify genes with large signal are labeled with purple, while the best statistics to identify reliably measured differentially expressed genes are labeled with blue. The results are summarized in Table 4.

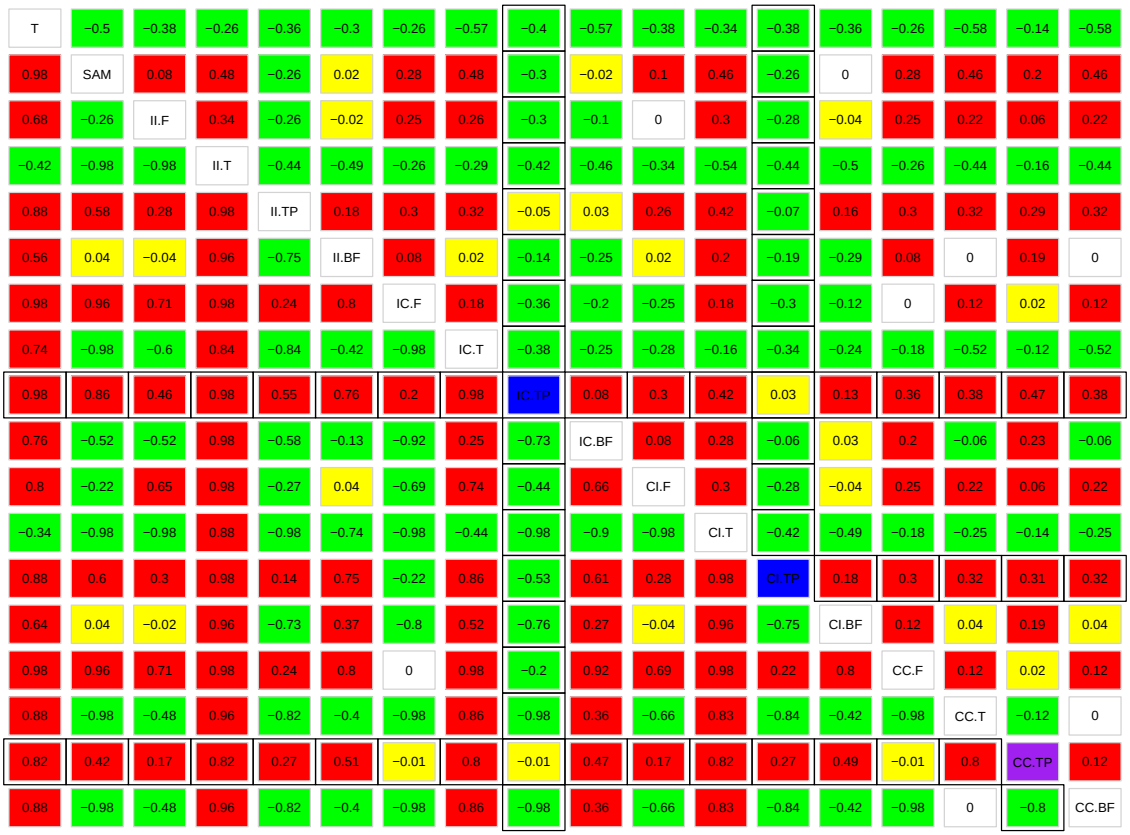

Figure 8: Heat map of the result for data simulated from the Independence of Abundance and Noise model with three replicates. This plot picks the difference between PU and PL for all of the statistics and labeled with colors. Green indicates PU<PL, red indicates PU>PL, while yellow indicates those cases where the difference was negligible, less than 0.05. The best statistics to identify genes with large signal are labeled with purple, while the best statistics to identify reliably measured differentially expressed genes are labeled with blue. The results are summarized in Table 4.

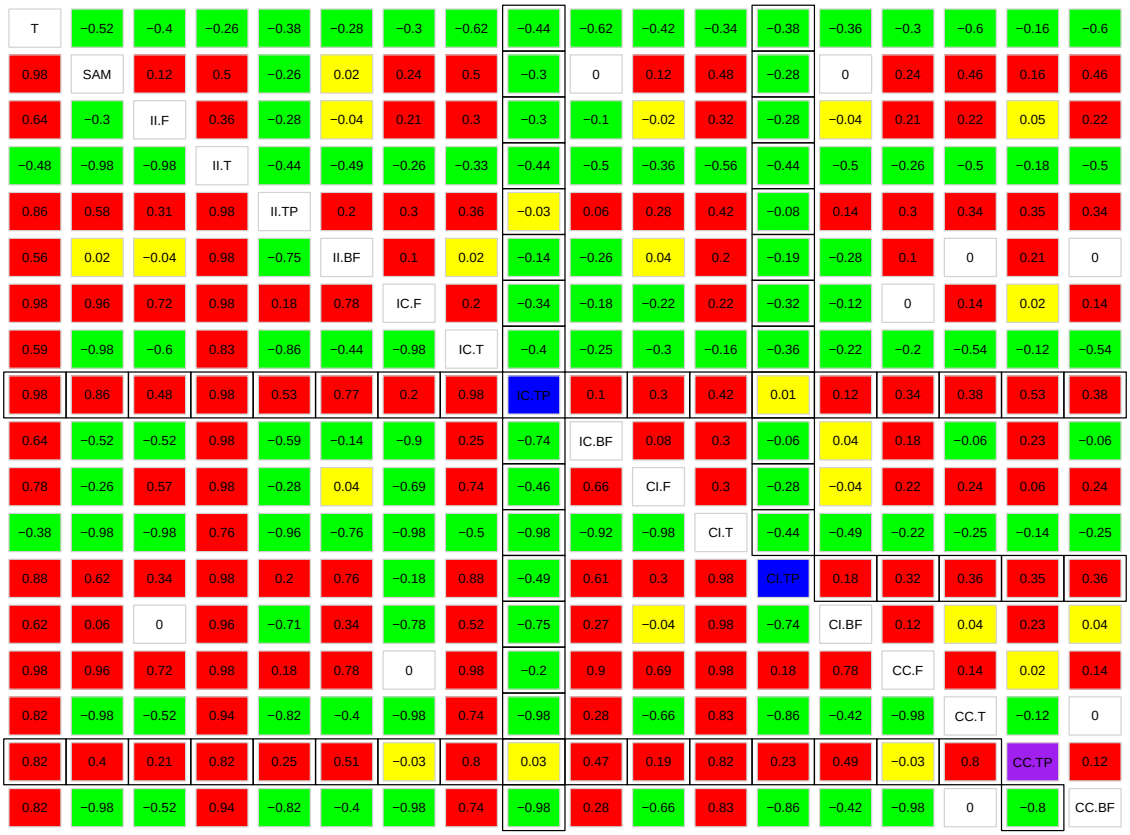

Figure 9: Heat map of the result for data simulated from the Complete Conjugacy model with three replicates. This plot picks the difference between PU and PL for all of the statistics and labeled with colors. Green indicates PU<PL, red indicates PU>PL, while yellow indicates those cases where the difference was negligible, less than 0.05. The best statistics to identify genes with large signal are labeled with purple, while the best statistics to identify reliably measured differentially expressed genes are labeled with blue. The results are summarized in Table 4.

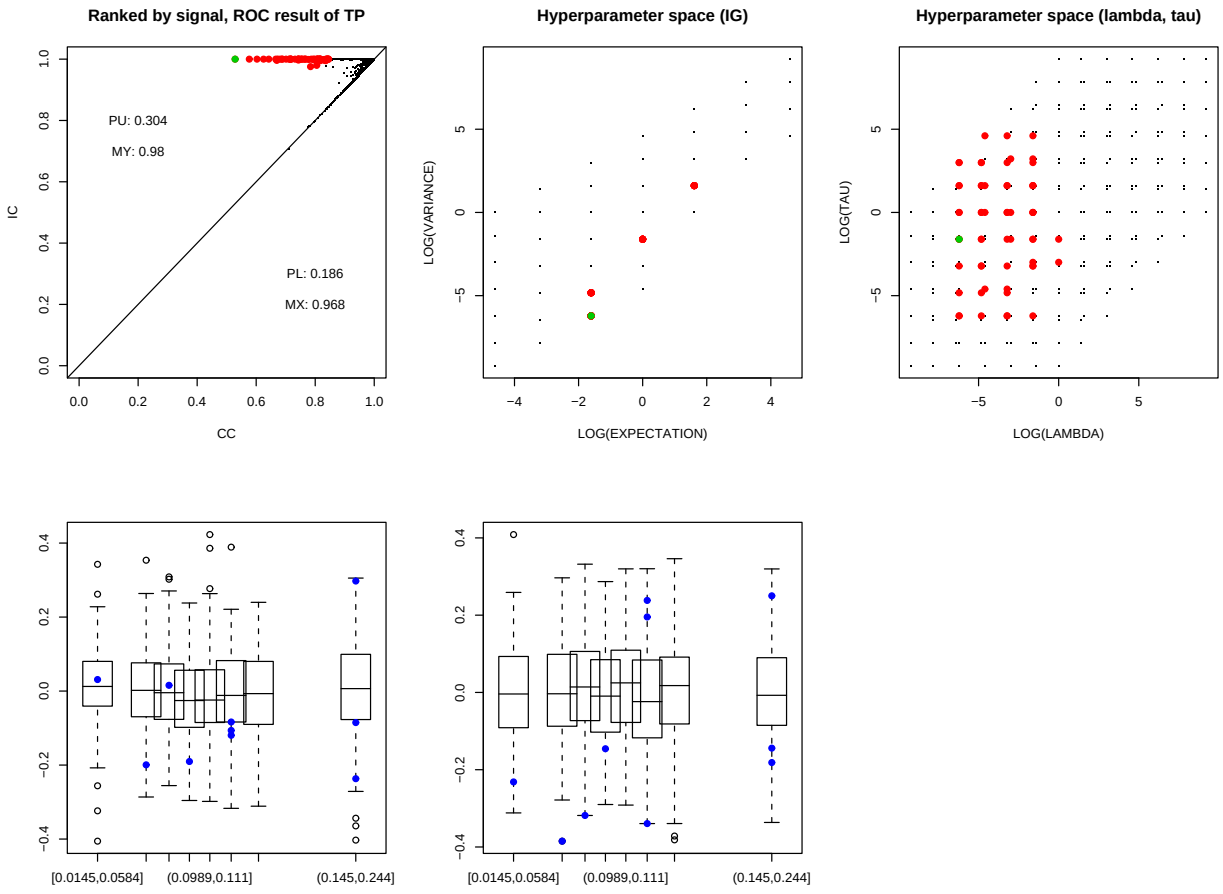

Figure 10: Further investigation of the counterintuitive result that IC.TP outperforms CC.TP on data simulated from the Complete Conjugacy model. The first graph is the scatter plot of IC.TP versus CC.TP. Cases where IC.TP does significantly better CC.TP are highlighted in red. The next two plots depict the hyperparameter space for the mean and variance of the inverse gamma prior and λ and τ respectively. In the first plot, note that all of the discrepant cases fall in 4 small neighborhoods. The diagnostic plot of the most discrepant point, labeled in green, are given in the final two plots. Note that the relationship strongly support the Complete Independence despite the fact that the data were simulated from Complete Conjugacy model.

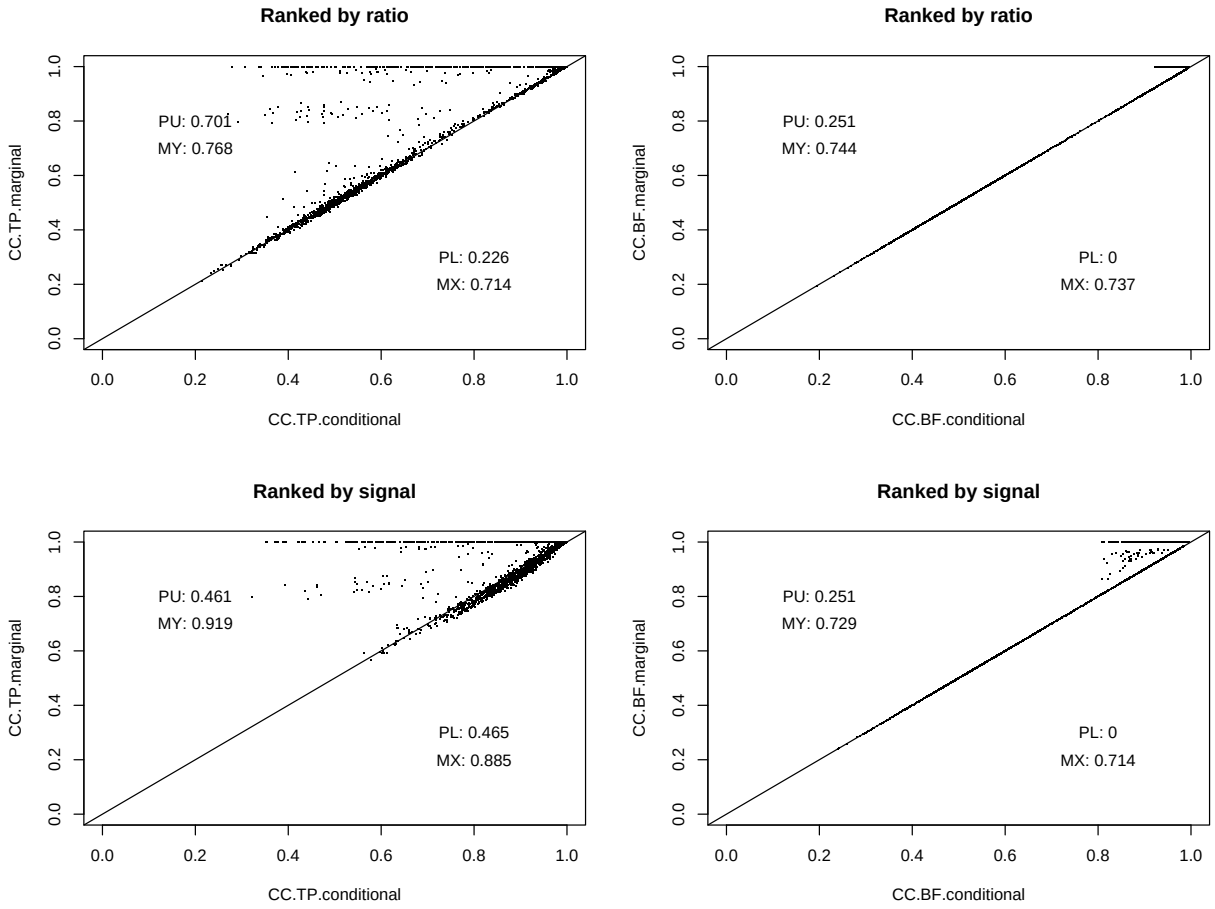

Figure 11: For data simulated from Complete Conjugacy model, a comparison of the performance of the CC.TP and CC.BF statistics calculated exactly by integrating over σ_g^2 (marginal) versus the real time approximation obtained by plugging in the posterior mode of σ_g^2 (conditional). In the bottom row, genes were ranked by signal change, while in the top row, genes were ranked by signal to noise ratio. Encouragingly, the real time statistics performs similarly to their exact counterparts.

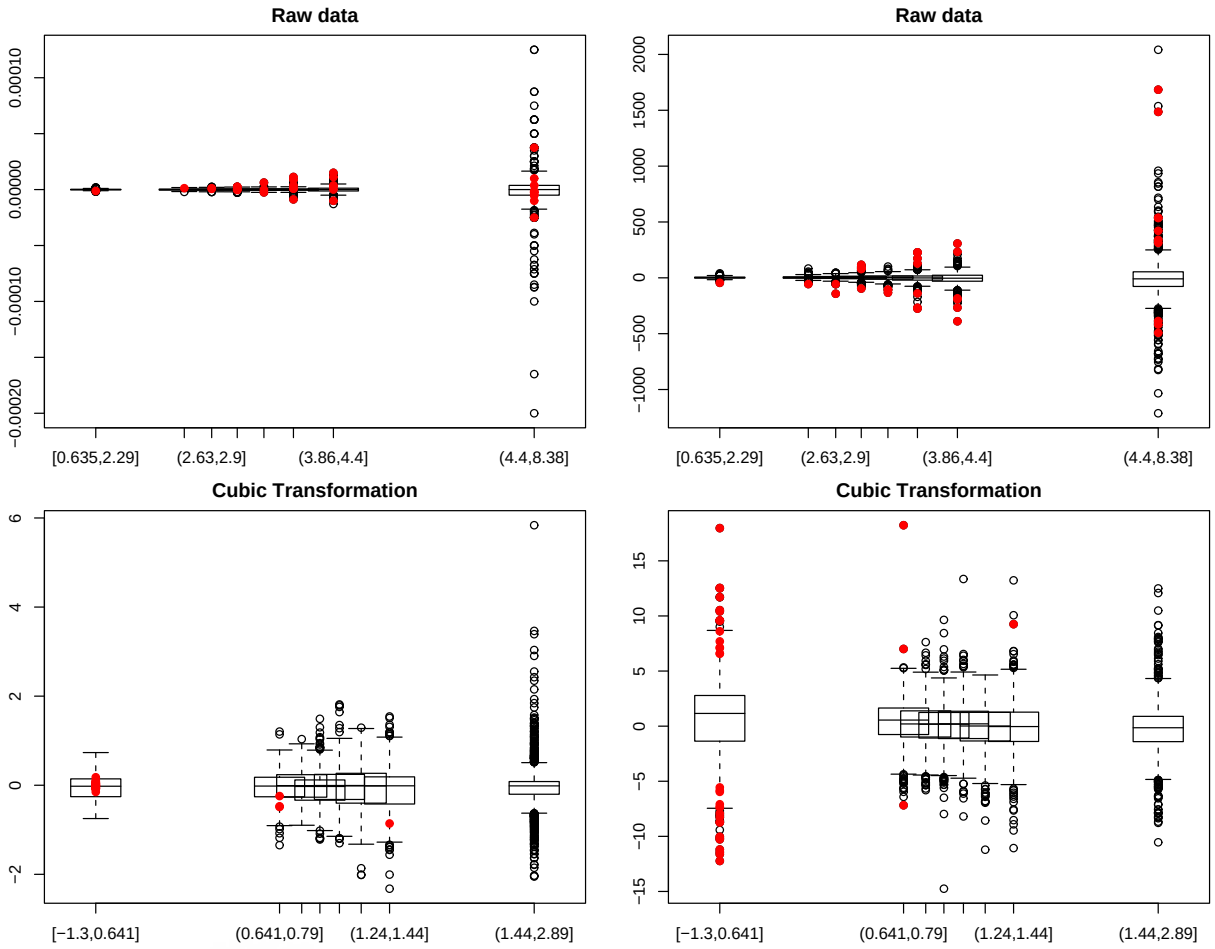

Figure 12: Signal to noise plot for the Tusher two class data under two transformations. Genes identified by SAM are highlighted in red. The horizontal axis is the binned gene specific variances while the vertical axis for the plots in the left column is the average expression across the two groups (abundance) while it is the the average difference in expression (signal) for the right column. Conjugacy appears appropriate for the raw data, and independence relationship appears more appropriate for the cubic transformed data.

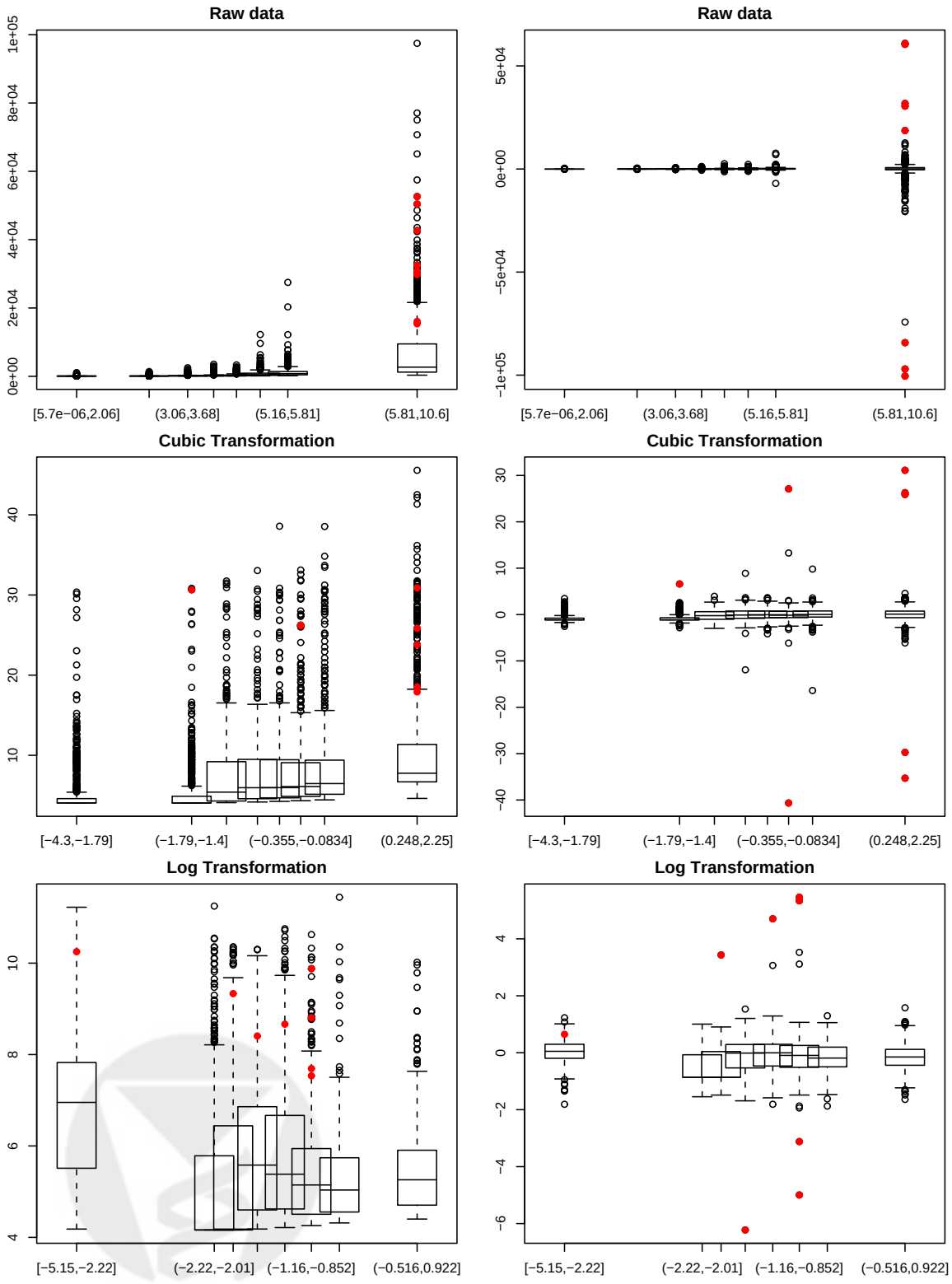

Figure 13: Signal to noise plot for the spike-in data under three transformations. Spiked-in genes are highlighted in red. The horizontal axis is the binned gene specific variances while the vertical axis for the plots in the left column is the average expression across the two groups (abundance) while it is the the average difference in expression (signal) for the right column. The apparent relationships between abundance and noise and signal and noise clearly change dependent on the transformation used. While conjugacy appears appropriate for the raw data, and independence relationship appears more appropriate for the log-transformed data.

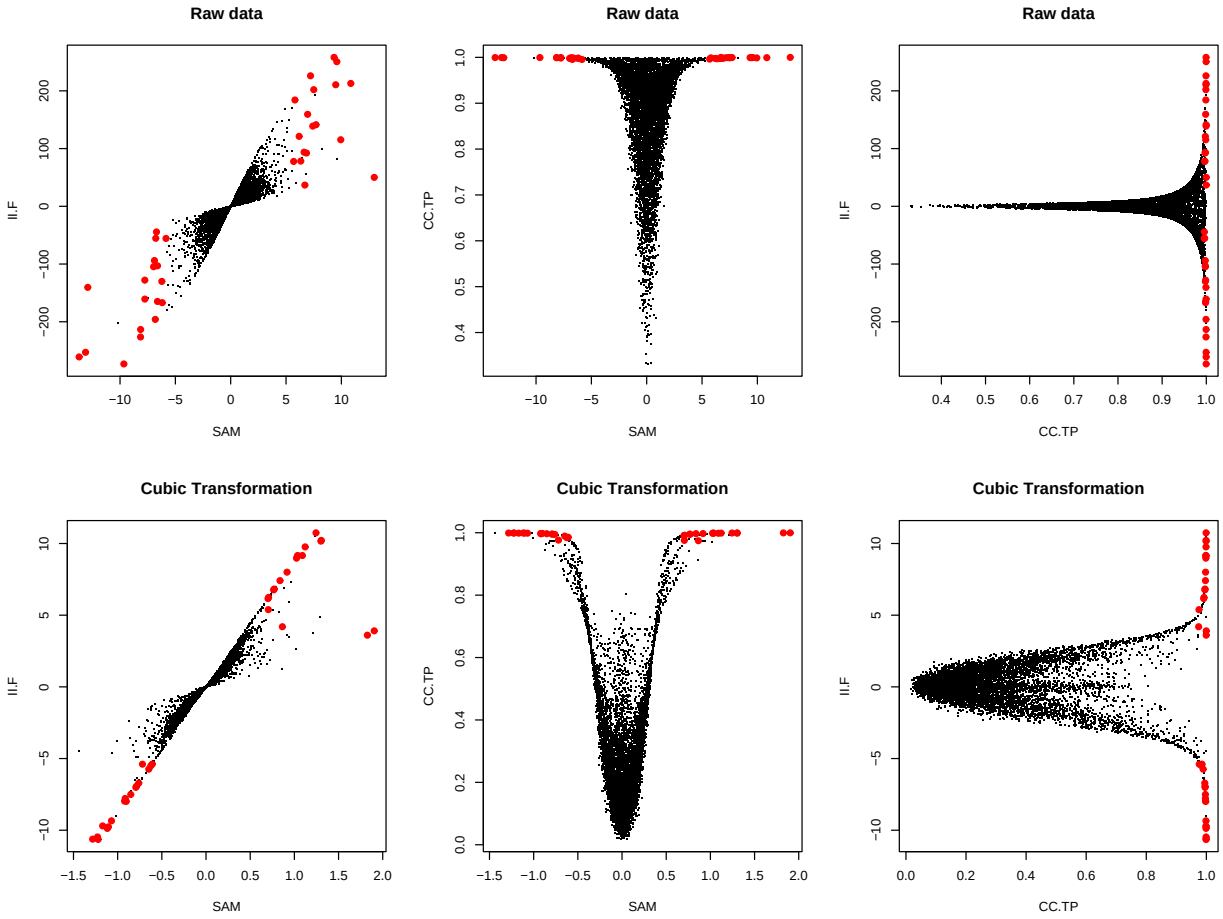

Figure 14: Comparisons of the II.F, CC.TP and SAM statistics on the Tusher data. Genes selected by SAM are highlighted in red for reference. Here the CC.F and II.TP statistics were chosen from Tables 4,5 and 6 as the optimal statistics for detecting large signal changes for the II and CC models. For this data set, the II model is supported under the cubic transformations while the CC model is supported for the raw data. A zoomed-in view of the comparison between CC.TP and SAM the raw data case is in Figure 19.

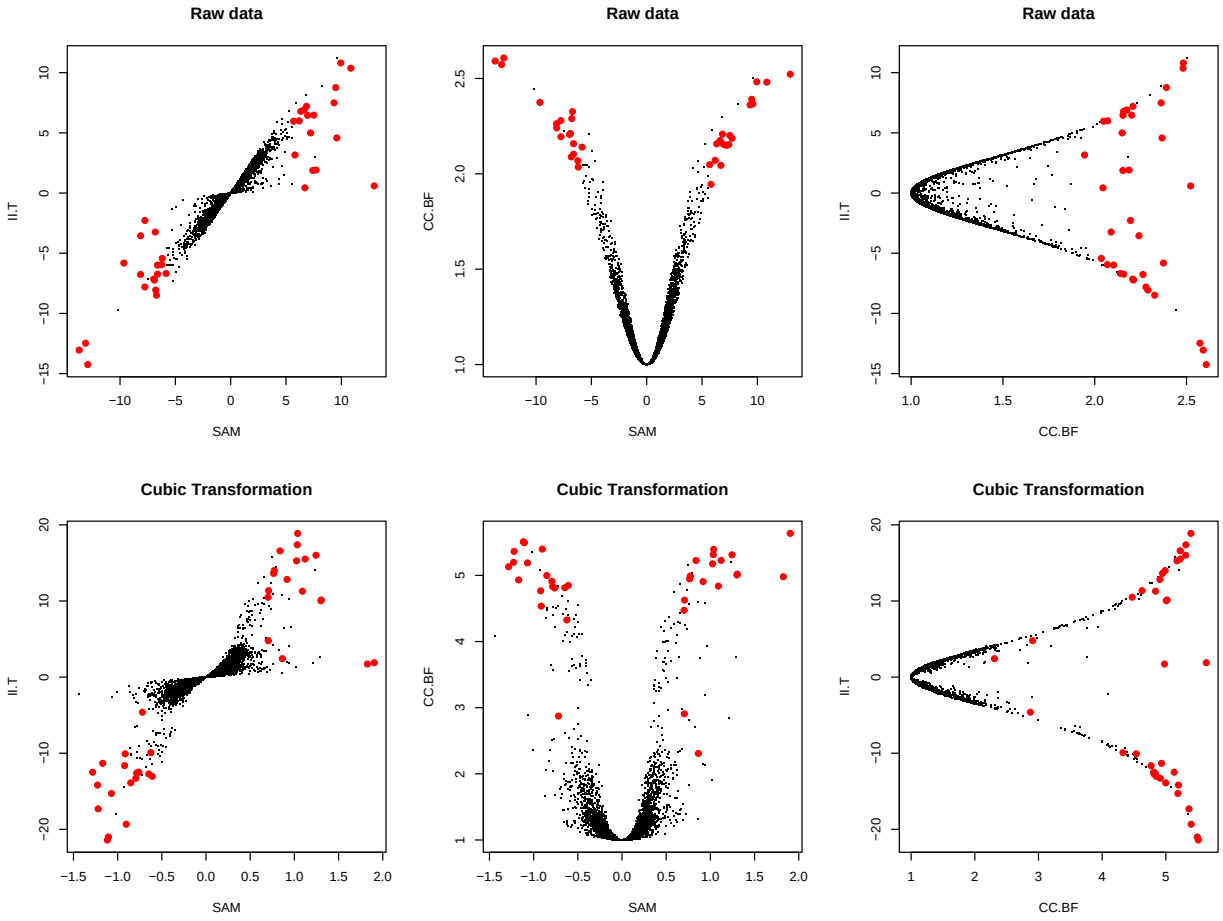

Figure 15: Comparisons of the II.T, CC.BF and SAM statistics on the Tusher data. Genes selected by SAM are highlighted in red. Here the CC.BF and II.T statistics were chosen from Tables 4,5 and 6 as the optimal statistics for detecting large signal to noise ratio changes for the II and CC models. For this data, the II model is supported under the log transformations while the CC model is supported for the raw and cubic transformed data.

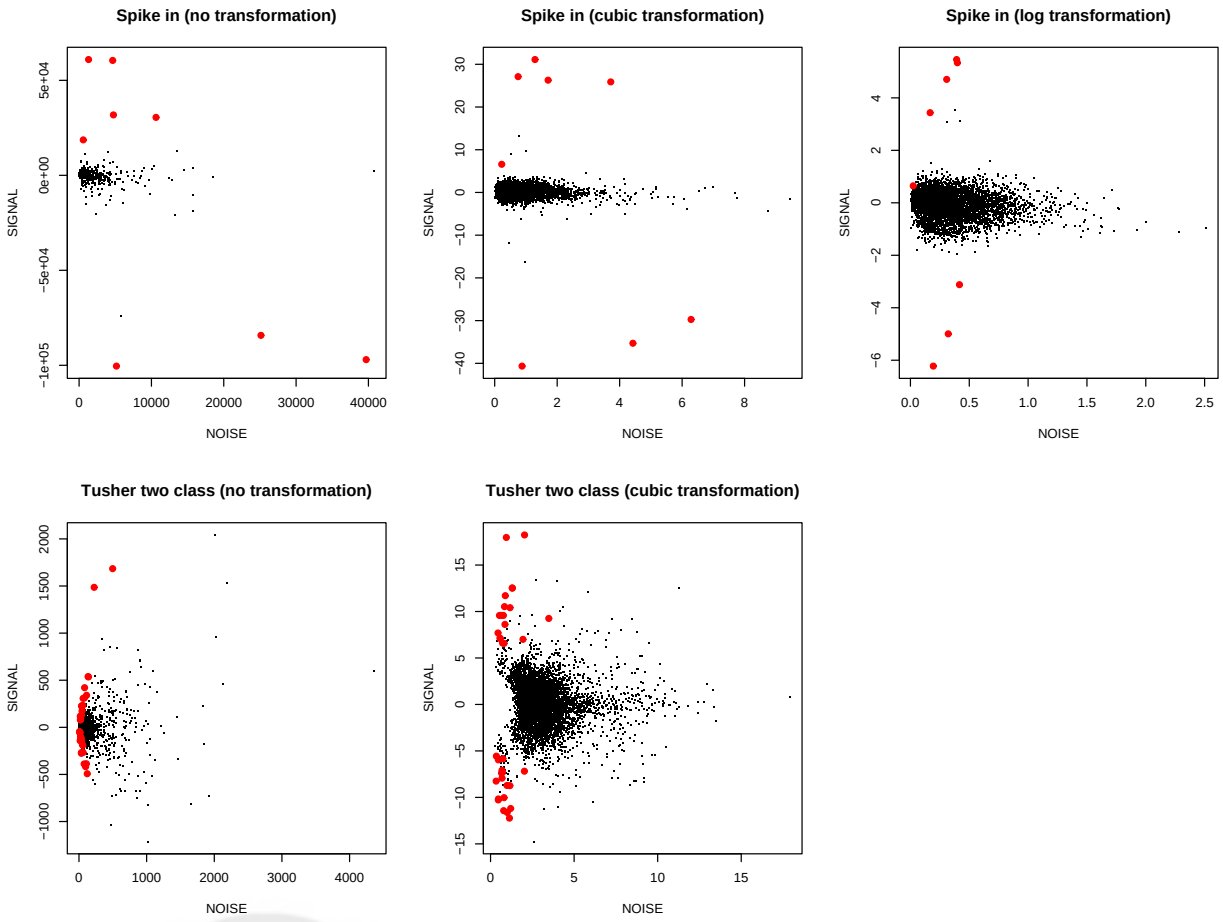

Figure 16: Scatterplot of the gene specific average difference in expression (signal) versus the gene specific variances for the spike-in and Tusher data under various transformations of the signal. The red highlighted genes correspond to the spiked in genes and the ones identified by SAM respectively.

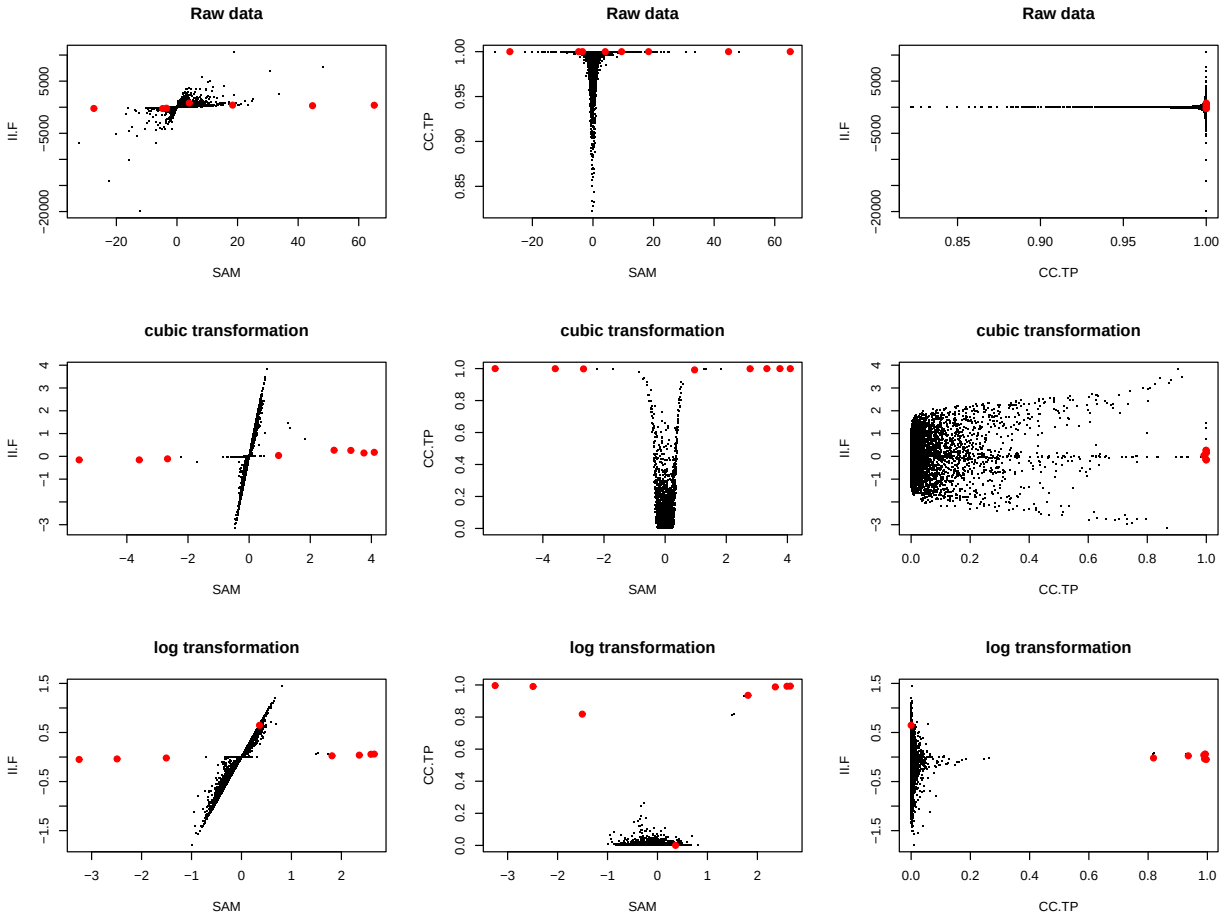

Figure 17: Comparisons of fold change statistics II.F and CC.TP with SAM in the spike-in data. Spiked-in genes are highlighted in red. Here the CC.F and II.TP statistics were chosen from Tables 4,5 and 6 as the optimal statistics for detecting large signal changes for the II and CC models. For this data, the II model is supported under the log transformations while the CC model is supported for the raw and cubic transformed data. The II.F statistic shrinks conservatively even for the log transformed data while SAM and CC.TP perform well for both transformed data sets. No statistic performs well on the original scale. A zoomed-in view of the comparison between CC.TP and SAM the raw data case is in Figure 19.

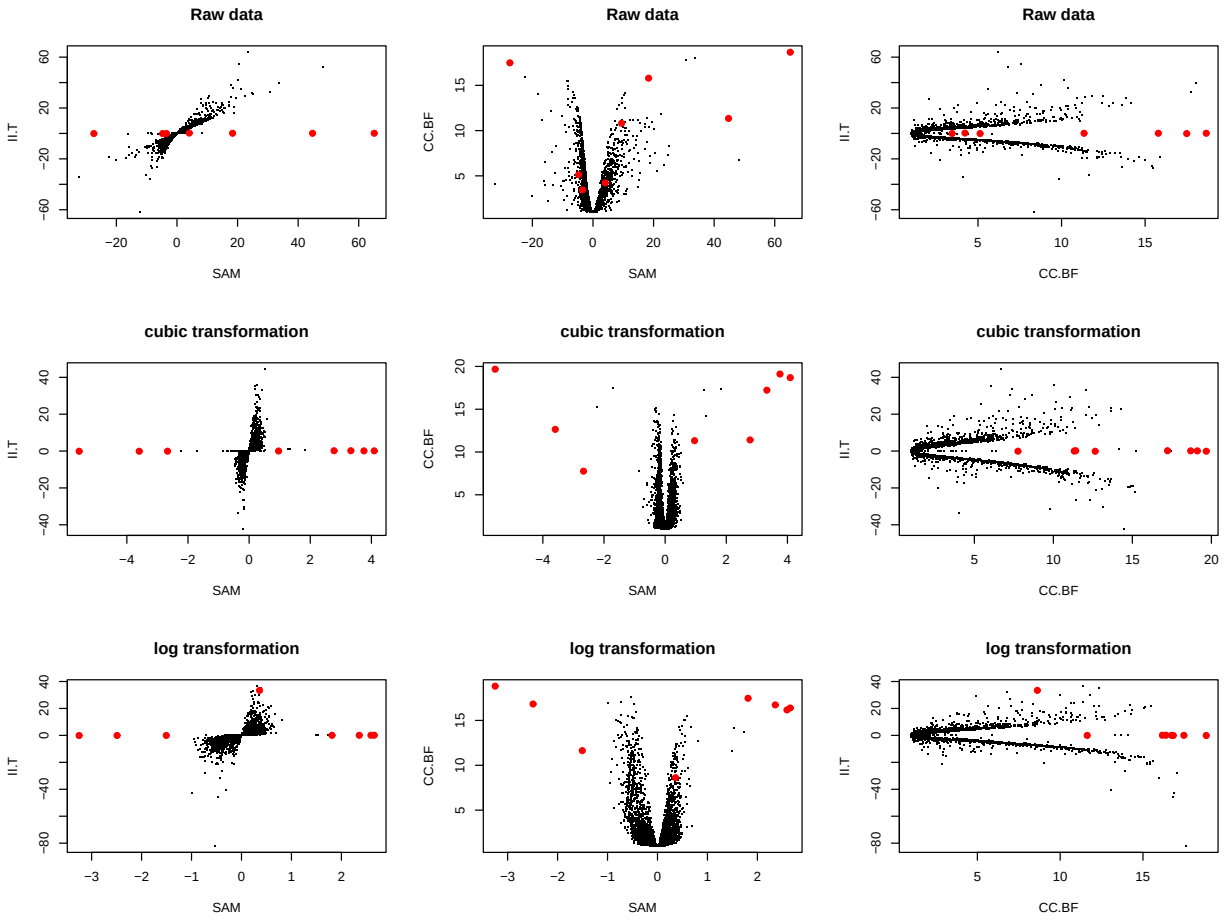

Figure 18: Comparisons of signal-to-noise statistics II.T and CC.BF with SAM in the spike-in data. Spiked-in genes are highlighted in red. Here the CC.BF and II.T statistics were chosen from Tables 4,5 and 6 as the optimal statistics for detecting large signal to noise ratio changes for the II and CC models. For this data, the II model is supported under the log transformation while the CC model is supported for the raw and cubic transformed data. The II.T statistic shrinks conservatively even for the log transformed data, while SAM and CC.BF perform well for both transformed data sets. No statistic performs well on the original scale.

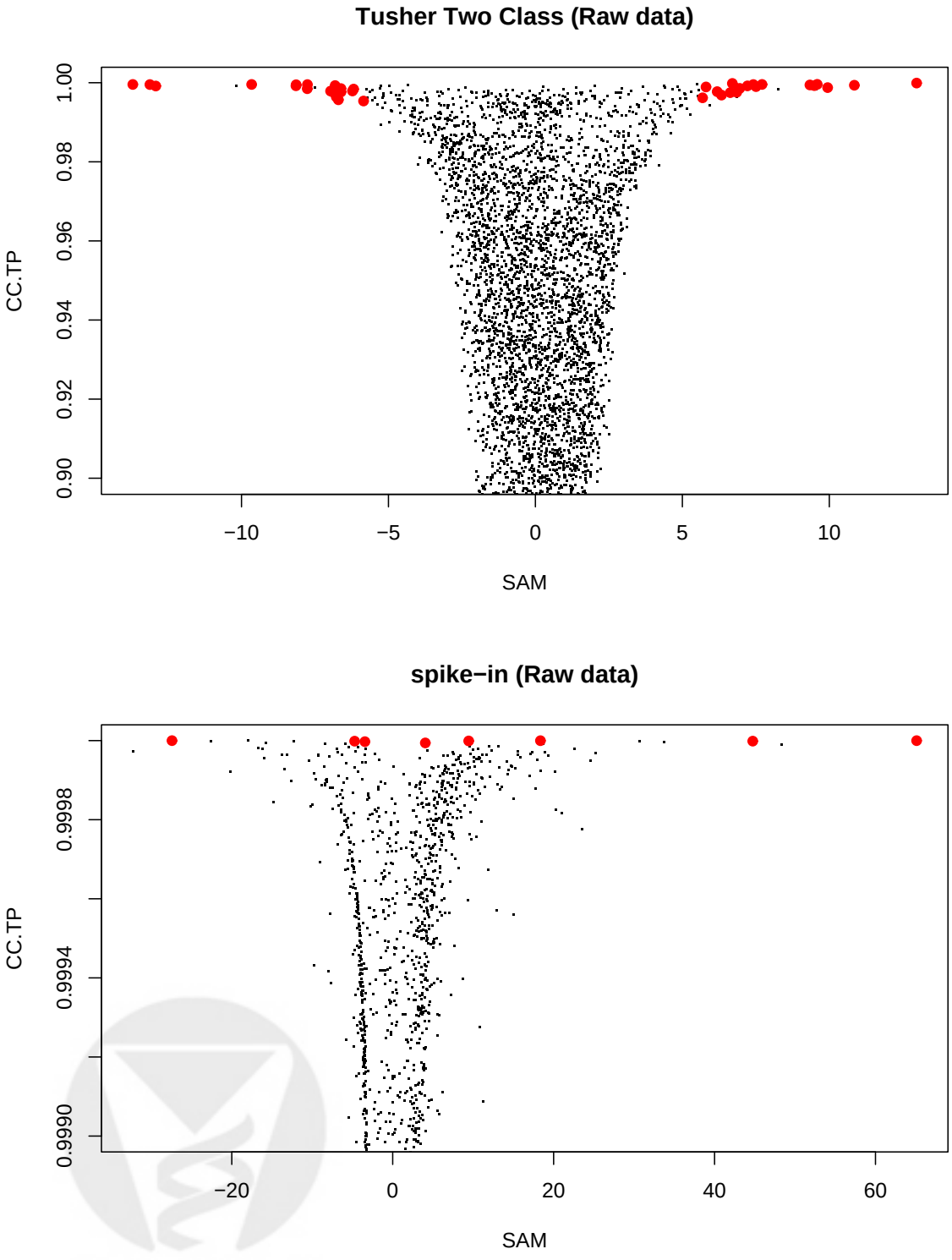

Figure 19: Zoomed-in view of the comparison between CC.TP and SAM the raw data case in Figure 14 and Figure 17.

Define

$$\begin{aligned}
\hat{\xi} &= (\nu, \beta, \lambda, \tau) \\
d_g &= \bar{X}_{2g} - \bar{X}_{1g} \\
a_g &= \frac{1}{2n}\sum_{i=1}^{2}\sum_{j=1}^{n} X_{igj} \\
s_g^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(X_{1gj} - \bar{X}_{1g})^2 + \frac{1}{n-1}\sum_{j=1}^{n}(X_{2gj} - \bar{X}_{2g})^2 \\
d_g &\sim N(\delta_g, \frac{2}{n}\sigma_g^2) \\
a_g &\sim N(\mu_g, \frac{1}{2n}\sigma_g^2) \\
\frac{n-1}{\sigma_g^2}s_g^2 &\sim \chi^2_{2(n-1)}
\end{aligned}$$

II: Independence model

$$\begin{aligned}
\mu_g|\tau^2 &\sim N(0, \tau^2) \\
\delta_g|\lambda^2 &\sim N(0, \lambda^2)
\end{aligned}$$

Posterior distribution of σ_g^2

$$\begin{aligned}
f(d_g|\delta_g, \sigma_g^2) f(\delta_g|\hat{\xi}) &= \int \frac{1}{\sqrt{2\pi\frac{2}{n}\sigma_g^2}} e^{-\frac{(d_g-\delta_g)^2}{2\frac{2}{n}\sigma_g^2}} \frac{1}{\sqrt{2\pi\lambda^2}} e^{-\frac{\delta_g^2}{2\lambda^2}} d\delta_g \\
&= \sqrt{\frac{1}{2\pi\left(\frac{2}{n}\sigma_g^2 + \lambda^2\right)}} e^{-\frac{d_g^2}{2\left(\frac{2}{n}\sigma_g^2+\lambda^2\right)}}
\end{aligned}$$

$$\begin{aligned}
f(a_g|\mu_g, \sigma_g^2) f(\mu_g|\hat{\xi}) &= \int \frac{1}{\sqrt{2\pi\frac{1}{2n}\sigma_g^2}} e^{-\frac{(a_g-\mu_g)^2}{2\frac{1}{2n}\sigma_g^2}} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{\mu_g^2}{2\tau^2}} d\mu_g \\
&= \sqrt{\frac{1}{2\pi\left(\frac{\sigma_g^2}{2n} + \tau^2\right)}} e^{-\frac{a_g^2}{2\left(\frac{\sigma_g^2}{2n}+\tau^2\right)}}
\end{aligned}$$

$$\pi(\sigma_g^2|d_g, a_g, s_g^2, \hat{\xi}) \propto \frac{s_g^{2(n-2)}}{\sigma_g^{2(n+\nu)}} e^{-\frac{2\beta+(n-1)s_g^2}{2\sigma_g^2}} \sqrt{\frac{1}{\left(\frac{2}{n}\sigma_g^2 + \lambda^2\right)}} e^{-\frac{d_g^2}{2\left(\frac{2}{n}\sigma_g^2+\lambda^2\right)}} \sqrt{\frac{1}{\left(\frac{\sigma_g^2}{2n} + \tau^2\right)}} e^{-\frac{a_g^2}{2\left(\frac{\sigma_g^2}{2n}+\tau^2\right)}}$$

Define the point esitmate of σ_g^2 to be the posterior mode. For all the summary statistics, plug in this point estimation of σ_g^2.

The posterior distribution of δ_g conditional on σ_g^2 is

$$\delta_g|d_g,\sigma_g^2 \sim N\left(\frac{\frac{nd_g}{2\sigma_g^2}}{\frac{n}{2\sigma_g^2}+\frac{1}{\lambda^2}},\frac{1}{\frac{n}{2\sigma_g^2}+\frac{1}{\lambda^2}}\right)$$

Define

$$\begin{aligned} II.F &= \frac{\frac{nd_g}{2\sigma_g^2}}{\frac{n}{2\sigma_g^2}+\frac{1}{\lambda^2}} \\ II.T &= CI.F\sqrt{\frac{n}{2\sigma_g^2}+\frac{1}{\lambda^2}} \propto \frac{\frac{nd_g}{2\sigma_g^2}}{\sqrt{\frac{n}{2\sigma_g^2}+\frac{1}{\lambda^2}}} \\ II.TP &= Prob(\delta_g > D|data,\sigma_g^2) \\ &\quad \text{based on the posterior distribution of } \delta_g \text{ conditional on } \sigma_g^2 \end{aligned}$$

The Bayes Factor of Independence model:

$$\begin{aligned} II.BF &= \frac{Pr(\delta_g=0|d_g,a_g,s_g^2,\hat{\xi})}{Pr(\delta_g\neq 0|d_g,a_g,s_g^2,\hat{\xi})} \\ &= \frac{Pr(d_g,a_g,s_g^2|\delta_g=0,\hat{\xi})}{Pr(d_g,a_g,s_g^2|\delta_g\neq 0,\hat{\xi})}\frac{Pr(\delta_g=0|\hat{\xi})}{Pr(\delta_g\neq 0|\hat{\xi})} \\ &\propto \frac{Pr(d_g,a_g,s_g^2|\delta_g=0,\hat{\xi})}{Pr(d_g,a_g,s_g^2|\delta_g\neq 0,\hat{\xi})} \end{aligned}$$

$$Pr(d_g,a_g,s_g^2|\delta_g=0,\hat{\xi}) = \int\int f(d_g|\delta_g=0,\sigma_g^2)f(a_g|\mu_g,\sigma_g^2)f(\mu_g|\hat{\xi})f(s_g^2|\sigma_g^2)f(\sigma_g^2|\hat{\xi})d\mu_g d\sigma_g^2$$

$$Pr(d_g,a_g,s_g^2|\delta_g=0,\hat{\xi}) \propto \int \sqrt{\frac{1}{\left(\frac{\sigma_g^2}{2n}+\tau^2\right)}}e^{-\frac{a_g^2}{2\left(\frac{\sigma_g^2}{2n}+\tau^2\right)}}\frac{s_g^{2(n-2)}}{\sigma_g^{2(n+\nu+\frac{1}{2})}}e^{-\frac{2\beta+(n-1)s_g^2+\frac{n}{2}d_g^2}{2\sigma_g^2}}d\sigma_g^2$$

$$Pr(d_g,a_g,s_g^2|\delta_g\neq 0,\hat{\xi}) = \int\int f(d_g|\delta_g\neq 0,\sigma_g^2)f(\delta_g|\lambda^2)f(a_g|\mu_g,\sigma_g^2)f(\mu_g|\hat{\xi})f(s_g^2|\sigma_g^2)f(\sigma_g^2|\hat{\xi})d\mu_g d\sigma_g^2$$

$$Pr(d_g,a_g,s_g^2|\delta_g\neq 0,\hat{\xi}) \propto \int \sqrt{\frac{1}{\left(\frac{\sigma_g^2}{2n}+\tau^2\right)}}e^{-\frac{a_g^2}{2\left(\frac{\sigma_g^2}{2n}+\tau^2\right)}}\sqrt{\frac{1}{\left(\frac{2}{n}\sigma_g^2+\lambda^2\right)}}e^{-\frac{d_g^2}{2\left(\frac{2}{n}\sigma_g^2+\lambda^2\right)}}\frac{s_g^{2(n-2)}}{\sigma_g^{2(n+\nu)}}e^{-\frac{2\beta+(n-1)s_g^2}{2\sigma_g^2}}d\sigma_g^2$$

The Bayes Factor of Independence model conditional on σ_g^2 is:

$$\begin{aligned} II.BF &\propto \frac{Pr(d_g, a_g, s_g^2|\delta_g = 0, \hat{\xi}, \sigma_g^2)}{Pr(d_g, a_g, s_g^2|\delta_g \neq 0, \hat{\xi}, \sigma_g^2)} \\ &= \sqrt{\frac{\left(\frac{2}{n}\sigma_g^2 + \lambda^2\right)}{\sigma_g^2}} e^{-\frac{\frac{n}{2}d_g^2}{2\sigma_g^2} + \frac{d_g^2}{2\left(\frac{2}{n}\sigma_g^2 + \lambda^2\right)}} \end{aligned}$$

CI: Independence of Signal and Noise

$$\begin{aligned} \mu_g|\tau^2, \sigma_g^2 &\sim N(0, \sigma_g^2\tau^2) \\ \delta_g|\lambda^2 &\sim N(0, \lambda^2) \end{aligned}$$

Posterior distribution of σ_g^2

$$\begin{aligned} f(d_g|\delta_g, \sigma_g^2) f(\delta_g|\lambda^2) &= \sqrt{\frac{1}{2\pi\left(\frac{2}{n}\sigma_g^2 + \lambda^2\right)}} e^{-\frac{d_g^2}{2\left(\frac{2}{n}\sigma_g^2 + \lambda^2\right)}} \\ f(a_g|\mu_g, \sigma_g^2) f(\mu_g|\hat{\xi}) &= \sqrt{\frac{1}{2\pi\left(\frac{1}{2n} + \tau^2\right)\sigma_g^2}} e^{-\frac{a_g^2}{2\left(\frac{1}{2n} + \tau^2\right)\sigma_g^2}} \\ \pi(\sigma_g^2|d_g, a_g, s_g^2, \hat{\xi}) &\propto \frac{s_g^{2(n-2)}}{\sigma_g^{2(n+\nu+\frac{1}{2})}} e^{-\frac{2\beta + (n-1)s_g^2 + \frac{a_g^2}{\frac{1}{2n}+\tau^2}}{2\sigma_g^2}} \sqrt{\frac{1}{\left(\frac{2}{n}\sigma_g^2 + \lambda^2\right)}} e^{-\frac{d_g^2}{2\left(\frac{2}{n}\sigma_g^2 + \lambda^2\right)}} \end{aligned}$$

The posterior distribution of δ_g conditional on σ_g^2 is

$$\delta_g|d_g, \sigma_g^2 \sim N\left(\frac{\frac{nd_g}{2\sigma_g^2}}{\frac{n}{2\sigma_g^2} + \frac{1}{\lambda^2}}, \frac{1}{\frac{n}{2\sigma_g^2} + \frac{1}{\lambda^2}}\right)$$

Define

$$\begin{aligned} CI.F &= \frac{\frac{nd_g}{2\sigma_g^2}}{\frac{n}{2\sigma_g^2} + \frac{1}{\lambda^2}} \\ CI.T &= CI.F\sqrt{\frac{n}{2\sigma_g^2} + \frac{1}{\lambda^2}} \propto \frac{\frac{nd_g}{2\sigma_g^2}}{\sqrt{\frac{n}{2\sigma_g^2} + \frac{1}{\lambda^2}}} \\ CI.TP &= Prob(\delta_g > D|data, \sigma_g^2) \end{aligned}$$

based on the posterior distribution of δ_g conditional on σ_g^2

The Bayes Factor of Independence of Signal and Noise model:

$$CI.BF \propto \frac{Pr(d_g, a_g, s_g^2|\delta_g = 0, \hat{\xi})}{Pr(d_g, a_g, s_g^2|\delta_g \neq 0, \hat{\xi})}$$

$$Pr(d_g, a_g, s_g^2|\delta_g = 0, \hat{\xi}) = \int\int f(d_g|\delta_g = 0, \sigma_g^2) f(a_g|\mu_g, \sigma_g^2) f(\mu_g|\hat{\xi}) f(s_g^2|\sigma_g^2) f(\sigma_g^2|\hat{\xi}) d\mu_g d\sigma_g^2$$

$$Pr(d_g, a_g, s_g^2|\delta_g = 0, \hat{\xi}) \propto s_g^{2(n-2)} \left(1 + \frac{1}{\beta}\left(\frac{(n-1)}{2}s_g^2 + \frac{n}{4}d_g^2 + \frac{\frac{a_g^2}{2}}{\frac{1}{2n} + \tau^2}\right)\right)^{-(n+\nu)}$$

$$Pr(d_g, a_g, s_g^2|\delta_g \neq 0, \hat{\xi}) = \int\int f(d_g|\delta_g \neq 0, \sigma_g^2) f(\delta_g|\lambda^2) f(a_g|\mu_g, \sigma_g^2) f(\mu_g|\hat{\xi}) f(s_g^2|\sigma_g^2) f(\sigma_g^2|\hat{\xi}) d\mu_g d\sigma_g^2$$

$$Pr(d_g, a_g, s_g^2|\delta_g \neq 0, \hat{\xi}) \propto \int \sqrt{\frac{1}{\left(\frac{2}{n}\sigma_g^2 + \lambda^2\right)}} e^{-\frac{d_g^2}{2\left(\frac{2}{n}\sigma_g^2 + \lambda^2\right)}} \frac{s_g^{2(n-2)}}{\sigma_g^{2(n+\nu+\frac{1}{2})}} e^{-\frac{2\beta + (n-1)s_g^2 + \frac{a_g^2}{\left(\frac{1}{2n} + \tau^2\right)}}{2\sigma_g^2}} d\sigma_g^2$$

The Bayes Factor of Independence of Signal and Noise model conditional on σ_g^2 is:

$$\begin{aligned} CI.BF &\propto \frac{Pr(d_g, a_g, s_g^2|\delta_g = 0, \hat{\xi}, \sigma_g^2)}{Pr(d_g, a_g, s_g^2|\delta_g \neq 0, \hat{\xi}, \sigma_g^2)} \\ &= \sqrt{\frac{\left(\frac{2}{n}\sigma_g^2 + \lambda^2\right)}{\sigma_g^2}} e^{-\frac{\frac{n}{2}d_g^2}{2\sigma_g^2} + \frac{d_g^2}{2\left(\frac{2}{n}\sigma_g^2 + \lambda^2\right)}} \end{aligned}$$

IC: Independence of Abundance and Noise

$$\begin{aligned} \mu_g|\tau^2 &\sim N(0, \tau^2) \\ \delta_g|\lambda^2, \sigma_g^2 &\sim N(0, \sigma_g^2\lambda^2) \end{aligned}$$

posterior distribution of σ_g^2

$$f(d_g|\delta_g, \sigma_g^2) f(\delta_g|\lambda^2, \sigma_g^2) = \sqrt{\frac{1}{2\pi\left(\frac{2}{n} + \lambda^2\right)\sigma_g^2}} e^{-\frac{d_g^2}{2\left(\frac{2}{n} + \lambda^2\right)\sigma_g^2}}$$

$$f(a_g|\mu_g, \sigma_g^2) f(\mu_g|\hat{\xi}) = \sqrt{\frac{1}{2\pi\left(\frac{\sigma_g^2}{2n} + \tau^2\right)}} e^{-\frac{a_g^2}{2\left(\frac{\sigma_g^2}{2n} + \tau^2\right)}}$$

$$\pi(\sigma_g^2|d_g, a_g, s_g^2, \hat{\xi}) \propto \frac{s_g^{2(n-2)}}{\sigma_g^{2(n+\nu+\frac{1}{2})}} e^{-\frac{2\beta + (n-1)s_g^2 + \frac{d_g^2}{\left(\frac{2}{n} + \lambda^2\right)}}{2\sigma_g^2}} \sqrt{\frac{1}{\left(\frac{\sigma_g^2}{2n} + \tau^2\right)}} e^{-\frac{a_g^2}{2\left(\frac{\sigma_g^2}{2n} + \tau^2\right)}}$$

The posterior distribution of δ_g conditional on σ_g^2 is

$$\delta_g|d_g,\sigma_g^2 \sim N\left(\frac{\frac{n}{2}d_g}{\frac{n}{2}+\frac{1}{\lambda^2}},\frac{\sigma_g^2}{\frac{n}{2}+\frac{1}{\lambda^2}}\right)$$

Define

$$\begin{aligned} IC.F &= \frac{\frac{n}{2}d_g}{\frac{n}{2}+\frac{1}{\lambda^2}} \propto d_g \\ IC.T &= IC.F\sqrt{\frac{\frac{n}{2}+\frac{1}{\lambda^2}}{\sigma_g^2}} \propto \frac{d_g}{\sqrt{\sigma_g^2}} \\ IC.TP &= Prob(\delta_g > D|data,\sigma_g^2) \\ &\quad \text{based on the posterior distribution of } \delta_g \text{ conditional on } \sigma_g^2 \end{aligned}$$

The Bayes Factor of Independence of Abundance and Noise model:

$$IC.BF \propto \frac{Pr(d_g,a_g,s_g^2|\delta_g=0,\hat{\xi})}{Pr(d_g,a_g,s_g^2|\delta_g\neq 0,\hat{\xi})}$$

$$\begin{aligned} Pr(d_g,a_g,s_g^2|\delta_g=0,\hat{\xi}) &= \int\int f(d_g|\delta_g=0,\sigma_g^2)f(a_g|\mu_g,\sigma_g^2)f(\mu_g|\hat{\xi})f(s_g^2|\sigma_g^2)f(\sigma_g^2|\hat{\xi})d\mu_g d\sigma_g^2 \\ Pr(d_g,a_g,s_g^2|\delta_g=0,\hat{\xi}) &\propto \int \sqrt{\frac{1}{\left(\frac{\sigma_g^2}{2n}+\tau^2\right)}}e^{-\frac{a_g^2}{2\left(\frac{\sigma_g^2}{2n}+\tau^2\right)}}\frac{s_g^{2(n-2)}}{\sigma_g^{2(n+\nu+\frac{1}{2})}}e^{-\frac{2\beta+(n-1)s_g^2+\frac{n}{2}d_g^2}{2\sigma_g^2}}d\sigma_g^2 \end{aligned}$$

$$\begin{aligned} Pr(d_g,a_g,s_g^2|\delta_g\neq 0,\hat{\xi}) &= \int\int f(d_g|\delta_g\neq 0,\sigma_g^2)f(\delta_g|\lambda^2,\sigma_g^2)f(a_g|\mu_g,\sigma_g^2)f(\mu_g|\hat{\xi})f(s_g^2|\sigma_g^2)f(\sigma_g^2|\hat{\xi})d\mu_g d\sigma_g^2 \\ Pr(d_g,a_g,s_g^2|\delta_g\neq 0,\hat{\xi}) &\propto \int \sqrt{\frac{1}{\left(\frac{\sigma_g^2}{2n}+\tau^2\right)}}e^{-\frac{a_g^2}{2\left(\frac{\sigma_g^2}{2n}+\tau^2\right)}}\frac{s_g^{2(n-2)}}{\sigma_g^{2(n+\nu+\frac{1}{2})}}e^{-\frac{2\beta+(n-1)s_g^2+\frac{d_g^2}{\frac{2}{n}+\lambda^2}}{2\sigma_g^2}}d\sigma_g^2 \end{aligned}$$

The Bayes Factor of Independence of Abundance and Noise model conditional on σ_g^2 is:

$$\begin{aligned} IC.BF &\propto \frac{Pr(d_g,a_g,s_g^2|\delta_g=0,\hat{\xi},\sigma_g^2)}{Pr(d_g,a_g,s_g^2|\delta_g\neq 0,\hat{\xi},\sigma_g^2)} \\ &= e^{\frac{-\frac{n}{2}d_g^2+\frac{d_g^2}{\frac{2}{n}+\lambda^2}}{2\sigma_g^2}} \end{aligned}$$

CC: Complete Conjugate

$$\begin{aligned}\mu_g|\lambda_\mu^2\sigma_g^2 &\sim N(0,\sigma_g^2\lambda_\mu^2)\\ \delta_g|\lambda^2\sigma_g^2 &\sim N(0,\sigma_g^2\lambda^2)\end{aligned}$$

Posterior distribution of σ_g^2

$$\begin{aligned}f(d_g|\delta_g,\sigma_g^2)f(\delta_g|\lambda^2,\sigma_g^2) &= \sqrt{\frac{1}{2\pi\left(\frac{2}{n}+\lambda^2\right)\sigma_g^2}}e^{-\frac{d_g^2}{2\left(\frac{2}{n}+\lambda^2\right)\sigma_g^2}}\\ f(a_g|\mu_g,\sigma_g^2)f(\mu_g|\hat{\xi}) &= \sqrt{\frac{1}{2\pi\left(\frac{1}{2n}+\tau^2\right)\sigma_g^2}}e^{-\frac{a_g^2}{2\left(\frac{1}{2n}+\tau^2\right)\sigma_g^2}}\\ \pi(\sigma_g^2|d_g,a_g,s_g^2,\hat{\xi}) &\propto \frac{s_g^{2(n-2)}}{\sigma_g^{2(n+\nu)}}e^{-\frac{2\beta+(n-1)s_g^2}{2\sigma_g^2}}\sqrt{\frac{1}{\left(\frac{2}{n}+\lambda^2\right)\sigma_g^2}}e^{-\frac{d_g^2}{2\left(\frac{2}{n}+\lambda^2\right)\sigma_g^2}}\sqrt{\frac{1}{\left(\frac{1}{2n}+\tau^2\right)\sigma_g^2}}e^{-\frac{a_g^2}{2\left(\frac{1}{2n}+\tau^2\right)\sigma_g^2}}\\ \pi(\sigma_g^2|d_g,a_g,s_g^2,\hat{\xi}) &\propto \frac{s_g^{2(n-2)}}{\sigma_g^{2(n+\nu+1)}}e^{-\frac{2\beta+(n-1)s_g^2+\frac{d_g^2}{\frac{2}{n}+\lambda^2}+\frac{a_g^2}{\frac{1}{2n}+\tau^2}}{2\sigma_g^2}}\\ \sigma_g^2|d_g,a_g,s_g^2,\hat{\xi} &\sim IG\left(\nu_n,\beta_n\right)\\ \text{where } \nu_n &= \nu+n\\ \beta_n &= \beta+\frac{(n-1)}{2}s_g^2+\frac{1}{2}\frac{d_g^2}{\frac{2}{n}+\lambda^2}+\frac{1}{2}\frac{a_g^2}{\frac{1}{2n}+\tau^2}\end{aligned}$$

The posterior mode of σ_g^2 is $\frac{\beta_n}{\nu_n+1}$.

The posterior distribution of δ_g conditional on σ_g^2 is

$$\delta_g|d_g,\sigma_g^2 \sim N\left(\frac{\frac{n}{2}d_g}{\frac{n}{2}+\frac{1}{\lambda^2}},\frac{\sigma_g^2}{\frac{n}{2}+\frac{1}{\lambda^2}}\right)$$

Define

$$\begin{aligned}CC.F &= \frac{\frac{n}{2}d_g}{\frac{n}{2}+\frac{1}{\lambda^2}} \propto d_g\\ CC.T &= CC.F\sqrt{\frac{\frac{n}{2}+\frac{1}{\lambda^2}}{\sigma_g^2}} \propto \frac{d_g}{\sqrt{\sigma_g^2}}\\ CC.TP &= Prob(\delta_g > D|data,\sigma_g^2)\end{aligned}$$

based on the posterior distribution of δ_g conditional on σ_g^2

The marginal posterior distribution of δ_g is

$$\delta_g|d_g, a_g, s_g^2, \hat{\xi} \sim St\left(\frac{\frac{n}{2}d_g}{\frac{n}{2}+\frac{1}{\lambda^2}}, \left(\frac{n}{2}+\frac{1}{\lambda^2}\right)\frac{\nu_n}{\beta_n}, 2(\nu+n)\right)$$

The marginal distribution of δ_g in this case has close form. The parameter for Student t distribution are location, precision, and degree of freedom respectively. So the tail probability could be defined based on the marginal distribution of δ_g. we also tried the conditional posterior distribution of δ_g in this case to check how different the TP based on conditional posterior distribution of δ_g is from the one based on marginal posterior distribution of δ_g.

The Bayes Factor of Complete Conjuagte model:

$$CC.BF \propto \frac{Pr(d_g, a_g, s_g^2|\delta_g = 0, \hat{\xi})}{Pr(d_g, a_g, s_g^2|\delta_g \neq 0, \hat{\xi})}$$

$$\begin{aligned} Pr(d_g, a_g, s_g^2|\delta_g = 0, \hat{\xi}) &= \int\int f(d_g|\delta_g = 0, \sigma_g^2) f(a_g|\mu_g, \sigma_g^2) f(\mu_g|\hat{\xi}) f(s_g^2|\sigma_g^2) f(\sigma_g^2|\hat{\xi}) d\mu_g d\sigma_g^2 \\ Pr(d_g, a_g, s_g^2|\delta_g = 0, \hat{\xi}) &\propto s_g^{2(n-2)}\left(1+\frac{1}{\beta}\left(\frac{(n-1)}{2}s_g^2+\frac{n}{4}d_g^2+\frac{\frac{a_g^2}{2}}{\frac{1}{2n}+\tau^2}\right)\right)^{-(n+\nu)} \end{aligned}$$

$$\begin{aligned} Pr(d_g, a_g, s_g^2|\delta_g \neq 0, \hat{\xi}) &= \int\int f(d_g|\delta_g \neq 0, \sigma_g^2) f(a_g|\mu_g, \sigma_g^2) f(\mu_g|\hat{\xi}) f(s_g^2|\sigma_g^2) f(\sigma_g^2|\hat{\xi}) d\mu_g d\sigma_g^2 \\ Pr(d_g, a_g, s_g^2|\delta_g \neq 0, \hat{\xi}) &\propto s_g^{2(n-2)}\left(1+\frac{1}{\beta}\left(\frac{(n-1)}{2}s_g^2+\frac{\frac{d_g^2}{2}}{\frac{2}{n}+\lambda^2}+\frac{\frac{a_g^2}{2}}{\frac{1}{2n}+\tau^2}\right)\right)^{-(n+\nu)} \end{aligned}$$

$$CC.BF = \left(\frac{1+\frac{1}{\beta}\left(\frac{(n-1)}{2}s_g^2+\frac{n}{4}d_g^2+\frac{\frac{a_g^2}{2}}{\frac{1}{2n}+\tau^2}\right)}{1+\frac{1}{\beta}\left(\frac{(n-1)}{2}s_g^2+\frac{\frac{d_g^2}{2}}{\frac{2}{n}+\lambda^2}+\frac{\frac{a_g^2}{2}}{\frac{1}{2n}+\tau^2}\right)}\right)^{-(n+\nu)}$$

The Bayes Factor of Complete Conjugacy model conditional on σ_g^2 is:

$$\begin{aligned} CC.BF &\propto \frac{Pr(d_g, a_g, s_g^2|\delta_g = 0, \hat{\xi}, \sigma_g^2)}{Pr(d_g, a_g, s_g^2|\delta_g \neq 0, \hat{\xi}, \sigma_g^2)} \\ &= e^{\frac{-\frac{n}{2}d_g^2+\frac{d_g^2}{\frac{2}{n}+\lambda^2}}{2\sigma_g^2}} \end{aligned}$$

CC.T is a function of CC.BF.

$$
\begin{aligned}
CC.BF &= \left(\frac{1+\frac{1}{\beta}\left(\frac{(n-1)}{2}s_g^2+\frac{n}{4}d_g^2+\frac{\frac{a_g^2}{2}}{\frac{1}{2n}+\tau^2}\right)}{1+\frac{1}{\beta}\left(\frac{(n-1)}{2}s_g^2+\frac{\frac{d_g^2}{2}}{\frac{2}{n}+\lambda^2}+\frac{\frac{a_g^2}{2}}{\frac{1}{2n}+\tau^2}\right)} \right)^{-(n+\nu)} \\
&= \left(\frac{\beta_n - \frac{\frac{d_g^2}{2}}{\frac{2}{n}+\lambda^2}+\frac{n}{4}d_g^2}{\beta_n} \right)^{-(n+\nu)} \\
&= \left(1 - \frac{\frac{4}{n}\lambda^2}{\frac{2}{n}+\lambda^2}\frac{d_g^2}{\frac{\beta_n}{\nu_n+1}}\frac{1}{\nu_n+1} \right)^{-(n+\nu)} \\
&= \left(1 - K\frac{d_g^2}{\sigma_g^2} \right)^{-(n+\nu)} = \left(1-K\times CC.T^2\right)^{-(n+\nu)}
\end{aligned}
$$

Where $K = \frac{\frac{4}{n}\lambda^2}{\frac{2}{n}+\lambda^2}\frac{1}{\nu_n+1}$.

The ranking of genes based on CC.BF and absolute value of CC.T are the same.