The Reproducibility of Lists of Differentially Expressed Genes in Microarray Studies

Leming Shi[1, *], Wendell D. Jones[2], Roderick V. Jensen[3], Stephen C. Harris[1], Roger G. Perkins[4], Federico M. Goodsaid[5], Lei Guo[1], Lisa J. Croner[6], Cecilie Boysen[7], Hong Fang[4], Shashi Amur[5], Wenjun Bao[8], Catalin C. Barbacioru[9], Vincent Bertholet[10], Xiaoxi Megan Cao[4], Tzu-Ming Chu[8], Patrick J. Collins[11], Xiao-hui Fan[1], Felix W. Frueh[5], James C. Fuscoe[1], Xu Guo[12], Jing Han[13], Damir Herman[14], Huixiao Hong[4], Ernest S. Kawasaki[15], Quan-Zhen Li[16], Yuling Luo[17], Yunqing Ma[17], Nan Mei[1], Ron L Peterson[18], Raj K. Puri[13], Feng Qian[4], Richard Shippy[19], Zhenqiang Su[1], Yongming Andrew Sun[9], Hongmei Sun[4], Brett Thorn[4], Yaron Turpaz[12], Charles Wang[20], Sue Jane Wang[5], Janet A. Warrington[12], James C. Willey[21], Jie Wu[4], Qian Xie[4], Liang Zhang[22], Lu Zhang[23], Sheng Zhong[24], James J. Chen[1], Russell D. Wolfinger[8], and Weida Tong[1]

[1]National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA;
[2]Expression Analysis, 2605 Meridian Parkway, Durham, NC 27713, USA;
[3]University of Massachusetts Boston, Department of Physics, 100 Morrissey Boulevard, Boston, MA 02125, USA;
[4]Z-Tech Corporation at NCTR/FDA, 3900 NCTR Road, Jefferson, AR 72079, USA;
[5]Center for Drug Evaluation and Research, US Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA;
[6]Biogen Idec, 5200 Research Place, San Diego, CA 92122, USA;
[7]ViaLogy, 2400 Lincoln Avenue, Altadena, CA 91001, USA;
[8]SAS Institute Inc., SAS Campus Drive, Cary, NC 27513, USA;
[9]Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA;
[10]Eppendorf Array Technologies, rue du Séminaire 20a, 5000 Namur, Belgium;
[11]Agilent, 5301 Stevens Creek Boulevard, Santa Clara, CA 95051, USA;
[12]Affymetrix Inc., 3420 Central Expressway, Santa Clara, CA 95051, USA;
[13]Center for Biologics Evaluation and Research, US Food and Drug Administration, 8800 Rockville Pike, Bethesda, MD 20892, USA;
[14] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA;
[15]National Cancer Institute Advanced Technology Center, 8717 Grovemont Circle, Gaithersburg, MD 20877, USA;
[16]University of Texas Southwestern Medical Center, 6000 Harry Hines Boulevard, Dallas, TX 75390, USA;
[17]Panomics, 6519 Dumbarton Circle, Fremont, CA 94555, USA;
[18]Novartis Institutes for Biomedical Research, 250 Massachusetts Avenue, Cambridge, MA 02139, USA;
[19]GE Healthcare, 7700 S River Parkway, Tempe, AZ 85284, USA;
[20]UCLA David Geffen School of Medicine, Transcriptional Genomics Core, Cedars-Sinai Medical Center, 8700 Beverly Boulevard, Los Angeles, CA 90048, USA;
[21]Ohio Medical University, 3000 Arlington Avenue, Toledo, OH 43614, USA;
[22]CapitalBio Corporation, 18 Life Science Parkway, Changping District, Beijing 102206, China;
[23]Solexa, 25861 Industrial Boulevard, Hayward, CA 94545, USA;
[24]University of Illinois at Urbana-Champaign, Department of Bioengineering, 1304 W. Springfield Avenue, Urbana, IL 61801, USA.

Running Title: Reproducibility of Microarray Gene Lists

*Corresponding Author:

Leming Shi
National Center for Toxicological Research
US Food and Drug Administration
3900 NCTR Road
Jefferson, Arkansas 72079, U.S.A.
Tel: +1-870-543-7387
Fax: +1-870-543-7854
Leming.Shi@fda.hhs.gov

ABSTRACT

Reproducibility is a fundamental requirement in scientific experiments and clinical contexts. Recent publications raise concerns about the reliability of microarray technology because of the apparent lack of agreement between lists of differentially expressed genes (DEGs). In this study we demonstrate that (1) such discordance may stem from ranking and selecting DEGs solely by statistical significance (*P*) derived from widely used simple *t*-tests; (2) when fold change (FC) is used as the ranking criterion, the lists become much more reproducible, especially when fewer genes are selected; and (3) the instability of short DEG lists based on *P* cutoffs is an expected mathematical consequence of the high variability of the *t*-values. We recommend the use of FC ranking plus a non-stringent *P* cutoff as a baseline practice in order to generate more reproducible DEG lists. The FC criterion enhances reproducibility while the *P* criterion balances sensitivity and specificity.

Abbreviations:

A: The MAQC sample A (Stratagene Universal Human Reference RNA);

ABI: Applied Biosystems microarray platform;

AFX: Affymetrix microarray platform;

AG1: Agilent one-color microarray platform;

B: The MAQC sample B (Ambion Human Brain Reference RNA);

C: The MAQC sample C (75%A+25%B);

CV: Coefficient of variation;

D: The MAQC sample D (25%A+75%B);

DEG: Differentially expressed genes;

FC: Fold change in expression levels;

GEH: GE Healthcare microarray platform;

ILM: Illumina microarray platform;

MAQC: MicroArray Quality Control project;

P: The *P*-value calculated from a two-tailed two-sample t-test assuming equal variance;

POG: Percentage of Overlapping (common) Genes between two lists of differentially expressed genes. It is used as a measure of concordance of microarray results.

INTRODUCTION

A fundamental step in most microarray experiments is determining one or more short lists of differentially expressed genes (DEGs) that distinguish biological conditions, such as disease from health. Challenges regarding the reliability of microarray results have largely been founded on the inability of researchers to replicate DEG lists across highly similar experiments. For example, Tan *et al*.[1] found only four common DEGs using an identical set of RNA samples across three popular commercial platforms. Independent studies by the groups of Ramalho-Santos[2] and Ivanova[3] of stem cell-specific genes using the same Affymetrix platform and similar study design found a disappointing six common DEGs among about 200 identified in each study[4]. A comparative neurotoxicological study by Miller *et al*.[5] using the same set of RNA samples found only 11 common DEGs among 138 and 425, respectively, from Affymetrix and CodeLink platforms. All these studies ranked genes by *P* from simple *t*-tests, used a *P* threshold to identify DEG lists, and used the Percentage of Overlapping Genes (POG) between DEG lists as the measure of reproducibility.

Criticism of and concerns about microarrays continue to appear in some of the most prestigious scientific journals[6-10], leading to a growing negative perception regarding microarray reproducibility, and hence reliability. However, in reanalyzing the data set of Tan *et al*.[1], Shi *et al*.[11] found that cross-platform concordance was markedly improved when either simple fold change (FC) or Significance Analysis of Microarrays (SAM)[12] methods were used to rank order genes before determining DEG lists. The awareness that microarray reproducibility is sensitive to how DEGs are identified was, in fact, a major motivator for the MAQC project[11].

Several plausible explanations and solutions have been proposed to interpret and address the apparent lack of reproducibility and stability of DEG lists from microarray studies. Larger sample sizes[13]; novel, microarray-specific statistical methods[14]; more accurate array annotation information by mapping probe sequences across platforms[1, 15]; eliminating absent call genes from data analysis[11, 16, 17]; improving probe design to minimize cross-hybridization[15]; standardizing manufacturing processes[1]; and improving data quality by fully standardizing sample preparation and hybridization procedures are among the suggestions for improvement[18].

The MAQC study was specifically designed to address these previously identified sources of variability in DEG lists. Two very different RNA samples, Stratagene Universal Human Reference RNA and Ambion Human Brain Reference RNA, with thousands of differentially expressed genes, were prepared in sufficient quantities and distributed to three different laboratories for each of the five different commercial whole genome microarray platforms participating in the study. For each platform, each sample was analyzed using five technical replicates with standardized procedures for sample processing, hybridization, scanning, data acquisition, data preprocessing, and data normalization at each site. The probe sequence information was used to generate a stringent mapping of genes across the different platforms and 906 genes were further analyzed with TaqMan® assays using the same RNA samples.

A careful analysis of these MAQC data sets, along with numerical simulations and mathematical arguments, demonstrates that the reported lack of reproducibility of DEG lists can be attributed in large part to identifying DEGs from simple *t*-tests without consideration of FC when sample numbers are small. The finding holds for intra-laboratory, inter-laboratory, and cross-platform comparisons independent of sample pairs and normalization methods, and is increasingly apparent with decreasing number of genes selected.

As a basic procedure for improving reproducibility while balancing specificity and sensitivity, choosing genes using a combination of FC ranking and *P* threshold was investigated. This joint criterion results in DEG lists with much higher POG, commensurate with better reproducibility, than lists generated by *t*-test *P* alone, even when selecting a relatively small numbers of genes. An FC criterion explicitly incorporates the measured quantity to enhance reproducibility, whereas a *P* criterion incorporates control of sensitivity and specificity. The results increase our confidence in the reproducibility of microarray studies while supporting a need for caution in the use of inferential statistics when sample numbers are small. While numerous more advanced statistical modeling techniques have been proposed and compared for selecting DEGs[14, 19, 20], the primary objectives here are to explain that the primary reason for microarray reproducibility concerns is failure to include an FC criterion during gene selection, and to recommend a simple and straightforward approach concurrently satisfying statistical and reproducibility requirements. It should be stressed that robust methods are needed to meet stringent clinical requirements for reproducibility, sensitivity and specificity of microarray applications in, for example, clinical diagnostics and prognostics

RESULTS

The POG for a number of gene selection scenarios employing *P* and/or FC are compared and a numerical example (see side box) is provided that shows how the simple *t*-test, when sample size is small, results in selection of different genes purely by chance. While the data lack biological variability, the results are supported by the toxicogenomic data of Guo *et al*.[21] While *P* could be computed from many different statistical methods, for simplicity and consistency, throughout this article *P* is calculated with the two-tailed *t*-test that is widely employed in microarray data analysis.

1. Inter-site Concordance for the Same Platform

Figure 1 gives plots of inter-site POG versus number of genes for each MAQC platform. Since there are three possible inter-site comparisons (S1-S2, S1-S3, and S2-S3) and six gene selection methods (see Methods), there are 18 POG lines for each platform. Figure 1 shows that inter-site reproducibility in terms of POG heavily depends on the number of chosen differential genes and the gene ranking criterion: Gene selection using FC ranking gives consistently higher POG than *P* ranking. The POG from FC ranking is near 90% for as few as 20 genes for most platforms, and remains at this high inter-site concordance level as the number of selected genes increases. In contrast, the POG from *P* ranking is in the range of 20-40% for as many as 100 genes, and then asymptotically approaches 90% only after several thousand genes are selected.

The POG is higher when the analyses are limited to the genes commonly detected ("Present" in the majority of replicates for each sample) by both test sites under comparison (Supplementary Fig. 1, available online). In addition to a slight increase (2-3%), the inter-site POG lines after noise filtering are more stable than those before noise filtering, particularly for ABI, AG1, and GEH. Furthermore, differences between the three ILM test sites are further decreased after noise filtering, as seen from the convergence of the POG for S1-S2, S1-S3, and S2-S3 comparisons. Importantly, noise filtering does not change either the trend or magnitude of the higher POG graphs for FC ranking compared with *P*-ranking.

Inter-site concordance for different FC and *P* ranking criteria were also calculated for other MAQC sample pairs having much smaller differences than for sample A versus sample B, and correspondingly lower FC. In general, POG is much lower for other sample pairs regardless of ranking method and ranking order varies more greatly, though FC ranking methods still consistently gives a higher POG than *P* ranking methods. Supplementary Figure 2 gives the plots of POG for Sample C versus Sample D[22, 23] for all inter-site comparisons.

The substantial difference in inter-site POG shown in Figure 1 and Supplementary Figure 1 is a direct result of applying different gene selection methods to the same data sets, and clearly depicts how perceptions of inter-site reproducibility can be affected for any microarray platform. While the emphasis here is on reproducibility in terms of POG, in practice, this criterion must be balanced against other desirable characteristics of gene lists, such as specificity and sensitivity (when the truth is binary) or mean squared error (when the truth is continuous), considerations that that are discussed further in later sections.

2. Cross-platform Concordance

Figure 2 shows the substantial effect that FC- and *P*-ranking based gene selection methods have on cross-platform POG. Similar to inter-site comparisons, *P*-ranking results in lower cross-platform POG than FC-ranking. When FC is used to rank DEGs from each platform, the cross-platform POG is around 70-85%, depending on the platform pair. The platforms themselves contribute about 15% differences in the cross-platform POG, as seen from the spread of the blue POG lines. Noise filtering improves FC-ranking cross-platform POG by about 5-10% and results in more stable POG when a smaller number of genes are selected (Fig. 2b). Importantly, the relative differences between FC- and *P*-ranking methods remain the same after filtering.

3. Concordance between Microarray and TaqMan® Assays

TaqMan® real-time PCR assays are widely used to validate microarray results[24, 25]. In the MAQC project, the expression levels of 997 genes randomly selected from available TaqMan® assays have been quantified in the four MAQC samples[22, 26]. Nine hundred and six (906) of the 997 genes are among the "12,091" set of genes found on all of the six genome-wide microarray platforms[22]. There are four TaqMan® assays technical replicates for each sample and the DEGs for TaqMan® assays were identified using the same six gene selection procedures as those used for microarray data. The DEGs calculated from the microarray data are compared with DEGs calculated from TaqMan® assay data. With noise filtering (*i.e*., focusing on the genes detected by both the microarray platform and TaqMan® assays), 80-85% concordance was observed (Fig. 3). Consistent with inter-site and cross-platform comparisons, POGs comparing microarray with TaqMan® assays also show that ranking genes by FC results in markedly higher POG than ranking by *P* alone, especially for short gene lists. POG results without noise filtering (Supplementary Fig. 3) are some 5% lower but the notable differences in POG between the FC- and *P*- ranking are unchanged.

4. Reproducibility of FC and t-statistic: Different Metrics for Identifying Differentially Expressed Genes (DEGs)

Figure 4 shows that the inter-site reproducibility of log2 FC (panel a) is much higher than that of log2 t-statistic (panel b). In addition, the relationship between log2 FC and log2 t-statistic from the same test site is non-linear and the correlation appears to be low (panel c). We see similar results when data from different microarray platforms are compared to each other or when microarray data are compared against TaqMan® assay data (results not shown). The differences between the reproducibility of FC and *t*-statistic observed here are consistent with the differences between POG results in inter-site (Fig. 1), cross-platform (Fig. 2), and microarray versus TaqMan® assay (Fig. 3) comparisons. The nonlinear relationship between log2 FC and log2 t-statistic (Fig. 5c) leads to low concordance of lists of DEGs derived from FC ranking when compared to a list derived from t-statistic (*P*) ranking (Supplementary Fig. 4); an expected outcome due to the different emphases of FC and *P*.

5. Joint Fold Change and *P* Rule Illustrated with a Volcano Plot: Ranking by FC, not by *P*

Supplementary Figure 5 is a volcano plot depicting how a joint FC and *P* rule works in gene selection. It uses the MAQC Agilent data, and plots negative log *P* on the y-axis versus log FC on the x-axis. A joint rule chooses genes that fall in the upper left and right sections of the plot (sections A and C of Fig. 5). Other possible cutoff rules for combining FC and *P* are apparent, but are precluded from inclusion due to space. An important conclusion from this study is that genes should be ranked and selected by FC (x-axis) with a non-stringent *P* threshold in order to generate reproducible lists of DEGs.

6. Concordance Using Other Statistical Tests

Numerous different statistical tests including rank tests (*e.g*., Wilcoxon rank-sum test) and shrunken t-tests (*e.g*., SAM) have been used for the identification of DEGs. Although this work is not intended to serve as a comprehensive performance survey of different statistical procedures, we set out to briefly examine a few examples due to their popularity. Figure 5 shows the POG results of several commonly used approaches including FC ranking, t-test statistic, Wilcoxon rank-sum test, and SAM using AFX site-site comparison as an example. The POG by SAM (pink line), although greatly improved over that of simple t-test statistic (purple line), approached, but did not exceed, the level of POG based on FC ranking (green line). In addition, the small numbers of replicates in this study rendered many ties in the Wilcoxon rank statistic, resulting in poor inter-site concordance in terms of rank-order of the DEGs between the two AFX test sites. Similar findings (data not shown) were observed using the toxicogenomics data set of Guo *et al*.[21]

7. Gene Selection in Simulated Datasets

The MAQC data, like data from actual experiments, allows evaluation of DEG list reproducibility, but not of truth. Statistics are used to estimate truth, often in terms of sensitivity and specificity, but the estimates are based on assumptions about data variance and error structure that are also unknown. Simulations where truth can be specified *a priori* are useful to conduct parametric evaluations of gene selection methods, and true false positives and negatives are then known. However, results are sensitive to assumptions regarding data structure and error that for microarrays remains poorly characterized.

Figure 6 gives POG versus number of genes for three simulated data sets (MAQC-simulated set, Small-Delta simulated set, and Medium-Delta simulated set, see Methods) that were prepared in order to compare the same gene selection methods as the MAQC data. The MAQC-simulated set was created to emulate the magnitude and structure of differential expression observed between the actual MAQC samples A and B (*i.e.*, thousands of genes with FC > 2). By comparison, the Small-Delta simulated data set had only 50 significant genes with FC > 2 and most genes had FC < 1.3. The Medium-Delta data set had FC profiles in between.

For the MAQC-simulated data, either FC ranking or FC ranking combined with any of the *P* threshold resulted in markedly higher POG than any *P* ranking method, regardless of gene list length and coefficient of variation (CV) of replicates. The POG is ~100%, ~95%, and ~75%, for replicate CV values of 2%, 10%, and 35% CV, respectively, decreasing to about 20-30% with an exceedingly high (100%) CV. In contrast, POG from *P* ranking alone varies from a few percent to only ~10% when 500 genes are selected.

For the Medium- and Small-Delta simulated data sets, we see differences start to emerge between using FC alone and FC with *P* cutoff. From Figure 6, when variances in replicates become larger (CV > 10%), we see that reproducibility is greatly enhanced using FC ranking with a suitable *P* cutoff versus FC or *P* by themselves. In addition, when variances are small (CV ≤ 10%), we see that reproducibility is essentially the same for FC with *P* or without. What is clear is that *P* by itself did not produce the most reproducible list for any condition simulated.

Although *P* ranking generally resulted in very low POG, a false positive was rarely produced, even for a list size of 500 (data not shown). Thus, the *P* criterion performed as expected, and identified mostly true positives. Unfortunately, the probability of selection of the same true positives with a fixed *P* cutoff in a replicated experiment appears small due to variation in the *P* statistic itself (see inset). FC ranking by itself resulted in a large number of false positives with a large number of genes for the Medium and Small-Delta sets when genes with small FCs are selected as differentially expressed. These false positives were greatly reduced to the same level as for the *P*-ranking alone when FC ranking was combined with a *P*-cutoff.

DISCUSSION

A fundamental requirement in microarray experiments is that the identification of DEG lists must be reproducible if the data and scientific conclusion from them are to be credible. DEG lists are normally developed by rank-ordering genes in accordance with a suitable surrogate value to represent biological relevance, such as the magnitude of the differential expression (*i.e*., FC) or the measure of statistical significance (*P*) of the expression change, or both. The results show that concurrent use of both FC ranking and *P*-cutoff as criteria to identify biological relevant genes can be essential to attain reproducible DEG lists across laboratories and platforms.

A decade since the microarray-generated differential gene expression results of Schena *et al*.[27] and Lockhart *et al*.[28] were published, microarray usage has become ubiquitous. Over this time, many analytical techniques for identifying DEGs have been introduced and used. Early studies predominantly relied on the magnitude of differential expression change in experiments done with few if any replicates, with an FC cutoff typically of two used to reduce false positives. Mutch *et al*.[29] recommended using intensity-dependent FC cutoffs to reduce biased selection of genes with low expression.

Gene selection using statistical significance estimates became more prevalent during the last few years as studies with replicates became possible. Incorporation of a t-statistic in gene selection was intended to compensate for the heterogeneity of variances of genes [30]. Haslett *et al.* [31] employed stringent values of both FC and *P* to determine DEGs. In recent years, there has been an increasing tendency to use *P* ranking for gene selection. Kittleson *et al*.[32] selected genes with a FC cutoff of two and a very restrictive Bonferroni corrected *P* of 0.05 in a quest for a short list of true positive genes. Tan *et al*.[33] used *P* to rank genes. Correlation coefficient, which behaves similarly to the t-statistic, has also been widely used as a gene selection method in the identification of signature genes for classification purposes[13, 34, 35].

New and widely employed methods have appeared in recent years and that implicitly correct for the large variance in the t-statistic that results when gene variance is estimated with a small number of samples. Allison *et al*.[14] collectively described these methods as "variance shrinkage" approaches. They include the popular permutation-based "SAM" procedure[5, 12, 36, 37], Bayesian-based approaches[30, 38] and others[39]. Qin *et al*.[19] compared several variance shrinkage methods with a simple t-statistic and FC for spike-in gene identification on a two-color platform, concluding that all methods except *P* performed well. All these methods have the effect of reducing a gene's variance to be between the average for the samples, and the average over the arrays.

In some cases, however, the use of FC for gene selection was criticized and entirely abandoned. For example, Callow *et al*.[40], using *P* alone for identifying DEGs, concluded that *P* alone eliminated the need for filtering low intensity spots because the t-statistic is uniformly distributed across the entire intensity range. Reliance on *P* alone to represent a gene's FC and variability in gene selection has become commonplace. Norris and Kahn[41] describe how false discovery rate (FDR) has become so widely used as to constitute a standard to which microarray analyses are held. However, FDR usually employs a shrunken t-statistic and genes are ranked and selected similar to *P* (see Figure 6).

Prior to MAQC, Irizarry *et al*.[42] compared data from five laboratories and three platforms using the CAT plots that are essentially the same as the POG graphs used in our study. Lists of less than 100 genes derived from FC ranking showed 30 to 80% intra-site, inter-site, and inter-platform concordance. Interestingly, important disagreements were attributable to a small number of genes with large fold change that they posit resulted from a laboratory effect due to inexperienced technicians and sequence-specific effects where some genes are not correctly measured.

Exactly how to best employ FC with *P* to identify genes is a function of both the nature of the data, and the inevitable tradeoff between sensitivity and specificity that is familiar across research, clinical screening and diagnostics, and even drug discovery. But how the tradeoff is made depends on the application. Fewer false negatives at the cost of more false positives may be desirable when the application is identifying a few hundred genes for further study, and FC ranking with a non-stringent *P* value cutoff from a simple t-test could be used to eliminate some noise. The gene list can be further evaluated in terms of gene function and biological pathway data, as illustrated in Guo *et al*.[21] for toxicogenomic data. Even for relatively short gene lists, FC ranking together with a non-stringent *P* cutoff should result in reproducible lists. In addition, DEG lists identified by the ranking of FC is much less susceptible to the impact of normalization methods. In fact, global scaling methods (*e.g*., median- or mean-scaling) do not change the relative ranking of genes based on FC; they do, however, impact gene ranking by *P*-value.

The tradeoffs between reproducibility, sensitivity, and specificity become pronounced when genes are selected by *P* alone without consideration of FC, especially when a stringent *P* cutoff is used to reduce false positives. When sample numbers are small, any gene's t-statistic can change considerably in repeated studies within or across laboratories or across platforms. Each study can select different significant genes, purely by chance. It is entirely possible that separately determined lists will have a small proportion of common genes even while each list comprises mostly true positives. This apparent lack of reproducibility of the gene lists is an expected outcome of statistical variation in the t-statistic for small numbers of samples. In other words, each study fails to produce some, but not all, of the correct results. The side box provides a numerical example of how gene list discordance can result from variation in the t-statistic across studies. Decreasing the *P* cutoff will increase the proportion of true positives, but also diminish the number of selected genes, diminish genes common across lists, and increase false negatives. Importantly, selecting genes based on a small *P* cutoff derived from a simple t-test without consideration of FC renders the gene list non-reproducible in many cases.

Additional insight is gained by viewing gene selection from the perspective of the biologist ultimately responsible for interpreting microarray results. Statistically speaking, a microarray experiment tests 10,000 or more null hypotheses where essentially all genes have non-zero differential expression. Statistical tests attempt to account for an unknowable error structure, in order to eliminate the genes with low probability of biological relevance. To the biologist, however, the variance of a gene with a large FC in one microarray study may be irrelevant if a similar experiment again finds the gene to have a large FC; the second experiment would probably be considered a validating reproduction. This conclusion would be reasonable since the gene's *P* depends on a poor estimate of variance across few samples, whereas a repeated FC measurement is tangible reproducibility which tends to increase demonstrably with increasing FC. The biological interpreter can also consider knowledge of gene function and biological pathways before finally assigning biological relevance, and will be well aware that either *P* or FC is only another indicator regarding biological significance.

This study shows that genes with smaller expression fold changes generated from one platform or laboratory are, in general, less reproducible in another laboratory with the same or different platforms. However, it should be noted that genes with small fold changes may be biologically important[43]. When a fixed FC cutoff is chosen, sensitivity could be sacrificed for reproducibility. Alternatively, when a high *P* cutoff (or no *P* cutoff) is used, specificity could be sacrificed for reproducibility. Ultimately, the acceptable trade-off is based on the specific question being asked or the need being addressed. When searching for a few reliable biomarkers, high FC and low *P* cutoffs can be used to produce a highly specific and reproducible gene list. When identifying the components of genetic networks involved in biological processes, a lower FC and higher *P* cutoff can be used to identify larger, more sensitive but less specific, gene lists. In this case additional biological information about putative gene functions can be incorporated to identify reliable gene lists that are specific to the biological process of interest.

Truly differentially expressed genes should be more likely identified as differentially expressed by different platforms and from different laboratories than those genes with no differential expression between sample groups. In the microarray field, we usually do not have the luxury of knowing the "truth" in a given study. Therefore, it is not surprising that most microarray studies and data analysis protocols have not been adequately evaluated against the "truth". A reasonable surrogate of such "truth" could be the consensus of results from different microarray platforms, from different laboratories using the same platform, or from independent methods such as TaqMan® assays, as we have extensively explored in this study.

The fundamental scientific requirement of reproducibility is a critical dimension to consider along with balancing specificity and sensitivity when defining a gene list. Irreproducibility would render microarray technology generally, and any research result, specifically, vulnerable to criticism. New methods for the identification of DEGs continue to appear in the scientific literature. These methods are typically promoted in terms of improved sensitivity (power) while retaining nominal rates of specificity. Reproducibility is seldom emphasized.

The results show that selecting DEGs based solely on *P* from a simple t-test most often predestines a poor concordance in DEG lists, particularly for small numbers of genes. In contrast, using FC ranking in conjunction with a *P*-cutoff results in more concordant gene lists concomitant with needed reproducibility, even for fairly small numbers of genes. Moreover, enhanced reproducibility holds for inter-site, cross-platform, and between microarray and TaqMan® assay comparisons, and is independent of platforms, sample pairs, and normalization methods (data not shown). The results should increase confidence in the reproducibility of data produced by microarray technology and should also expand awareness that gene lists identified solely based on *P* will tend to be discordant. This work demonstrates the need for a shift from the common practice of selecting differentially expressed genes solely on the ranking of a statistical significance measure (*e.g.*, t-statistic) to an approach that emphasizes fold-change, a quantity actually measured by microarray technology.

Conclusions and Recommendations:

1. A fundamental step of microarray studies is the identification of a small subset of DEGs from among tens of thousands of genes probed on the microarray. DEG lists must be concordant to satisfy the scientific requirement of reproducibility, and must also be specific and sensitive for scientific relevance. A baseline practice is needed for properly assessing reproducibility/concordance alongside specificity and sensitivity.
2. Reports of DEG list instability in the literature are often a direct consequence of comparing DEG lists derived from a simple t-statistic when the sample size is small and variability in variance estimation is large. Therefore, the practice of using *P* alone for gene selection should be discouraged.
3. A DEG list should be chosen in a manner that concurrently satisfies scientific objectives in terms of inherent tradeoffs between reproducibility, specificity, and sensitivity.
4. Using FC and *P* together balances reproducibility, specificity, and sensitivity. Control of specificity and sensitivity can be accomplished with a *P* criterion, while reproducibility is enhanced with an FC criterion. Sensitivity can also be improved by better platforms with greater dynamic range and lower variability or by increased sample sizes.
5. FC ranking should be used in combination with a non-stringent *P* threshold to select a DEG list that is reproducible, specific, and sensitive, and a joint rule is recommended as a baseline practice.
6. These conclusions and recommendations are further supported by toxicogenomic results from Guo *et al*.[21]

METHODS

MAQC Data Sets

Analyses identified differentially expressed genes between the primary samples A (Stratagene Universal Human Reference RNA, Catalog #740000) and B (Ambion Human Brain Reference RNA, Catalog #6050) of the MAQC study. Analyses are additionally limited to data sets from the following five commercial genome-wide microarray platforms: ABI (Applied Biosystems), AFX (Affymetrix), AG1 (Agilent one-color), GEH (GE Healthcare), and ILM (Illumina), and to the subset of “12,091” genes commonly probed by them. TaqMan® assay data for 906 genes are used to examine gene list comparability between microarrays and TaqMan® assays. For more information about the MAQC project and the data sets, refer to Shi *et al*.[22].

Normalization Methods

The following manufacturer’s preferred normalization methods were used: quantile normalization for ABI and ILM, PLIER for AFX, and median-scaling for AG1 and GEH[22]. For quantile normalization (including PLIER), each test site is independently considered.

Gene Ranking (Selection) Rules

Six gene ranking (selection) methods were examined: (1) FC (fold change ranking); (2) FC_*P*0.05 (FC ranking with *P* cutoff of 0.05); (3) FC_*P*0.01 (FC ranking with *P* cutoff of 0.01); (4) *P (P* ranking, simple t-test assuming equal variance); (5) *P*_FC2 (*P* ranking with FC cutoff of 2); (6) *P*_FC1.4 (*P* ranking with FC cutoff of 1.4). When a cutoff value (*e.g*., *P*<0.05) is imposed for a ranking metric (*e.g*., FC), it is likely that the lists of candidate genes that meet the cutoff value may not be the same for the two test sites or two platforms as a result of differences in inter-site or cross-platform variations. Such differences are part of the gene selection process and have been carried over to the gene ranking/selection stage.

Evaluation Criterion - POG (Percentage of Overlapping Genes)

The POG (percentage of overlapping genes) criterion was applied in three types of comparisons: (1) Inter-site comparison using data from the three test sites of each platform; (2) Cross-platform comparison between ABI, AFX, AG1, GEH, and ILM using data from test site 1; for each sample pair, there are ten cross-platform pairs for comparison; (3) Microarray versus TaqMan® assay comparisons.

POG is calculated for several different cutoffs that can be considered as arbitrary.
The number of genes considered as differentially expressed is denoted as 2L, where L is both the number of genes up- and down-regulated. The number of genes available for ranking and selection in one direction, L, varies from 1 to 6000 (with a step of one) or when there are no more genes in one regulation direction, corresponding to 2L varying from 2 to 12,000. Directionality of gene regulation is considered in POG calculations; genes selected by two sites or platforms but with different regulation directionalities are considered as discordant.

The formula for calculating POG is: POG = 100*(DD+UU)/2L, where DD and UU are the number of commonly down- or up-regulated genes, respectively, from the two lists, and L is the number of genes selected from the up- or down-regulation directionality. To overcome the confusion of different numbers for the denominator, in our POG calculations we deliberately selected an equal number of up-regulated and down-regulated genes, L.

Noise-filtering

Most of the analyses in this study exclude flagging information; that is, the entire set of "12,091" genes is used in the analyses. Some calculations are limited to subsets of genes commonly detectable ("common present") by the two test sites or two platforms under comparison. To be denoted as "commonly present", the gene is detected ("present") in the majority of replicates (*e.g.*, three or more when there are five replicates) for each sample in a sample-pair comparison and for each test site or platform.

Gene Selection Simulation

A simulation was created to emulate the characteristics of the MAQC dataset. Fifteen thousand simulated genes were created where 5,000 were undifferentiated in expression between simulated biological samples A and B and 10,000 were differentiated but at various levels (exponential distribution for the log ratio, where almost 4,000 are differentiated two-fold or higher, similar to a typical platform in the MAQC study, divided equally into up and down regulated genes). To simulate instances of technical or biological replicates, multiplicative noise (error) was added to the signal for each gene for each of five simulated replicates for each sample using an error distribution that would produce a replicate CV similar to that typically seen in the MAQC data set (ie, the mean replicate CV would be roughly 10%). The CV for any given gene was randomly selected from a trimmed exponential distribution. To address a variety of additional error scenarios but preserving the same distribution of fold change, we also considered three additional mean CV values (2%, 35%, and 100%). To understand the impact of gene list size on the stability of the DEG list, list sizes of 10, 25, 100, and 500 genes were examined for each mean CV scenario. Several gene selection rules were compared: FC ranking only, *P* ranking only, and shrunken t-statistic ranking. Note: *P* ranking is equivalent to t-statistic ranking as well as ranking based on FDR that monotonically transforms the *P*-value. In addition, shrunken t-statistic ranking is equivalent to ranking based on the test statistic used by SAM and related methods. In addition, rules based on FC ranking with a *P* threshold were also compared (for *P*=0.1, 0.01, 0.001, and 0.0001). Finally, to simulate differences in the variation patterns of analytes between platforms and even between laboratories, covariance between laboratories/platforms of the variance for each gene was included in the simulations. For a given mean CV, 20 or more simulated instances of (5) replicates of simulated biological samples A and B were created and DEG lists were prepared for each instance that were rank ordered using the methods described above. To determine reproducibility of a given method for a given mean CV using a given gene list size, the rank-ordered gene lists from these 20 instances were pair-wise compared for consistency and reproducibility. The results presented in the graphs are averages from those pair-wise comparisons.

The MAQC actual data is characterized by large magnitudes of differential expression among the vast majority of the 12,091 common genes, with some 4000 exhibiting FC > 2 and hundreds with FC > 10. As such, the data may be atypical of commonplace microarray experiments with biological effects. Consequently, two other simulation data sets were created with far fewer genes with large FC, as might be expected in some actual microarray experiments. Specifically, the Small-Delta data set was created with fewer than 50 genes with FC > 2, and a FC < 1.3 for most differentiated genes, and 10,000 undifferentiated genes. In addition, the variances of the genes were correlated similar to that observed in the MAQC data. The third simulated dataset, termed the Medium-Delta set, had a large number of differentiated genes similar to the MAQC simulated dataset, but with small FC similar to the Small-Delta set. Again, gene variances were correlated similar to that observed in the MAQC data.

ACKNOWLEDGMENTS

We thank participants of the MicroArray Quality Control (MAQC) project for generating the large data sets that were used in this study. Many MAQC participants contributed to the sometimes-heated discussions on the topic of this paper during MAQC teleconferences and face-to-face project meetings. The common conclusions and recommendations evolved from this extended discourse.

DISCLAIMER

This document has been reviewed in accordance with United States Food and Drug Administration (FDA) policy and approved for publication. Approval does not signify that the contents necessarily reflect the position or opinions of the FDA nor does mention of trade names or commercial products constitute endorsement or recommendation for use. The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the FDA. James C. Willey is a consultant for and has significant financial interest in Gene Express, Inc.

Variability of the two-sample t-statistic

In a two-sample t-test comparing the mean of sample A to the mean of sample B, the t-statistic is given as follows:

$$t = \frac{\overline{X}_B - \overline{X}_A}{\sqrt{{S_p}^2/n_B + {S_p}^2/n_A}}$$

where $\overline{X}_A$ is the average of the log2 expression levels of sample A with n_A replicates, and $\overline{X}_B$ is the average of the log2 expression levels of sample B with n_B replicates, and $S_p{}^2$=(SS_A+SS_B)/(n_A+n_B-2) is the pooled variance of samples A and B, and *SS* denotes the sum of squared errors. The numerator of the t-statistic is the fold-change (FC) in log2 scale and represents the signal level of the measurements (*i.e*., the magnitude of the difference between the expression levels of sample A and sample B). The denominator represents the noise components from the measurements of samples A and B. Thus, the t-statistic represents a measure of the signal-to-noise ratio. Therefore, the FC and the t-statistic (*P*) are two measures for the differences between the means of sample A and sample B. The t-statistic is intrinsically less reproducible than FC when the variance is small.

Assume the data are normally distributed, the variances of samples A and B are equal (σ^2), the numbers of replicates in samples A and B are equal ($n = n_A = n_B$), and that there is a real difference in the mean values between samples A and B, d (the true FC in log2 scale). Then the t-statistic has a non-central t-distribution with non-central parameter

$\delta = (d/\sigma)(\sqrt{n/2})$,

and the mean and variance of the *t*-statistic (Johnson and Kotz, 1970) are

$$E(t) = \left(\frac{1}{2}\nu\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{1}{2}(\nu-1)\right)}{\Gamma\left(\frac{1}{2}\nu\right)}\delta \quad , \qquad Var(t) = \frac{\nu}{\nu-2} + \left(\frac{\nu}{\nu-2} - \left[\left(\frac{1}{2}\nu\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{1}{2}(\nu-1)\right)}{\Gamma\left(\frac{1}{2}\nu\right)}\right]^2\right)\delta^2$$

where ν= (*2n-2*) and is the degrees of freedom of the non-central t-distribution. When d = 0 (the two means are equal), then the t-statistic has a t-distribution with mean $E(t) = 0$ and $Var(t) = \nu/(\nu-2)$. The variance of the t-statistic depends on the sample size *n*, the magnitude of the difference between the two means *d,* and the variance σ^2. On the other hand, the variance of the mean difference for the FC is $(2/n)\sigma^2$. That is, the variance of the FC depends only on the sample size *n* and the variance σ^2, regardless of the magnitude of the difference *d* between the two sample means.

In an MAQC data set, a typical sample variance for the log2 expression levels is approximately $\sigma^2 = 0.15^2$. With n = 5, the standard deviation of the FC (in log2 scale) is 0.09. For a differentially expressed gene with a 4-fold change between 5 replicates of sample A and 5 replicates of sample B, d = 2 and the t-values have a non-central t-distribution with (ν=n_A+n_B-2) = 8 degrees of freedom and δ = 21.08. From the equations above, the mean and the variance of the t-values are E(t) = 23.35 and Var(t) = 6.96^2. Within two standard deviations the expected value of the t-value ranges from 9.43 (=23.35-2 x 6.96) to 37.27 (=23.35+2 x 6.96), corresponding to *P*s from 1 x 10^{-5} to 3 x 10^{-10}, based on the Student's two-sided t-test with 8 degrees of freedom. In contrast when $n=5$ the standard deviation of the FC (in log2 scale) is 0.09. The expected value of the FC ranges only from 3.53 (= $2^{1.82}$) to 4.53 (=$2^{2.18}$) within two standard deviations. In this case, this gene would be selected as differentially expressed using either a FC cutoff of 3.5 or a *P* cutoff of 1 x 10^{-5}. On the other hand, for a gene with a 2-fold change (d = 1), the t-statistic has a non-central t-distribution with δ = 10.54. The mean and the variance of the t-statistic are E(t) = 11.68 and Var(t) = 3.62^2 with a corresponding *P* of 3 x 10^{-6} at t = 11.68. Using the same *P* cutoff, 1 x 10^{-5}, this gene is likely to be selected with the probability greater than 0.5. For the FC criterion, the expected value of the FC ranges from 1.76 (= $2^{0.82}$) to 2.26 (=$2^{1.18}$). Using the same FC cutoff of 3.5, this gene is very unlikely to be selected. Thus, the top ranked gene list based on the FC is more reproducible than the top ranked gene list based on the *P*. The top ranked genes selected by a *P* cutoff may not be reproducible between experiments although both lists may contain mostly differentially expressed genes.

Reference: Johnson and Kotz (1970). Continuous Univariate Distributions - 2. Houghton Mifflin, Boston.

REFERENCES

1. Tan, P.K. et al. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31**, 5676-5684 (2003).
2. Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R.C. & Melton, D.A. "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science* **298**, 597-600 (2002).
3. Ivanova, N.B. et al. A stem cell molecular signature. *Science* **298**, 601-604 (2002).
4. Fortunel, N.O. et al. Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". *Science* **302**, 393; author reply 393 (2003).
5. Miller, R.M. et al. Dysregulation of gene expression in the 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine-lesioned mouse substantia nigra. *J Neurosci* **24**, 7445-7454 (2004).
6. Miklos, G.L. & Maleszka, R. Microarray reality checks in the context of a complex disease. *Nat Biotechnol* **22**, 615-621 (2004).
7. Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**, 488-492 (2005).
8. Ioannidis, J.P. Microarrays and molecular research: noise discovery? *Lancet* **365**, 454-455 (2005).
9. Frantz, S. An array of problems. *Nat Rev Drug Discov* **4**, 362-363 (2005).
10. Marshall, E. Getting the noise out of gene arrays. *Science* **306**, 630-631 (2004).
11. Shi, L. et al. Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics* **6 Suppl 2**, S12 (2005).
12. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-5121 (2001).
13. Ein-Dor, L., Zuk, O. & Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A* **103**, 5923-5928 (2006).
14. Allison, D.B., Cui, X., Page, G.P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* **7**, 55-65 (2006).
15. Mecham, B.H. et al. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res* **32**, e74 (2004).
16. Barczak, A. et al. Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res* **13**, 1775-1785 (2003).
17. Shippy, R. et al. Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics* **5**, 61 (2004).
18. Hoffman, E.P. Expression profiling--best practices for data generation and interpretation in clinical trials. *Nat Rev Genet* **5**, 229-237 (2004).

19. Qin, L.X. & Kerr, K.F. Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Res* **32**, 5471-5479 (2004).
20. Kim, S. & Lee, J. Comparison of various statistical methods for identifying differential gene expression in replicated microarray data. *Statistical Methods in Medical Research* **15**, 1-18 (2006).
21. Guo, L. et al. Biological response is preserved across microarray platforms. *Nat Biotechnol* **24**, MS-13, submitted (2006).
22. Shi, L. et al. MicroArray Quality Control (MAQC) Project: A comprehensive survey demonstrates concordant results between gene expression technology platforms. *Nat Biotechnol* **24**, MS-3, submitted (2006).
23. Shippy, R. et al. The use of RNA sample titrations for assessing microarray platform performance and normalization techniques. *Nat Biotechnol* **24**, MS-8, submitted (2006).
24. Wang, Y. et al. Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. *BMC Genomics* **7**, 59 (2006).
25. Kuo, W.P. et al. A sequence oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat Biotechnol* **in press** (2006).
26. Canales, R.D. et al. Validation of DNA Microarray Data Relative to Quantitative Gene Expression Measurement Technologies. *Nat Biotechnol* **24**, MS-7, submitted (2006).
27. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470 (1995).
28. Lockhart, D.J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**, 1675-1680 (1996).
29. Mutch, D.M., Berger, A., Mansourian, R., Rytz, A. & Roberts, M.A. Microarray data analysis: a practical approach for selecting differentially expressed genes. *Genome Biol* **2**, PREPRINT0009 (2001).
30. Baldi, P. & Long, A.D. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17**, 509-519 (2001).
31. Haslett, J.N. et al. Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle. *Proc Natl Acad Sci U S A* **99**, 15000-15005 (2002).
32. Kittleson, M.M. et al. Gene expression in giant cell myocarditis: Altered expression of immune response genes. *Int J Cardiol* **102**, 333-340 (2005).
33. Tan, F.L. et al. The gene expression fingerprint of human heart failure. *Proc Natl Acad Sci U S A* **99**, 11387-11392 (2002).
34. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21**, 171-178 (2005).
35. Tan, Y., Shi, L., Tong, W. & Wang, C. Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data. *Nucleic Acids Res* **33**, 56-65 (2005).

36. Kerr, M.K. & Churchill, G.A. Experimental design for gene expression microarrays. *Biostatistics* **2**, 183-201 (2001).

37. Wellmer, F., Riechmann, J.L., Alves-Ferreira, M. & Meyerowitz, E.M. Genome-wide analysis of spatial gene expression in Arabidopsis flowers. *Plant Cell* **16**, 1314-1326 (2004).

38. Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. & Tsui, K.W. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol* **8**, 37-52 (2001).

39. Cui, X., Hwang, J.T, Qui, J, Blades, N.J, Churchill, G.A Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 59-75 (2005).

40. Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P. & Rubin, E.M. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res* **10**, 2022-2029 (2000).

41. Norris, A.W. & Kahn, C.R. Analysis of gene expression in pathophysiological states: balancing false discovery and false negative rates. *Proc Natl Acad Sci U S A* **103**, 649-653 (2006).

42. Irizarry, R.A. et al. Multiple-laboratory comparison of microarray platforms. *Nat Methods* **2**, 345-350 (2005).

43. Hughes, T.R. et al. Functional discovery via a compendium of expression profiles. *Cell* 102, 109-126 (2000).

FIGURE LEGENDS

Figure 1: Concordance for inter-site comparisons. Each panel represents the POG results for a commercial platform comparing inter-site consistency in terms of DEGs between samples B and A. For each of the six gene selection methods, there are three possible inter-site comparisons: S1-S2, S1-S3, and S2-S3. Therefore, each panel consists of 18 POG lines that are colored based on gene ranking/selection method. Results shown here are based on the entire set of "12,091" genes commonly mapped across the microarray platforms without noise (absent call) filtering. Results are slightly improved when the analyses are performed using the subset of genes that are commonly detectable by the two test sites, as shown in the Supplementary Figure 1. The x-axis represents the number of selected DEGs, and the y-axis is the percentage (%) of genes common to the two gene lists derived from two test sites at a given number of DEGs.

Figure 2: Concordance for cross-platform comparisons. Panel a: Based on the "12,091" data set (without noise filtering); Panel b: Based on subsets of genes commonly detected ("Present") by two platforms. For each platform, the data from test site1 are used for cross-platform comparison. Each POG line corresponds to comparison of the DEGs from two microarray platforms using one of the six gene selection methods. There are ten platform-platform comparison pairs, resulting in 60 POG lines for each panel. The x-axis represents the number of selected DEGs, and the y-axis is the percentage (%) of genes common to the two gene lists derived from two platforms at a given number of DEGs. POG lines circled by the blue oval are from FC based gene selection methods with or without a *P* cutoff, whereas POG lines circled by the teal oval are from *P* based gene selection methods with or without an FC cutoff. Shown here are results for comparing sample B and sample A.

Figure 3: Concordance between microarray and TaqMan® assays. Each panel represents the comparison of one microarray platform to TaqMan® assays. For each microarray platform, the data from test site 1 are used for comparison to TaqMan® assays. Each POG line corresponds to comparison of the DEGs from one microarray platform and those from the TaqMan® assays using one of the six gene selection methods. The x-axis represents the number of selected DEGs, and the Y-axis is the percentage (%) of genes common to DEGs derived from a microarray platform and those from TaqMan® assays. Shown here are results for comparing sample B and sample A using a subset of genes that are detectable by both the microarray platform and TaqMan® assays. Results based on the entire set of 906 genes are provided in Supplementary Figure 2.

Figure 4: Inter-site reproducibility of log2 FC and log2 t-statistic. a: log2 FC of site 1 versus log2 FC of site 2; b: log2 t-statistic of test site 1 versus log2 t-statistic of test site 2; and c: log2 FC of test site 1 versus log2 t-statistic of test site 1. Shown here are results for comparing sample B and sample A for all "12,091" genes commonly probed. The inter-site reproducibility of log2 FC (a) is much higher than that of log2 t-statistic (b). The relationship between log2 FC and log2 t-statistic from the same test site is non-linear and the correlation appears to be low (c).

Figure 5: Inter-site concordance based FC, t-test, Wilcoxon rank-sum test, and SAM. Affymetrix data on samples A and B from site 1 and site 2 for the "12,091" commonly mapped genes were used. No flagged ("Absent") genes were excluded in the analysis. For the Wilcoxon rank-sum tests, there were many ties, *i.e*., many genes exhibited the same level of statistical significance because of the small sample sizes (five replicates for each group). The tied genes from each test site were broken (ranked) by random ordering. Concordance between genes selected completely by random choice is shown in red and reaches 50% when all candidate genes are declared as differentially expressed. SAM improves inter-site reproducibility compared to t-test, and approaches, but does not exceed that of fold-change.

Figure 6: Gene selection and percentage of agreement in gene lists in simulated data sets. Illustrations of the effect of biological context, replicate CV distribution, gene list size, and gene selection rules/methods on the reproducibility of gene lists using simulated microarray data. In some sense, these three graphs represent some extremes as well as typical scenarios in differential expression assays. However, FC sorting with low *P* thresholds (0.001 or 0.0001; *light and medium gray boxes*) consistently performed better overall than the other rules, even when FC ranking or *P* ranking by itself did not perform as well.

SUPPLEMENTARY INFORMATION

Supplementary Figure 1: Concordance for inter-site comparisons based on genes commonly detectable by the two test sites compared. Each panel represents the POG results for a commercial platform comparing inter-site consistency in terms of DEGs between samples B and A. For each of the six gene selection methods, there are three possible inter-site comparisons: S1-S2, S1-S3, and S2-S3. Therefore, each panel consists of 18 POG lines that are colored based on gene ranking/selection method. The x-axis represents the number of selected DEGs, and the y-axis is the percentage (%) of genes common to the two gene lists derived from two test sites at a given number of DEGs.

Supplementary Figure 2: Concordance for inter-site comparison with samples C and D. The largest fold change between samples C and D is small (three-fold). For each platform, DEG lists from sites 1 and 2 are compared. Analyses are performed using the subset of genes that are commonly detectable by the two test sites.

Supplementary Figure 3: Concordance between microarray and TaqMan® assays without noise-filtering. Each panel represents the comparison of one microarray platform to TaqMan® assays. The x-axis represents the number of selected DEGs, and the y-axis is the percentage (%) of genes common to DEGs derived from a microarray platform and those from TaqMan® assays. Shown here are results for comparing sample B and sample A using the entire set of 906 genes for which TaqMan® assay data are available.

Supplementary Figure 4: Concordance between FC and *P* based gene ranking methods ("12,091 genes"; test site 1). Each POG line represents a platform using data from its first test site. The x-axis represents the number of selected DEGs, and the y-axis is the percentage (%) of genes common in the DEGs derived from FC and *P* ranking. Shown here are results for comparing sample B and sample A for all "12,091" genes commonly probed. When a smaller number of genes (up to a few hundreds or thousands) are selected, POG for cross selection method comparison (FC vs. *P*) is low. For example, there are only about 50% genes in common for the top 500 genes selected by FC and *P* separately, indicating that FC and *P* rank order DEGs dramatically differently. When the number of selected DEGs increases, the overlap between the two methods increases, and eventually approach to 100% in common, as expected. The low concordance between FC- and *P*-based gene ranking methods is not unexpected considering their different definitions.

Supplementary Figure 5: Volcano plot illustration of joint FC and *P* gene selection rule. Genes in sectors A and C are selected as significant. The colors correspond to the negative $\log_{10} P$ and $\log_2$ fold change values:
Red (●): $20<-\log_{10} P<50$ and $3<\log_2$ fold<9 or $-9<\log_2$ fold <-3
Blue (●): $10<-\log_{10} P<50$ and $2<\log_2$ fold<3 or $-3<\log_2$ fold<-2
Yellow (●): $4<-\log_{10} P<50$ and $1<\log_2$ fold<2 or $-2<\log_2$ fold<-1
Pink (●): $10<-\log_{10} P<20$ and $3<\log_2$ fold or $\log_2$ fold<-3
Light blue (●): $4<-\log_{10} P<10$ and $2<\log_2$ fold or $\log_2$ fold<-2
Light green (●): $2<-\log_{10} P<4$ and $1<\log_2$ fold or $\log_2$ fold<-1
Gray (●): $-\log_{10} P<2$ or $\log_2$ fold<1 and $\log_2$ fold>-1

Figure 1: Inter-site concordance.

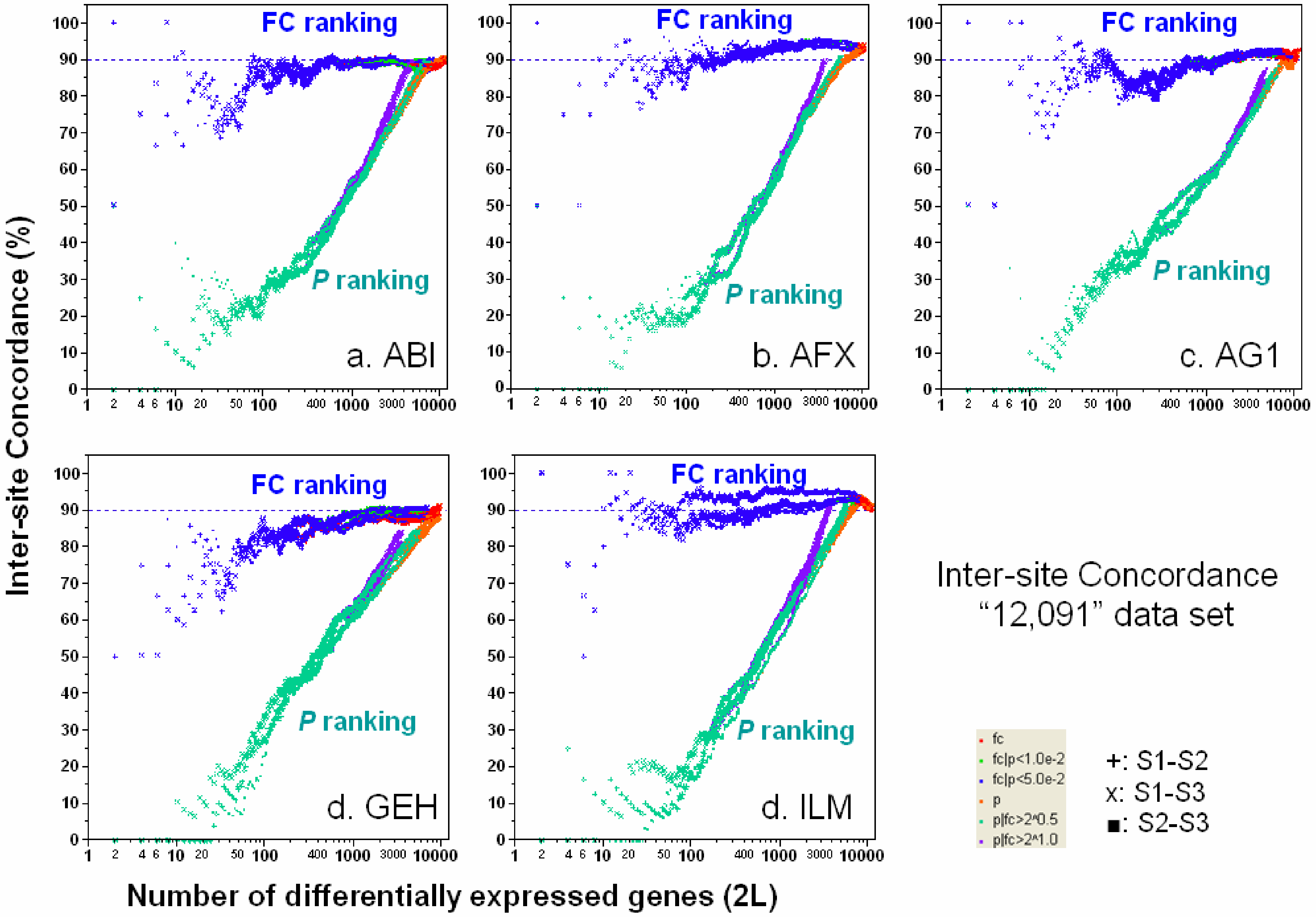


Figure 2: Cross-platform concordance.

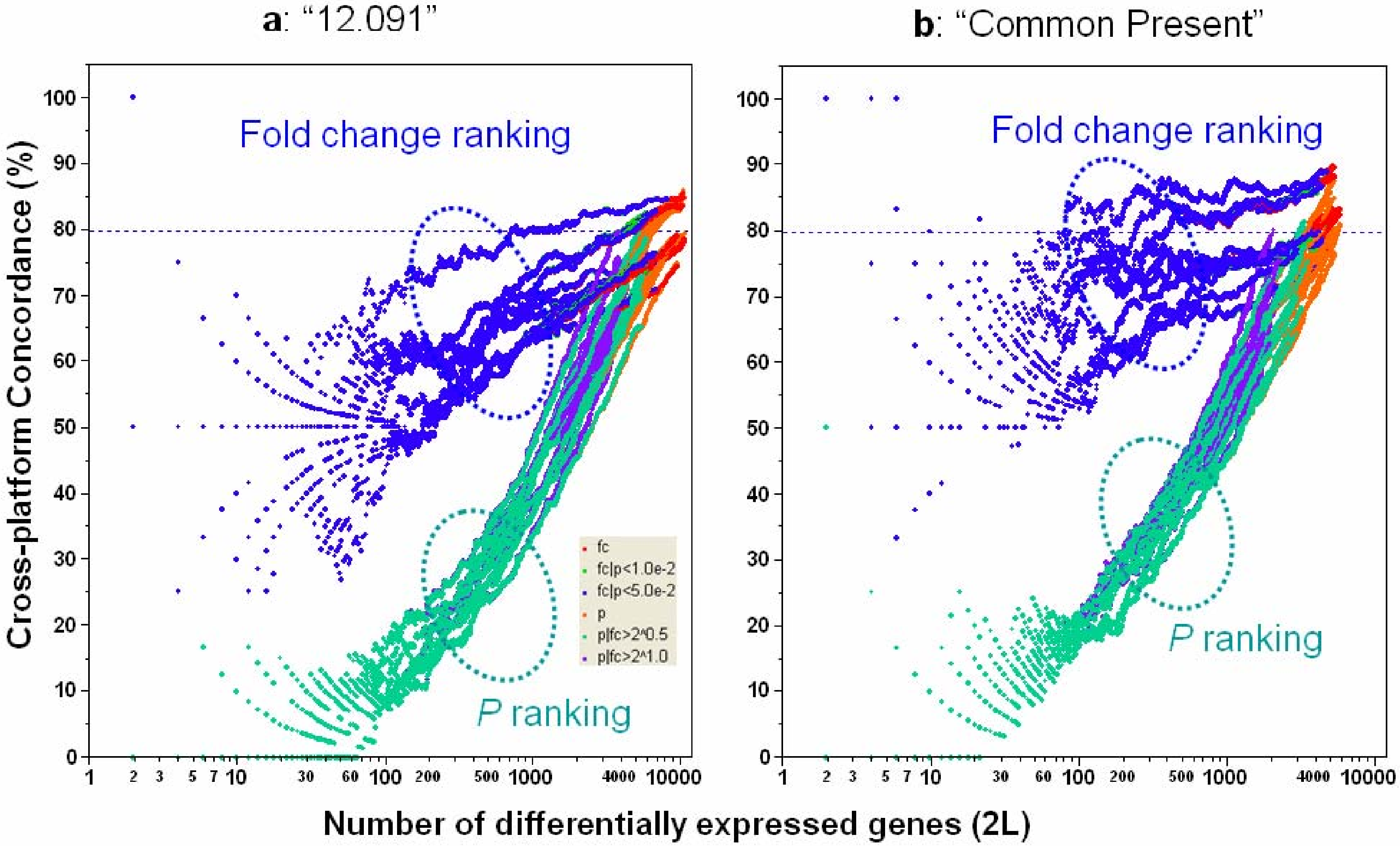
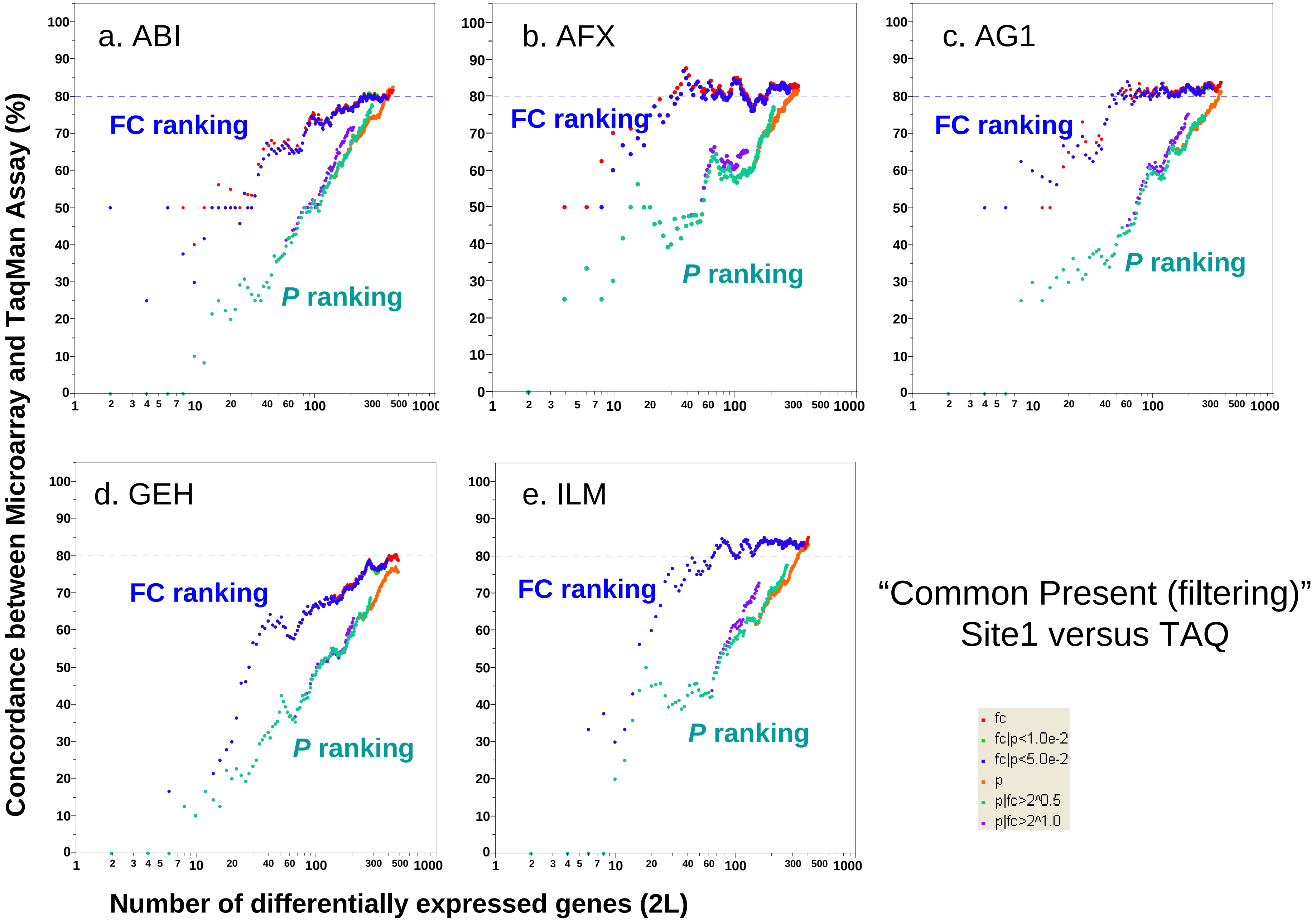


Figure 3: Concordance between microarray and TaqMan® assays with noise-filtering.



“Common Present (filtering)”
Site1 versus TAQ

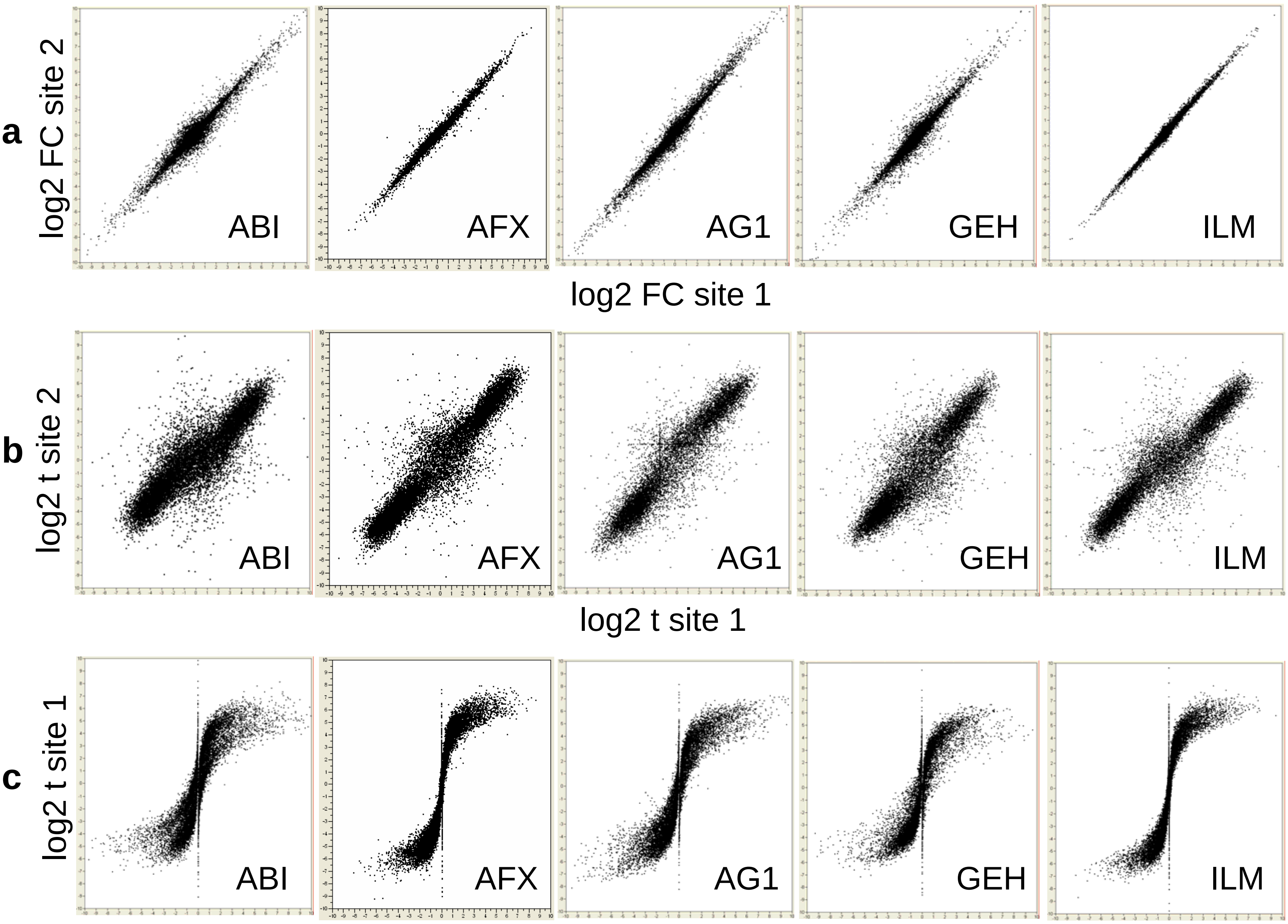
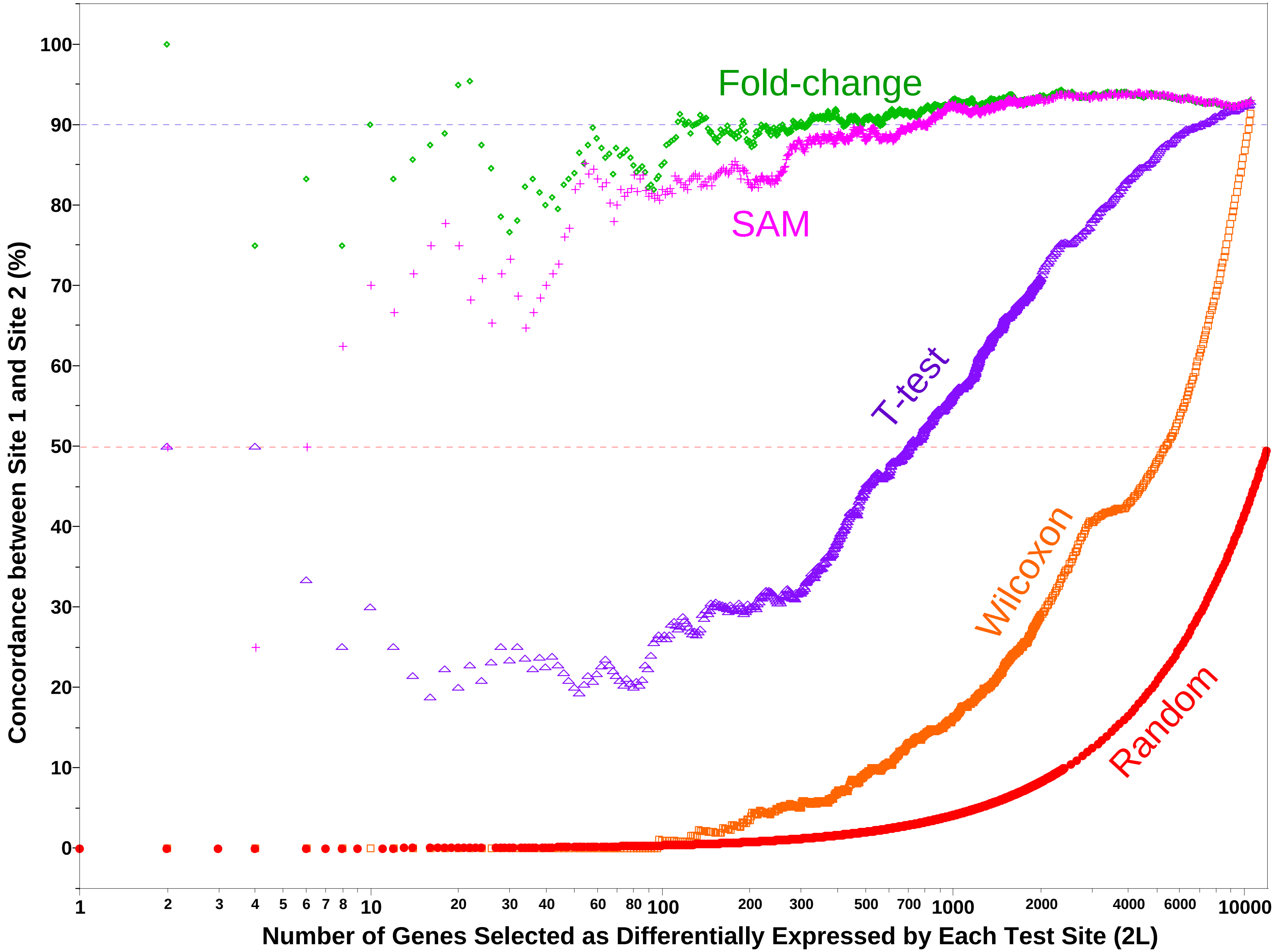


Figure 4: Inter-site reproducibility of log2 FC and log2 t-statistic.

Figure 5


Concordance between Site 1 and Site 2 (%)
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
10
20
30
40
60
80
100
200
300
500
700
1000
2000
4000
6000
10000
Number of Genes Selected as Differentially Expressed by Each Test Site (2L)
Fold-change
SAM
T-test
Wilcoxon
Random

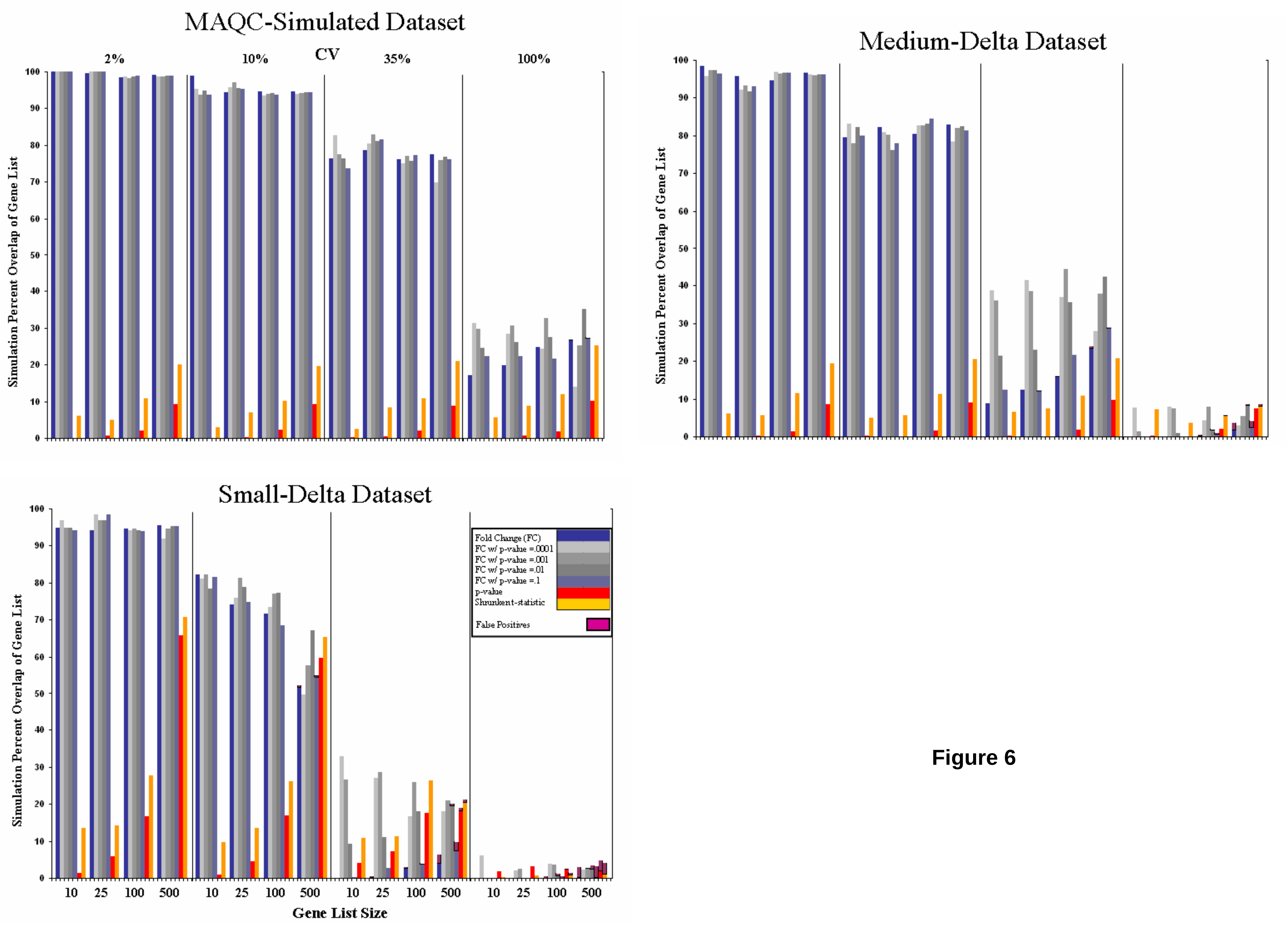
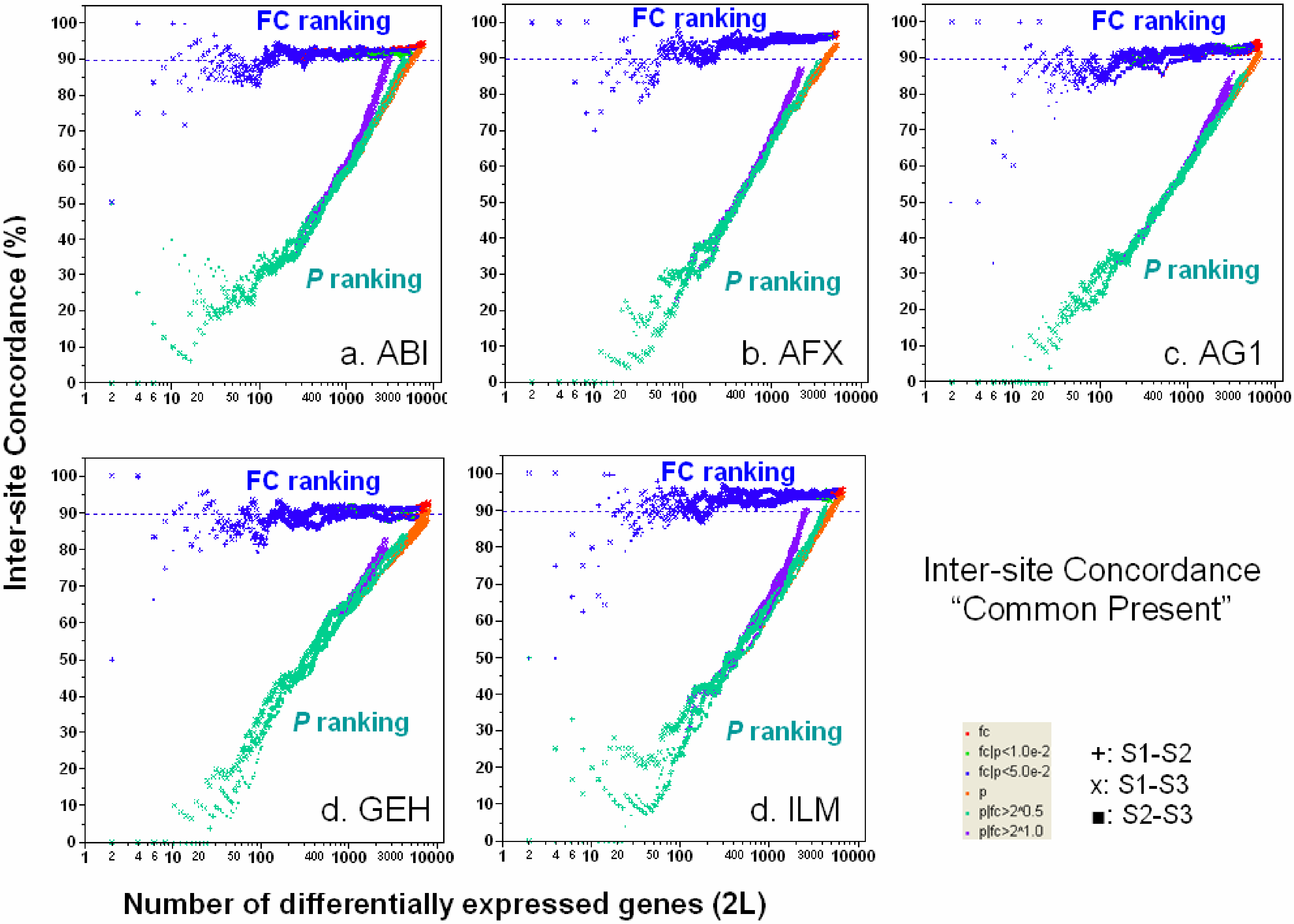
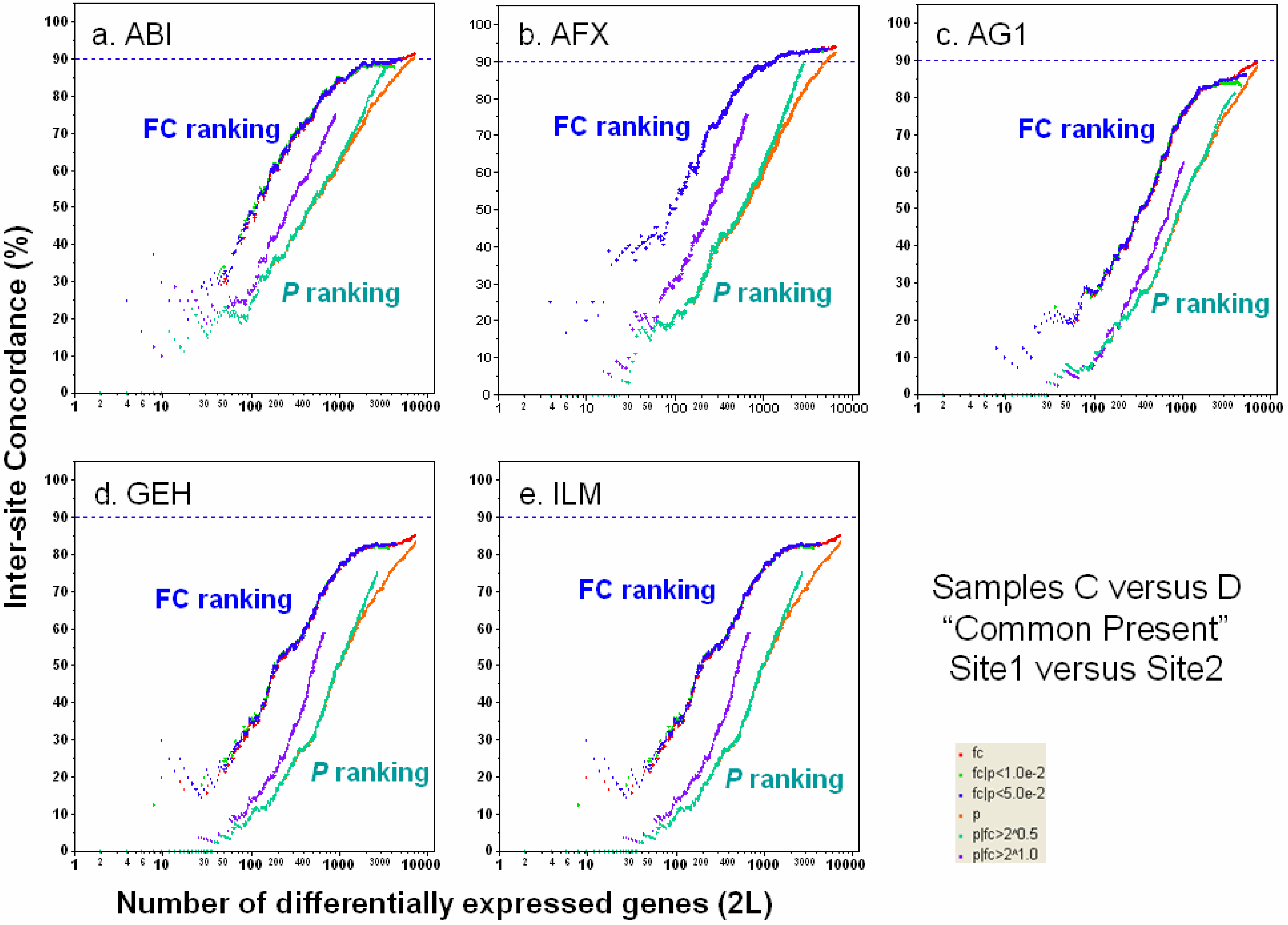

MAQC-Simulated Dataset
CV
2%
10%
35%
100%
Simulation Percent Overlap of Gene List
100
90
80
70
60
50
40
30
20
10
0
Medium-Delta Dataset
Simulation Percent Overlap of Gene List
Small-Delta Dataset
Simulation Percent Overlap of Gene List
Fold Change (FC)
FC w/ p-value =.0001
FC w/ p-value =.001
FC w/ p-value =.01
FC w/ p-value =.1
p-value
Shrunkent-statistic
False Positives
10
25
100
500
10
25
100
500
10
25
100
500
10
25
100
500
Gene List Size

Figure 6

Supplementary Figure 1: Inter-site concordance based on genes commonly detectable by the two test sites.

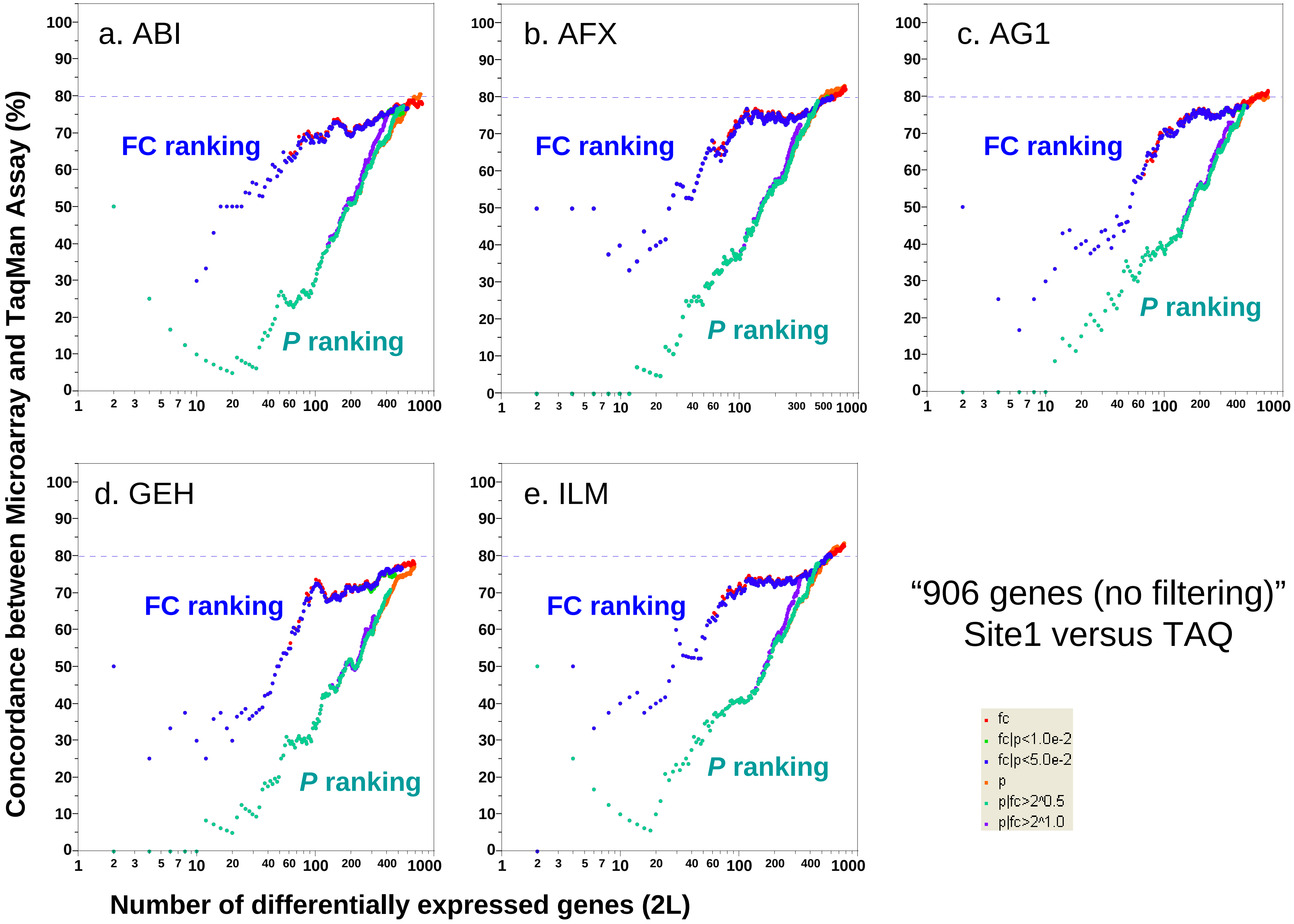


Supplementary Figure 2: Inter-site concordance for comparing samples C and D.



Samples C versus D
"Common Present"
Site1 versus Site2

Supplementary Figure 3: Concordance between microarray and TaqMan® assays without noise-filtering.



Supplementary Figure 4: Concordance between FC and *P*-value based gene ranking methods (“12,091 genes”; site 1).

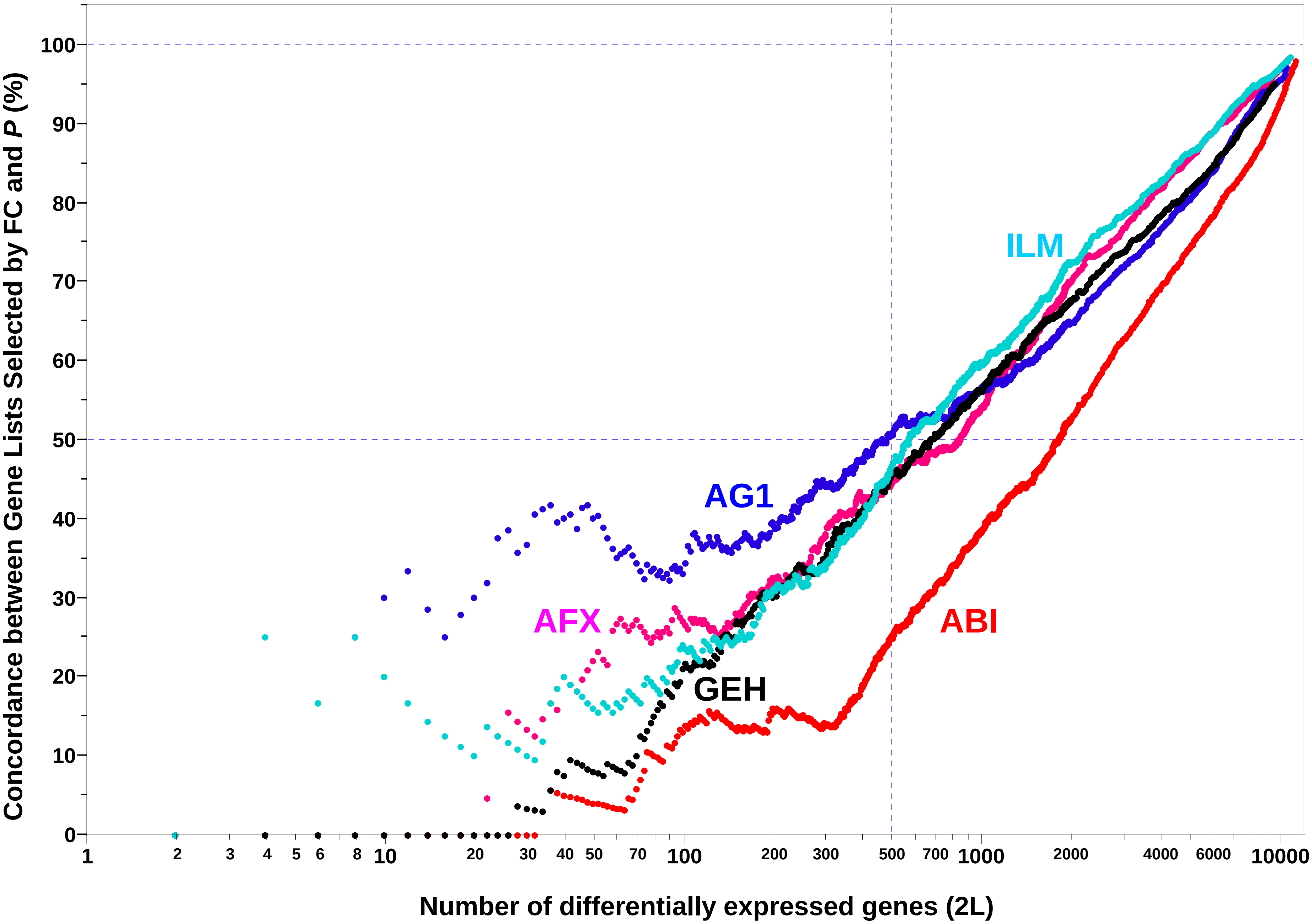
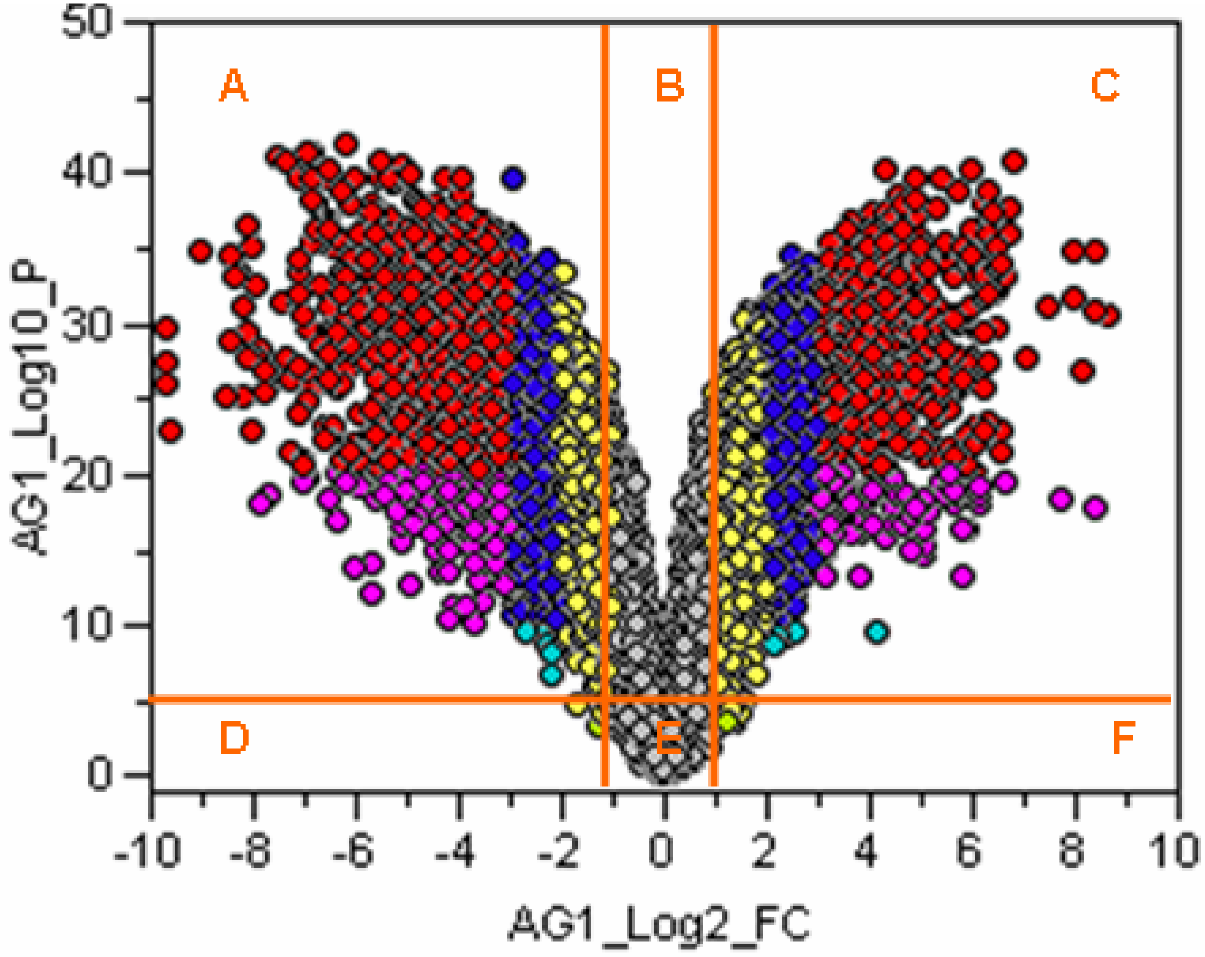


Figure 5: Volcano plot illustration of joint FC and *P*-value gene selection rule. Genes in sectors A and C are selected as significant.



DEPARTMENT OF HEALTH & HUMAN SERVICES

U.S. Food and Drug Administration
Jefferson Laboratories

Leming Shi, PhD, Principal Investigator
Phone: +1-870-543-7387 (Voice)
+1-870-543-7736 (Fax)
Leming.Shi@fda.hhs.gov
http://edkb.fda.gov/MAQC/

National Center for Toxicological Research
Division of Systems Toxicology
3900 NCTR Road
Jefferson, AR 72079-9502
U.S.A.

July 13, 2006

Manuscript Number: [MAQC MS-6]
Manuscript Title: *The reproducibility of lists of differentially expressed genes in microarray studies*
Authors: Shi L, Jones WD, *et al*.

Dear [Editor]:

Thank you very much for considering our manuscript and providing us an opportunity to address the concerns and comments raised by the three reviewers. We also appreciate the efforts of the first reviewer whose specific comments and suggestions provided valuable feedback helping us clarify the aims of the study and improve the manuscript.

We respectfully disagree with the assessment of the second reviewer that the use of the standard t-test (the most common statistical test used in microarray analysis) and the use of a careful simulation of microarray data to complement our mathematical analysis constitute “serious flaws in methodology”. The second reviewer’s comments did compel us to clarify the description of the data simulations to improve the manuscript. The third reviewer was gratuitously negative, and some of his/her comments were even sarcastic. It was clear to us that reviewer #3 did not carefully read our manuscript before drawing a conclusion and making bold statements throughout this review. In sharp contrast to the other two reviewers, reviewer #3 did not even consider the topic of our work an issue. We wonder whether or not we received a fair and unbiased review from reviewer #3. In the three separate files attached, we provide a point-by-point response to the three reviewers’ concerns in which the original comment is in Arial font and our response immediately follows in Roman font. Modifications to the manuscript are colored in blue when possible.

We do not conclude that the simple t-tests should not be used for the analysis of microarray data. In fact, we use t-test for the analysis of microarray data in our manuscript in order to generate a significance measure for the fold-change of each gene. **We do not think t-test itself is wrong in microarray data analysis**. Instead, **the problem is the use of t-statistic (*P*) as the <u>ranking</u> criterion** for the identification of differentially expressed genes, with or without a FC threshold. Our work demonstrates the need for a shift from the common practice of selecting differentially expressed genes solely by ranking a statistical significance measure (*e.g.*, t-statistic) to an approach that emphasizes fold-change, a quantity measured by microarray technology. We also would like to bring to your attention that none of the reviewers’ favorite methods (*e.g.*, SAM and rank test) performed as well as simple fold-change ranking in selecting reproducible DEGs, as demonstrated with additional analyses shown in the point-by-point responses and the revised manuscript.

All three referees seems to share a statistical perspective, a background shared by many of the MAQC participants who include well established statisticians and applied mathematicians as well as biologists. The first two reviewers touch on issues that arose during the course of the project that were extensively discussed and considered by the more than fifty study participants. We are grateful to the reviewers for their acknowledgement of the complexity of the study. The negative comments from the reviewers are testimony to the many interesting challenges facing the community in dealing with reproducibility in terms of lists of DEGs. The overly negative tone of reviewers #2 and #3 should simply reinforce how provocative and significant our results are. The questions regarding the importance and relevance of reproducibility in evaluating data quality are greatly appreciated as they led us to more clearly define this issue and give it greater emphasis in the Introduction and Discussion.

The initial presentation on POG results during MAQC face-to-face meetings literally enraged many statisticians on the MAQC consortium. Some of them seriously doubted that we would be able to produce a manuscript that was acceptable to both biological/chemical and statistical communities. Several of them were even to the point of removing their names from the author list. Through heavy debate that extended over several months, we achieved what could be considered as a breakthrough in understanding: that reproducibility is actually a third dimension needing optimization along with classical sensitivity and specificity. Although we still have some differences of opinion regarding the relative importance of these three dimensions and the POG metric, just recognizing reproducibility as a third factor -- especially from a regulatory perspective -- has been a very important outgrowth of our interactions. In retrospect, it seems that we did not make this point clear enough in the original submission of this work, and in this sense the critical replies of the statistically-oriented referees are understandable. We have tried to make this much more prominent in the revised manuscript and our point-by-point replies, thereby hoping to assuage their concerns.

New statistical methods for the identification of DEGs continue to appear in the scientific literature. In fact, the variety of existing and emerging methods has caused some confusion in the research community. These methods are typically promoted in terms of improved sensitivity (power) under various assumptions or conditions while retaining nominal rates of specificity. Reproducibility is a fundamental requirement in scientific experiments and clinical contexts, but is seldom emphasized in microarray literature. Reproducibility is equally if not more important than sensitivity and specificity in certain experimental and clinical contexts. Until recently ***reproducibility has not adequately been used as an essential criterion for evaluating the pros and cons of statistical methods for identifying DEGs***.

The focus of our work is the reproducibility of lists of putatively differentially expressed genes in microarray studies. The apparent lack of reproducibility of such DEGs has been used as scientific evidence to criticize microarray technology. Despite the availability of numerous statistical methods for the identification of DEGs, the simple t-statistic (and slight variations) is arguably still the most widely used test statistic, and many of the various methods that exist to create lists of DEGs primarily improve upon the inference from this basic test statistic. Our work was not intended to serve as a comprehensive performance survey of different statistical procedures; such a survey is not within the scope of our work and by itself is another large study.

We are not claiming that a concurrent use of fold-change ranking combined with a *P* threshold is the ultimate and best way of identifying DEGs in all circumstances. Instead, it appears to be a reasonable, straightforward (baseline) analysis procedure that can be used to enhance the reproducibility of DEG lists.

A good understanding of these factors is critical for the peer reviewers and readers to better appreciate the urgency of the critical issue addressed in this work and its important contribution to the microarray field. We hope that, despite the controversial nature of this work, you will consider the potential positive impact of increasing the salience of the reproducibility issue through the publication of this work.

We would like to alert you that since the original submission of this manuscript, we discovered that the normalization of the Affymetrix data was not performed exactly according to the manufacturer's recommended procedure, which was an established guideline for the MAQC project. For this reason, we have renormalized the Affymetrix data and all figures and tables impacted by this modification have been regenerated (specifically, Figure 1b, Figure 2, Figure 3b, Figure 4 (AFX panel), and Supplementary Figures S1b, S2b, S3b, and S4). In addition, the text was updated when necessary. While this modification had minor impact on the results and thereby the visualization of these data, there was no impact to the findings and conclusions that were represented in our original version of the manuscript.

Again, we appreciate your time and effort in considering our manuscript for publication and hope that you will find that the revised manuscript, which thoroughly addresses the reviewers' comments and suggestions, is suitable for publication in *[your journal]*.

Yours sincerely,

Leming Shi, Ph.D.
FDA/NCTR

Point-by-Point Response to Peer Reviewer #1' Comments

Manuscript: [MAQC MS-6]
Title: The reproducibility of lists of differentially expressed genes in microarray studies
Corresponding author: Leming Shi (leming.shi@fda.hhs.gov)
Date: July 13, 2006

General Response to Peer Reviewer #1

New statistical methods for the identification of differentially expressed genes (DEGs) continue to appear in the scientific literature. In fact, the variety of existing and emerging methods has caused some confusion in the research community. These methods are typically promoted in terms of improved sensitivity (power) under various assumptions or conditions while retaining nominal rates of specificity. Reproducibility is a fundamental requirement in scientific experiments and clinical contexts, but is seldom emphasized in microarray literature. Reproducibility is a critical third dimension that is distinct from specificity and sensitivity. It is equally if not more important than sensitivity and specificity in certain experimental and clinical contexts. Until recently reproducibility has not adequately been used as an essential criterion for evaluating the pros and cons of statistical methods for identifying DEGs. Demonstrating reproducible performance is critical to the acceptance of microarray-based data in clinical and regulatory environments. We anticipate that the editors of [*the journal*] will consider the potential positive impact on the scientific community in considering this work for publication.

We would like to emphasize the following:

1. The focus of our work is the reproducibility of lists of putatively differentially expressed genes (DEGs) in microarray studies.
2. The apparent lack of reproducibility of such DEGs has been used as scientific evidence to criticize microarray technology.
3. Despite the availability of numerous statistical methods for the identification of DEGs, the simple t-statistic (and slight variations) is arguably still the most widely used test statistic, and many of the various methods that exist to create lists of DEGs primarily improve upon the inference from this basic test statistic. This includes the simple unmodified two-sample t-test, Bonferroni and step-up/step-down procedures applied to the t-test, and others. We also note that a ranking criterion based on the t-statistic or the *P*-value derived from it is equivalent.
4. Statistical significance (*P*) derived from the simple two-group t-test has historically been widely used as the only criterion to identify DEGs, often with disappointing results related to reproducibility when it has been measured.
5. Our work was not intended to serve as a comprehensive performance survey of different statistical procedures. Such a survey is not within the scope of our work and by itself constitutes a separate large study.
6. We are NOT claiming that a concurrent use of FC ranking combined with a *P* threshold is the ultimate and best way of identifying DEGs in all circumstances. Instead, it appears to be a reasonable, straightforward (baseline) analysis procedure that can be used to enhance the

reproducibility of DEG lists, especially if the microarray-based experiment is to be reviewed in a clinical or regulatory environment.

A good understanding of these factors is critical for the peer reviewers, editors, and readers to better appreciate the urgency of the issue being addressed in this work and its important contribution to the microarray field.

To clarify the overall goals of this paper and of the MAQC study as a whole, we have made some modifications throughout the manuscript to emphasize the focus on the reproducibility of lists of differentially expressed genes. We also now provide a self-contained description of the design of the MAQC study. For example, we have modified the Abstract to read:

Abstract:
Reproducibility is a fundamental requirement in scientific experiments and clinical contexts. Recent publications raise concerns about the reliability of microarray technology because of the apparent lack of agreement between lists of differentially expressed genes (DEGs). In this study we demonstrate that (1) such discordance may stem from ranking and selecting DEGs solely by statistical significance (*P*) derived from widely used simple *t*-tests; (2) when fold change (FC) is used as the ranking criterion, the lists become much more reproducible, especially when fewer genes are selected; and (3) the instability of short DEG lists based on *P* cutoffs is an expected mathematical consequence of the high variability of the *t*-values. We recommend the use of FC ranking plus a non-stringent *P* cutoff as a baseline practice in order to generate more reproducible DEG lists. The FC criterion enhances reproducibility while the *P* criterion balances sensitivity and specificity.

Additionally, we added a paragraph and modified a sentence in the Introduction that clearly states that:

“The MAQC study was specifically designed to address these previously identified sources of variability in DEG lists. Two very different RNA samples, Stratagene Universal Human Reference RNA and Ambion Human Brain Reference RNA, with thousands of differentially expressed genes, were prepared in sufficient quantities and distributed to three different laboratories for each of the five different commercial whole genome microarray platforms participating in the study. For each platform, each sample was analyzed using five technical replicates with standardized procedures for sample processing, hybridization, scanning, data acquisition, data preprocessing, and data normalization at each site. The probe sequence information was used to generate a stringent mapping of genes across the different platforms and 906 genes were further analyzed with TaqMan® assays using the same RNA samples.

A careful analysis of these MAQC data sets, along with numerical simulations and mathematical arguments, demonstrates that the reported lack of reproducibility of DEG lists can be attributed in large part to identifying DEGs from simple *t*-tests without consideration of FC when sample numbers are small. The finding holds for intra-laboratory, inter-laboratory, and cross-platform comparisons independent of sample pairs and normalization methods, and is increasingly apparent with decreasing number of genes selected.”

Point-by-Point Response to Reviewer #1

Note: The reviewer's original comments/questions are in Arial font and the authors' response is in Roman font. The reviewer's comments are numbered for convenience in authors' response. Text changes to the manuscript are indicated in blue fonts.

1. It is well known that the microarray-based gene expression profiling experiments can result in very different lists of differentially expressed genes (DEGs), depending on what microarray platform is used and on which laboratory (or individual experimenter) performs the microarray experiments. Since differences in the lists of genes reported as being differentially expressed for a given type of experiment is at times a topic of very hot debate when expression profiling studies conflict, this study sheds light on the reasons for possible observed differences.

Response:

We appreciate the reviewer's recognition of the importance of the topic that has been addressed in our manuscript on the reproducibility of lists of differentially expressed genes (DEGs) and the general assessment of our manuscript that "*sheds light on the reasons for possible observed differences [in DEG lists]*".

2. In this manuscript the authors focus on the reproducibility of lists of DEGs. In particular, they claim that such discordance in DEG lists is due to "ranking and selecting DEGs solely by statistical significance such as by P from simple t-tests". Although the authors state that their objective is to explain a major reason for lack of reproducibility in lists of DEGs, they actually conclude with recommendations to use a combination of fold-change ranking and P-value cutoff, but they haven't gone so far as to discuss how exactly such a combination would be set based upon an optimization of sensitivity and specificity; it is not appropriate to simply set these cutoffs based on optimizations of percentage of overlapping genes (POG). Simply improving the reproducibility of DEGs is not of itself what the scientific community needs most. Instead, there is a need for more accurate lists of DEGs, and these lists may differ from platform to platform, or laboratory to laboratory. Optimizing for the POG is in essence simply identifying the "lowest common denominator". Using approaches that simply make the lists of DEGs more uniform across platforms and laboratories may reduce the number of biologically significantly DEGs that are reported, and that could be a real loss in terms of identification of important DEGs. Although it is interesting to see what kinds of combined cutoffs may improve the reproducibility of lists of DEGs, this gets around the issue of how to accurately report the biologically significantly DEGs. Indeed, many important, biologically significantly DEGs may be changed at subtle fold change (FC) levels, including those with less than 2-fold changes (see Hughes et al., Cell, 2000 Jul 7;102(1):109-26). The authors actually conclude with a recommendation that "the practice of using P alone for gene selection should be discouraged". What is the tradeoff between loss in sensitivity &

specificity, and DEG list reproducibility obtained by concurrent use of fold-change ranking and P value?

Response:

Previously, the focus of microarray data analysis has been on specificity and sensitivity. Reproducibility is a third, critical dimension that is distinct from specificity and sensitivity, and is equally if not more important in certain experimental and regulatory contexts. Unfortunately, reproducibility has not been used as an essential criterion for evaluating the pros and cons of statistical methods for identifying DEGs.

The reviewer states that *"they haven't gone so far as to discuss how exactly such a combination would be set based upon an optimization of sensitivity and specificity*." This point is strongly related to another question that the reviewer asked later in the same paragraph, and so our response to this statement and the later question is combined and provided at the end of this section.

While we agree with the reviewer that "*there is a need for more accurate lists of DEGs*", the second half of the sentence "*and these lists may differ from platform to platform, or laboratory to laboratory*" appears to imply that it is normal for different platforms or laboratories to generate different DEG results from the same RNA samples. In discovery, it is reasonable to continually encounter partial answers which may lead to further investigations or spawn an experiment that ultimately leads to a larger truth. However, there are other contexts which we have stated previously where such variation is at best undesired and at worst unacceptable.

We agree with the reviewer that genes with *"subtle fold change (FC)"* may be indeed biologically important. However, what our study shows is that genes with smaller fold changes from one platform or laboratory are, in general, less likely reproducible in another laboratory with the same or different platforms. This is in fact an issue of the assumptions and criteria used to establish the detection limit for FC estimation by different microarray technologies. Not unlike other methodologies for genes identified at the threshold of detection and/or reflecting small perturbations in biological levels and/or based on small number of samples, we may have to acknowledge the reality of variability in the results. Importantly, in a microarray study there are usually many genes on an array representing genetic networks that can be utilized in confirmatory work to build confidence in a finding. A "screening or filtering" procedure that enhances reproducibility is practical and essential for derivation of optimized robust signatures for specific applications, *e.g*. diagnostics. A real challenge for microarrays is the development of improved methods that can reliably and repeatedly differentiate truly biologically important genes with small FCs from those genes with small FCs as a result of random fluctuations or by chance.

We believe that truly differentially expressed genes should be more likely identified as differentially expressed by different platforms and from different laboratories than those genes with no differential expression between sample groups. In the microarray field, we usually do not have the luxury of knowing the "truth" in a given study. Therefore, it is not surprising that most microarray studies and data analysis protocols have not been adequately evaluated against the "truth". A reasonable surrogate of such "truth" could be the consensus of results from different microarray platforms, from different laboratories using the same platform, or from independent methods such as TaqMan® assays, as we have extensively explored in this study.

The reviewer revised an earlier stated concern and asked a very good question on the *"tradeoff between loss in sensitivity & specificity, and DEG list reproducibility obtained by concurrent use of fold-change ranking and P value"* as well as the concern that *"Optimizing for the POG is in essence simply identifying the 'lowest common denominator.'"* We feel that reproducibility is not "simple" as it is the subject of so many scientific papers. The focus of our study has been the exploration of the issue of reproducibility, the apparent lack of which has been used at times to question the reliability of microarray technologies. The limitation of the scope of our study prevented us from going any further on the tradeoff between loss of sensitivity & specificity and the gain in DEG reproducibility. However, there are cases where there is no loss or tradeoff in sensitivity or specificity when one sets a limit on the number of genes for further consideration (as is often done in many biological studies) when faced with hundreds or thousands of putative DEGs. Our recommendation of a combined FC ranking and *P*-value cutoff for identifying DEGs enhances reproducibility due to the use of FC, the quantity measured by microarrays, in ranking genes and the use of a reasonable *P*-value cutoff to address significance and specificity/sensitivity. If sensitivity is the primary focus, a less stringent *P*-value cutoff should be applied. A more stringent *P*-value cutoff increases specificity. Ultimately the trade-off one accepts is based on the specific question one is asking or the need being addressed. For instance, in diagnostic development, robust signatures that are highly reproducible and accurate may be developed that completely omit biological information that is at the limit of detection or representing small changes in expression. Although that information may be of further interest in the realm of drug target discovery, the signature used in the diagnostic assay serves a different purpose. Finally, identifying the "lowest common denominator" has potentially both negative and positive attributes depending on context. In our context, this attribute would be positive, implying that there is enhanced probability of independent confirmation of the result.

Actions taken:

(1) We have revised the text in Abstract, Introduction, and Discussion to further emphasize the focus of the study: the reproducibility of DEG lists.

(2) A paragraph has been added to Discussion (p.10) regarding "subtle fold change" with a citation to the work of TR Hughes *et al*. mentioned by the reviewer:

"This study shows that genes with smaller expression fold changes generated from one platform or laboratory are, in general, less reproducible in another laboratory with the same or different platforms. However, it should be noted that genes with small fold changes may be biologically important[43]. When a fixed FC cutoff is chosen, sensitivity could be sacrificed for reproducibility. Alternatively, when a high P cutoff (or no P cutoff) is used, specificity could be sacrificed for reproducibility. Ultimately, the acceptable trade-off is based on the specific question being asked or the need being addressed. When searching for a few reliable biomarkers, high FC and low P cutoffs can be used to produce a highly specific and reproducible gene list. When identifying the components of genetic networks involved in biological processes, a lower FC and higher P cutoff

can be used to identify larger, more sensitive but less specific, gene lists. In this case additional biological information about putative gene functions can be incorporated to identify reliable gene lists that are specific to the biological process of interest."

(3) A paragraph has been added to Discussion (p.10) regarding "accurate list of DEGs":

"Truly differentially expressed genes should be more likely identified as differentially expressed by different platforms and from different laboratories than those genes with no differential expression between sample groups. In the microarray field, we usually do not have the luxury of knowing the "truth" in a given study. Therefore, it is not surprising that most microarray studies and data analysis protocols have not been adequately evaluated against the "truth". A reasonable surrogate of such "truth" could be the consensus of results from different microarray platforms, from different laboratories using the same platform, or from independent methods such as TaqMan® assays, as we have extensively explored in this study."

Some more specific comments follow below:

3. The aspect of this study that examines the impact of CV (coefficient of variation) on reproducibility due to the combined use of FC and P-value cutoff is useful. I'd like to see a more in-depth analysis of the tradeoff between loss in sensitivity and DEG list reproducibility obtained by concurrent use of fold-change ranking and P value, at various CV values? The authors even make an intriguing statement that touches on this issue, in discussing analysis of their simulated data: "Although P ranking generally resulted in very low POG, a false positive was rarely produced, even for a list size of 500 (data not shown)."

Response:

The simulation part of the study was not designed to examine trade-offs between sensitivity and DEG list reproducibility. For example, sensitivity as defined by

Sensitivity = # true positives / (# true positives + # false negatives)

would be very similar for most methods as the number of false negatives would tend to be very large in all cases. That is, each simulation scenario had thousands of true DEGs, but list sizes were restricted (10, 50, 100, 500) to the point that false negatives would dominate the sensitivity measure.

Perhaps a later paper could consider a related metric to sensitivity called PPV (or positive predicted value) which is

PPV = # true positives / (# true positives + # false positives)

4. Various investigators analyzing microarray data are aware that genes with low transcript levels are going to have more highly variable data, and thus impose a

> minimum signal intensity threshold requirement (beyond that of just the array manufacturers' minimal criteria for a "Present" call), so that any genes below that threshold are filtered from further consideration. Did the authors explore the effect of such filters on the reproducibility of lists of DEGs using just P-value cutoffs, or possibly calculating a P-value that is dependent upon the magnitude of the signal intensity?

Response:

We agree that it is a common practice of using a more or less arbitrary intensity threshold as a filter to exclude more variable data points in microarray data analysis. We compared the impact of this further filtering procedure in addition to the “majority present” filtering procedure on POGs for *P*-value and FC based gene selection methods. As expected, there is an increase in POG for FC based gene selection. In addition, FC ranking continues to produce much better POG results compared to t-test *P*-value. An example figure for inter-site comparing AFX test sites is provided for the reviewer’s information (Figure R1-A).

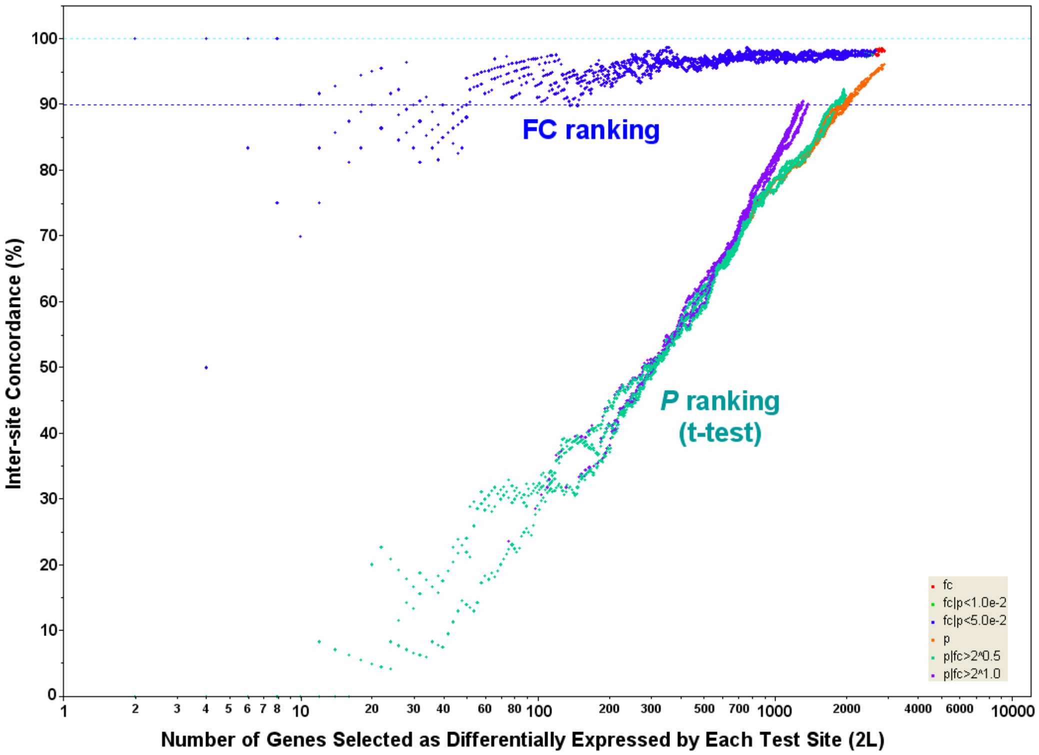


Figure R1-A. Gene selection methods determine the inter-site concordance of differentially expressed genes even __after aggressive data filtering__. Affymetrix data on samples A and B from the three test sites for the “12,091” commonly mapped genes were used. Two rounds of “filtering” were applied: first, genes that were called “Absent” by manufacturer’s criteria in the majority of replicates (three of five) for either sample A or sample B were excluded; second, an average intensity was calculated across the ten arrays for each of the remaining genes, and 50% of which with the lowest average intensity were further excluded.

5. No formula is provided to inform readers how POG is calculated. If the overlap between one list of 100 genes and another list of 200 genes is 50, then is the POG 50%, 25%, or 50/250 = 20%?

Response:

We thank the reviewer for bringing this to our attention. The formula for calculating POG is as follows:

$$POG = 100*(DD+UU)/2L$$

where DD and UU are the number of commonly down- or up-regulated genes, respectively, from the two lists, and L is the number of genes selected from the up- or down-regulation directionality. To overcome the kind of confusion (*i.e.*, different numbers for the denominator) raised by the reviewer, in our POG calculations we deliberately selected an equal number of genes, L, in the up- and down-regulation directionalities.

Actions taken:

(1) A paragraph has been added to Methods to describe the formula (p.13):

"The formula for calculating POG is: POG = 100(DD+UU)/2L, where DD and UU are the number of commonly down- or up-regulated genes, respectively, from the two lists, and L is the number of genes selected from the up- or down-regulation directionality. To overcome the confusion of different numbers for the denominator, in our POG calculations we deliberately selected an equal number of genes, L, in the up- and down-regulation directionalities."*

6. The authors show data for POG when comparing to a Taqman datase for 906 genes, but no information is provided for how this dataset was generated, in particular, what criteria were used to identify the list of genes differentially expressed according to the Taqman data.

Response:

We thank the reviewer for bringing this to our attention.

Actions taken:

(1) A sentence was modified on p.5 to more clearly describe the mapping of TaqMan assays to microarrays.

(2) A sentence was added on p.6 to describe how DEGs were identified for TaqMan data:

"There are four TaqMan® assays technical replicates for each sample and the DEGs for TaqMan® assays were identified using the same six gene selection procedures as those used for microarray data."

7. It would be highly informative to see a direct comparison of the use of false discovery rate (FDR) criterias versus results obtained using the authors' recommended use of a combination of fold-change and P-value cutoffs.

Response:

This is a good suggestion and was discussed intensively during the early stage of manuscript preparation. There are many methods for estimating FDR. Many of these methods start with test statistics which are either the t-statistic itself or a slightly modified version. However, typically these same FDR methods do not change the ranking order of genes determined by the *P*-values. They instead monotonically transform the *P*-value or test statistic information into the more useful q or FDR. So, in terms of DEG ranking order, many FDR methods are often equivalent to using the *P*-value. For SAM, which uses the modified or shrunken t-statistic, we have addressed this directly by including the shrunken t-statistic in our simulations and in some examples. We have seen that lists generated from the shrunken t-statistic generally improves reproducibility over using the *P*-value by itself but any list using FDR ranking from the shrunken t-statistic would still be less reproducible than using lists based on FC ordering with a *P*-value threshold for the scenarios studied and simulated.

Some minor comments follow below:

8. This manuscript needs to be proofread. There is at least one sentence that can not be understood: "Importantly, noise filtering does not either trend or magnitude of higher POG graphs for FC ranking compared with P-ranking."

Response:

We thank the reviewer for bringing this to our attention. The revised manuscript has been carefully proofread by several coauthors.

Actions taken:

(1) Several grammatical and typo errors have been corrected.

9. Why is the number of DEGs (the label for the x-axis in Figure 1) set equal to 2L? The number of up-regulated genes need not equal the number of down-regulated genes.

Response:

Please refer to our response to comment #5.

10. There is no explanation of the dotted lines in Figures 1, 2, and 3. There is also no explanation of what is shown in a dotted ovals in Figure 2.

Response:

We thank the reviewer for bringing this to our attention.

Actions taken:

(1) A sentence was added to the Fig. 3 legend (p.19) to describe the dotted POG lines (explanations on dotted POG lines were available for Figs. 1 and 2 in the original manuscript's figure legends):

"Each POG line corresponds to comparison of the DEGs from one microarray platform and those from the TaqMan® assays using one of the six gene selection methods."

(2) A sentence was added to the Fig. 2 legend (p.19) to describe the ovals that indicate the FC and *P*-value based POG lines:

"POG lines circled by the blue oval are from FC based gene selection methods with or without a P-value cutoff, and POG lines circled by the teal oval are from P-value based gene selection methods with or without an FC cutoff."

11. There is no explanation of the color scheme used in Figure 5.

Response:

We thank the reviewer for bringing this to our attention.

Actions taken:

(1) A sentence was added to the figure legend (p.21) to describe the color scheme. Please note that this figure is now moved to the supplementary information as Supplementary Figure 5:

"The colors correspond to the negative $\log_{10} P$ *and* $\log_2$ *fold change values.*
Red (●): $20<-\log_{10} P<50$ *and* $3<\log_2 fold<9$ *or* $-9< \log_2 fold <-3$
Blue (●): $10<-\log_{10} P<50$ *and* $2<\log_2 fold<3$ *or* $-3<\log_2 fold<-2$
Yellow (●): $4<-\log_{10} P<50$ *and* $1<\log_2 fold<2$ *or* $-2<\log_2 fold<-1$
Pink (●): $10<-\log_{10} P<20$ *and* $3<\log_2 fold$ *or* $\log_2 fold<-3$
Light blue (●): $4<-\log_{10} P<10$ *and* $2<\log_2 fold$ *or* $\log_2 fold<-2$
Light green (●): $2<-\log_{10} P<4$ *and* $1<\log_2 fold$ *or* $\log_2 fold<-1$
Gray (●): $-\log_{10} P<2$ *or* $\log_2 fold<1$ *and* $\log_2 fold>-1$*"*

Begin of original comments:

Reviewer #1(Remarks to the Author):

It is well known that the microarray-based gene expression profiling experiments can result in very different lists of differentially expressed genes (DEGs), depending on what microarray platform is used and on which laboratory (or individual experimenter) performs the microarray experiments. Since differences in the lists of genes reported as being differentially expressed for a given type of experiment is at times a topic of very hot debate when expression profiling studies conflict, this study sheds light on the reasons for possible observed differences.

In this manuscript the authors focus on the reproducibility of lists of DEGs. In particular, they claim that such discordance in DEG lists is due to "ranking and selecting DEGs solely by statistical significance such as by P from simple t-tests". Although the authors state that their objective is to explain a major reason for lack of reproducibility in lists of DEGs, they actually conclude with recommendations to use a combination of fold-change ranking and P-value cutoff, but they haven't gone so far as to discuss how exactly such a combination would be set based upon an optimization of sensitivity and specificity; it is not appropriate to simply set these cutoffs based on optimizations of percentage of overlapping genes (POG). Simply improving the reproducibility of DEGs is not of itself what the scientific community needs most. Instead, there is a need for more accurate lists of DEGs, and these lists may differ from platform to platform, or laboratory to laboratory. Optimizing for the POG is in essence simply identifying the "lowest common denominator". Using approaches that simply make the lists of DEGs more uniform across platforms and laboratories may reduce the number of biologically significantly DEGs that are reported, and that could be a real loss in terms of identification of important DEGs. Although it is interesting to see what kinds of combined cutoffs may improve the reproducibility of lists of DEGs, this gets around the issue of how to accurately report the biologically significantly DEGs. Indeed, many important, biologically significantly DEGs may be changed at subtle fold change (FC) levels, including those with less than 2-fold changes (see Hughes et al., Cell, 2000 Jul 7;102(1):109-26). The authors actually conclude with a recommendation that "the practice of using P alone for gene selection should be discouraged". What is the tradeoff between loss in sensitivity & specificity, and DEG list reproducibility obtained by concurrent use of fold-change ranking and P value?

Some more specific comments follow below:

The aspect of this study that examines the impact of CV (coefficient of variation) on reproducibility due to the combined use of FC and P-value cutoff is useful. I'd like to see a more in-depth analysis of the tradeoff between loss in sensitivity and DEG list reproducibility obtained by concurrent use of fold-change ranking and P value, at various CV values? The authors even make an intriguing statement that touches on this issue, in discussing analysis of their simulated data: "Although P ranking generally resulted in very low POG, a false positive was rarely produced, even for a list size of 500 (data not shown)."

Various investigators analyzing microarray data are aware that genes with low transcript levels are going to have more highly variable data, and thus impose a minimum signal intensity threshold requirement (beyond that of just the array manufacturers' minimal criteria for a "Present" call), so that any genes below that threshold are filtered from further consideration. Did the authors explore the effect of such filters on the reproducibility of lists of DEGs using just P-value cutoffs, or possibly calculating a P-value that is dependent upon the magnitude of the signal intensity?

No formula is provided to inform readers how POG is calculated. If the overlap between one list of 100 genes and another list of 200 genes is 50, then is the POG 50%, 25%, or 50/250 = 20%?

The authors show data for POG when comparing to a Taqman datase for 906 genes, but no information is provided for how this dataset was generated, in particular, what criteria were used to identify the list of genes differentially expressed according to the Taqman data.

It would be highly informative to see a direct comparison of the use of false discovery rate (FDR) criterias versus results obtained using the authors' recommended use of a combination of fold-change and P-value cutoffs.

Some minor comments follow below:

This manuscript needs to be proofread. There is at least one sentence that can not be understood: "Importantly, noise filtering does not either trend or magnitude of higher POG graphs for FC ranking compared with P-ranking."

Why is the number of DEGs (the label for the x-axis in Figure 1) set equal to 2L? The number of up-regulated genes need not equal the number of down-regulated genes.

There is no explanation of the dotted lines in Figures 1, 2, and 3. There is also no explanation of what is shown in a dotted ovals in Figure 2.

There is no explanation of the color scheme used in Figure 5.

End of original comments

Point-by-Point Response to Peer Reviewer #2' Comments

Manuscript: [MAQC MS-6]
Title: The reproducibility of lists of differentially expressed genes in microarray studies
Corresponding author: Leming Shi (leming.shi@fda.hhs.gov)
Date: July 13, 2006

General Response to Peer Reviewer #2

New statistical methods for the identification of differentially expressed genes (DEGs) continue to appear in the scientific literature. In fact, the variety of existing and emerging methods has caused some confusion in the research community. These methods are typically promoted in terms of improved sensitivity (power) under various assumptions or conditions while retaining nominal rates of specificity. Reproducibility is a fundamental requirement in scientific experiments and clinical contexts, but is seldom emphasized in microarray literature. Reproducibility is a critical third dimension that is distinct from specificity and sensitivity. It is equally if not more important than sensitivity and specificity in certain experimental and clinical contexts. Until recently reproducibility has not adequately been used as an essential criterion for evaluating the pros and cons of statistical methods for identifying DEGs. Demonstrating reproducible performance is critical to the acceptance of microarray-based data in clinical and regulatory environments. We anticipate that the editors of [*the journal*] will consider the potential positive impact on the scientific community in considering this work for publication.

We would like to emphasize the following:

1. The focus of our work is the reproducibility of lists of putatively differentially expressed genes (DEGs) in microarray studies.
2. The apparent lack of reproducibility of such DEGs has been used as scientific evidence to criticize microarray technology.
3. Despite the availability of numerous statistical methods for the identification of DEGs, the simple t-statistic (and slight variations) is arguably still the most widely used test statistic, and many of the various methods that exist to create lists of DEGs primarily improve upon the inference from this basic test statistic. This includes the simple unmodified two-sample t-test, Bonferroni and step-up/step-down procedures applied to the t-test, and others. We also note that a ranking criterion based on the t-statistic or the *P*-value derived from it is equivalent.
4. Statistical significance (*P*) derived from the simple two-group t-test has historically been widely used as the only criterion to identify DEGs, often with disappointing results related to reproducibility when it has been measured.
5. Our work was not intended to serve as a comprehensive performance survey of different statistical procedures. Such a survey is not within the scope of our work and by itself constitutes a separate large study.
6. We are NOT claiming that a concurrent use of FC ranking combined with a *P* threshold is the ultimate and best way of identifying DEGs in all circumstances. Instead, it appears to be a reasonable, straightforward (baseline) analysis procedure that can be used to enhance the

reproducibility of DEG lists, especially if the microarray-based experiment is to be reviewed in a clinical or regulatory environment.

A good understanding of these factors is critical for the peer reviewers, editors, and readers to better appreciate the urgency of the issue being addressed in this work and its important contribution to the microarray field.

To clarify the overall goals of this paper and of the MAQC study as a whole, we have made some modifications throughout the manuscript to emphasize the focus on the reproducibility of lists of differentially expressed genes. We also now provide a self-contained description of the design of the MAQC study. For example, we have modified the Abstract to read:

Abstract:
Reproducibility is a fundamental requirement in scientific experiments and clinical contexts. Recent publications raise concerns about the reliability of microarray technology because of the apparent lack of agreement between lists of differentially expressed genes (DEGs). In this study we demonstrate that (1) such discordance may stem from ranking and selecting DEGs solely by statistical significance (*P*) derived from widely used simple *t*-tests; (2) when fold change (FC) is used as the ranking criterion, the lists become much more reproducible, especially when fewer genes are selected; and (3) the instability of short DEG lists based on *P* cutoffs is an expected mathematical consequence of the high variability of the *t*-values. We recommend the use of FC ranking plus a non-stringent *P* cutoff as a baseline practice in order to generate more reproducible DEG lists. The FC criterion enhances reproducibility while the *P* criterion balances sensitivity and specificity.

Additionally, we added a paragraph and modified a sentence in the Introduction that clearly states that:

"The MAQC study was specifically designed to address these previously identified sources of variability in DEG lists. Two very different RNA samples, Stratagene Universal Human Reference RNA and Ambion Human Brain Reference RNA, with thousands of differentially expressed genes, were prepared in sufficient quantities and distributed to three different laboratories for each of the five different commercial whole genome microarray platforms participating in the study. For each platform, each sample was analyzed using five technical replicates with standardized procedures for sample processing, hybridization, scanning, data acquisition, data preprocessing, and data normalization at each site. The probe sequence information was used to generate a stringent mapping of genes across the different platforms and 906 genes were further analyzed with TaqMan® assays using the same RNA samples.

A careful analysis of these MAQC data sets, along with numerical simulations and mathematical arguments, demonstrates that the reported lack of reproducibility of DEG lists can be attributed in large part to identifying DEGs from simple *t*-tests without consideration of FC when sample numbers are small. The finding holds for intra-laboratory, inter-laboratory, and cross-platform comparisons independent of sample pairs and normalization methods, and is increasingly apparent with decreasing number of genes selected."

Point-by-Point Response to Reviewer #2

Note: The reviewer's original comments/questions are in Arial font and the authors' response is in Roman font. The reviewer's comments are numbered for convenience in authors' response. Text changes to the manuscript are indicated in blue fonts.

1. This manuscript, produced by the MAQC project, addresses a very important question of reproducibility of groups of markers identified from microarray studies. Although the question addressed is a critical one, the manuscript falls far short of addressing it to any extent due to serious flaws in the methodology used.

Response:

We appreciate the reviewer's recognition of the importance of the topic that has been addressed in our manuscript on the reproducibility of lists of differentially expressed genes (DEGs). Several PhD statisticians in the MAQC project initially expressed similarly strong doubts and arguments that the proposed methodology is flawed and/or ignores well trusted statistical principles. After review of data and extensive discourse, a critical consensus was reached that reproducibility is a third dimension needing optimization together with sensitivity and specificity. When high reproducibility is of primary concern, giving increased weight to the estimated fold change is necessary. If the rationale presented here is flawed, we would very much appreciate receiving a more detailed rebuttal.

2. The manuscript's main claim is that fold change, not p-values, should be used to order differentially expressed genes (and p-values used to evaluate sens/spec). This claim is weak to begin with due to numerous statistical and practical arguments, and has been previously published by the authors in proceedings of Second Annual MidSouth Computational Biology and Bioinformatics Society Conference. This new dataset, while impressive, does not support their claim further due to flaws in their comparison methodology.

Response:

We agree with the reviewer's assessment that our new data set is "impressive", and we would argue this data set and associated simulations provide a reasonable context within which to discuss the critical issue of reproducibility. We recommend a joint fold change / p-value rule, and when it is applied in practice, researchers are free to order the resulting gene list by either fold change or p-value, depending on ranking objectives. When the reproducibility of lists of DEGs is the major objective, FC ranking produces much more reproducible results. The arguments and points made in this paper represent significant refinements over those in the proceedings of Second Annual MidSouth Computational Biology and Bioinformatics Society Conference (Shi L *et al*., *BMC Bioinformatics*. 2005 Jul 15;6 Suppl 2:S12).

3. First and foremost, the authors use 2-tailed t-test as the statistical test to compare against. As the authors correctly note, this is indeed not the right test to use for small

> datasets (because normality assumption doesn't hold), and in general this is not a state-of-the-art statistical method for this problem. What about trying a rank-based test or one of the newer methods such as SAM etc?

Response:

We believe that the reviewer has misunderstood our claim "*The MAQC data sets and simulations are used to demonstrate that the reported lack of reproducibility of DEG lists can be attributed in large part to identifying DEGs from simple t-tests without consideration of FC when sample numbers are small*." (p.3, Introduction of the original manuscript). We are not saying that the t-test is "not right" or inappropriate. What we are saying is that t-tests used as the sole ranking criterion for generating gene lists is inappropriate if the desire is to also create reproducible results. To reiterate, the simple t-test has often been used previously for identifying DEGs in published reports that criticized the microarray technologies due to the apparent lack of reproducibility of DEG lists.

In our study, we set out to demonstrate and explain why the instability of short gene lists based on the t-test alone is a fundamental mathematical problem (as described in the "Insert" on p.14 of the original manuscript) that we clearly illustrate with the MAQC data and the simulations. This issue is independent of platform or site-to-site variability.

We agree with the reviewer that, in a variety of contexts, DEGs may be identified using numerous different statistical tests including rank tests (*e.g.*, Wilcoxon rank-sum test) and shrunken t-tests (*e.g.*, SAM). These methods are typically promoted in terms of improved sensitivity (power) while retaining nominal rates of specificity. Reproducibility is a fundamental requirement in scientific experiments and clinical contexts, but is seldom emphasized in microarray literature. However, our work was not intended to serve as a comprehensive performance survey of different statistical procedures. Although valuable, such a survey is out of the scope of this work and would be a different large study.

It should also be emphasized that despite the publication of numerous new statistical methods for the identification of DEGs, the simple t-test is arguably still the most widely used approach by the general microarray community. Also, with five technical replicates at each site, our sample sizes are not extremely small, especially when modeling data from all sites simultaneously. Analysis of residuals on log2 normalized signals (not shown) reveals that the assumption of normality--separately for each gene--is not unreasonable for these data.

<u>Rank-based test</u>:

We agree with the reviewer that a rank-based test (*e.g.*, Wilcoxon rank-sum test) is a better choice than the simple t-test when the normality assumption is violated. As mentioned above, this assumption does not appear to be violated here; but we did still explore rank-based analyses. When considering data from only one site (five replicates for each group in the microarray experiments); there are many ties in the rank test statistic. In fact, the Wilcoxon rank-sum test statistic takes on only 26 distinct values (from 15 to 40), and the smallest *P*-value is 0.0079 (two-sided, exact tests). Therefore, using a Wilcoxon rank-sum test for data sets of such small numbers of replicates (5) would create too many ties and would not effectively differentiate the differences among the 12,091 genes used in our analysis. Nevertheless, in a sincere attempt to satisfy the reviewer, we created a POG graph using AFX inter-site comparison as an example (Figure 5 in the revised manuscript). It is easy to see that the rank-sum test did not perform as expected by the reviewer.

SAM:

SAM incorporates both an FDR estimation procedure partnered with a modified (shrunken) t-statistic (along with permutation/resampling). If one ignores the FDR estimation method associated with SAM and instead focuses on the rank of the test statistic (which is essentially our focus) used by SAM, then one sees that SAM is indeed considered in our simulations that compare methods. That is, the top 10 genes that result from SAM are the same top 10 genes that come from rank ordering the shrunken t-statistic. Unfortunately, in sharp contrast to another reviewer, this reviewer did not appear to appreciate the value of the simulations. Therefore, it is not surprising that he/she appears to have overlooked the inclusion of SAM's test statistic in the simulation.

Actions taken:

(1) A few extra sentences justifying our choice of using simple t-test have been added to the Abstract, Introduction, and Conclusion.

(2) In a sincere attempt to satisfy the reviewer, we created a POG graph (Figure 5, added to the revised manuscript) by including Wilcoxon rank-sum test using AFX site-site comparison as an example. As can be easily seen from Figure 5, the SAM POG (pink line), although greatly improved over that of simple t-test (purple line), approached, but did not exceed, the level of POG based on FC ranking (green line).

(3) Scatterplots of SAM d values and FCs (Figure R2-A), and a POG graph (Figure 5, added to revised manuscript) using AFX site-site comparison were created as an example. As can be easily seen from Figure 5, the SAM POG (pink line), although greatly improved over that of simple t-test (purple line), approached, but did not exceed, the level of POG based on FC ranking (green line). This is consistent with the correlation of the log2 FC and SAM d values (Table R2-A). In summary, SAM did not appear to make microarray data (in terms of DEG lists or SAM d values) more reproducible across laboratories. Interestingly, in one case (the AFX data) the stabilization factor used in the denominator of SAM became large. One consequence of this is that the denominators in the SAM test statistic become more homogeneous for all genes, and ranking by SAM approaches ranking by FC. In addition, to address the reviewer's concerns on the use of an "artificial data set", we created a POG graph (Figure R2-B) from a real rat toxicogenomics data set (Guo *et al*.) using FC and SAM ranking for identifying differentially expressed genes. **Compared to fold change ranking, SAM reduced inter-site concordance in this case, a finding that is consistent with what was observed from the MAQC data sets.**

Table R2-A: Correlation matrix of log2 FC and SAM d in inter-site comparison.

	log2 FC (site 1)	**log2 FC (site 2)**	**log2 FC (site 3)**	**SAM_d (site 1)**	**SAM_d (site 2)**	**SAM_d (site 3)**
log2 FC (site 1)	**1.0000**	**0.9895**	**0.9934**	**0.9989**	**0.9885**	**0.9931**
log2 FC (site 2)	**0.9895**	**1.0000**	**0.9944**	**0.9889**	**0.9987**	**0.9947**
log2 FC (site 3)	**0.9934**	**0.9944**	**1.0000**	**0.9919**	**0.9925**	**0.9993**
SAM_d (site 1)	**0.9989**	**0.9889**	**0.9919**	**1.0000**	**0.9890**	**0.9924**
SAM_d (site 2)	**0.9885**	**0.9987**	**0.9925**	**0.9890**	**1.0000**	**0.9938**
SAM_d (site 3)	**0.9931**	**0.9947**	**0.9993**	**0.9924**	**0.9938**	**1.0000**

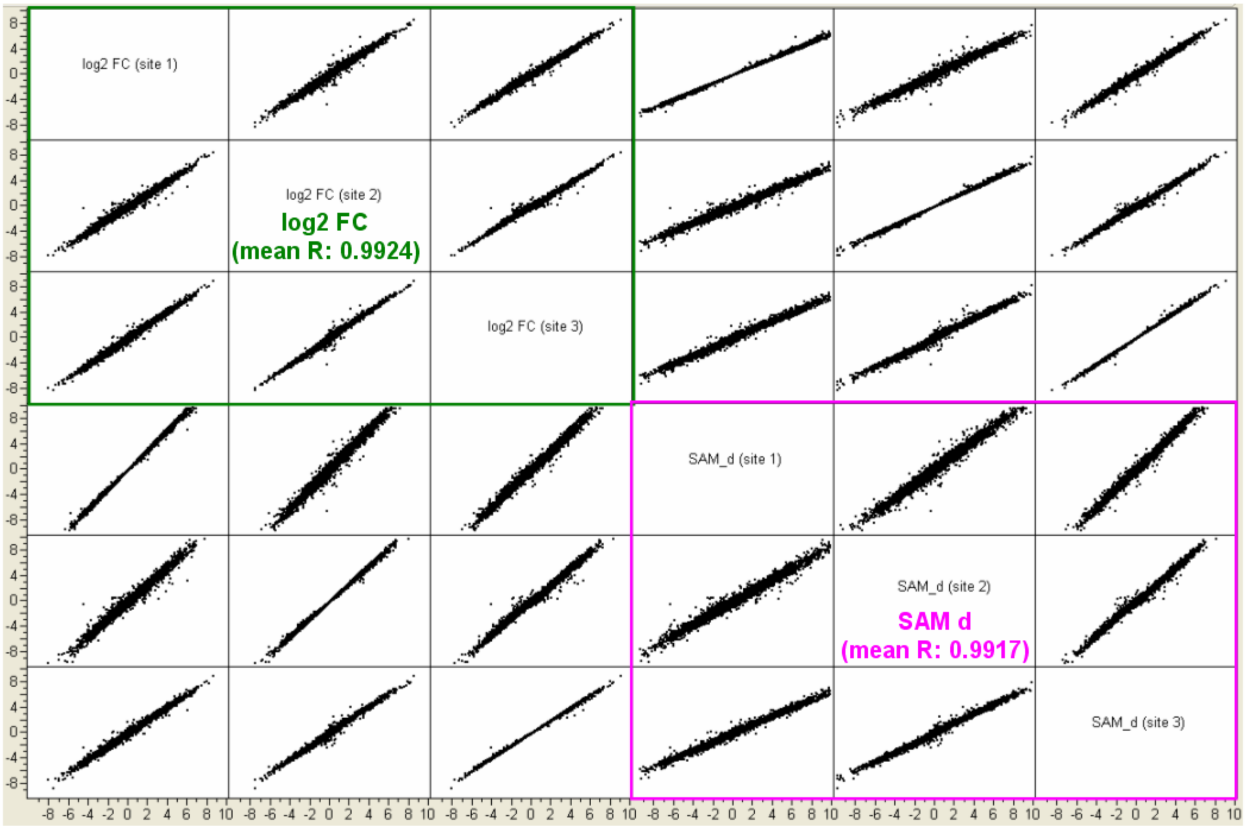


Figure R2-A: Scatter plots of SAM d values and log2FC in inter-site comparison. Affymetrix data on samples A and B from three test sites for the “12,091” commonly mapped genes were used. No flagged (“Absent”) genes were excluded in the analysis.

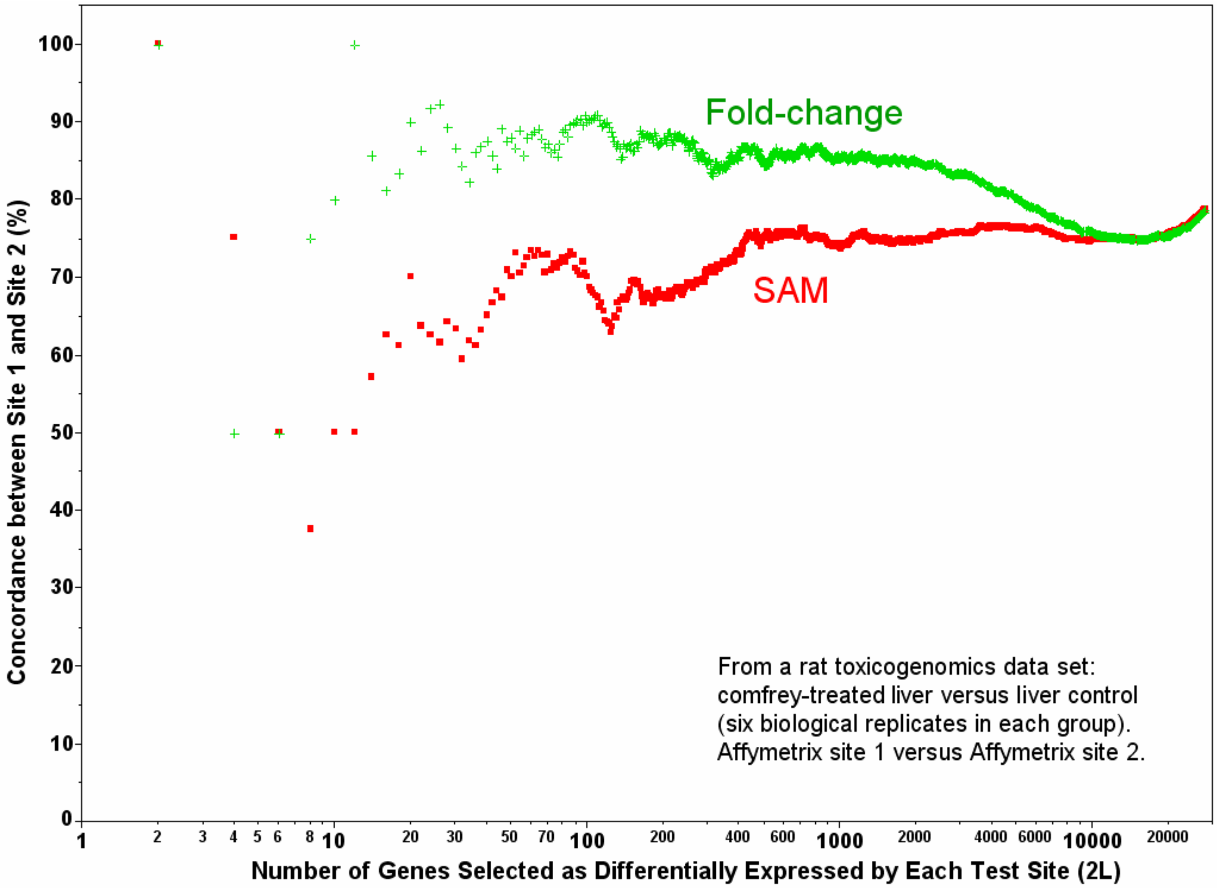


Figure R2-B: SAM reduces inter-site concordance in a real toxicogenomics data set.

4. Second, their simulated data are generated with fold change in mind. There is a strong basis to generation of sensible microarray data simulations in the literature (e.g. David Rocke's group work and others). If the authors want to use simulated data, they should simulate it based on distributions and error models that have been shown to simulate real microarray data. Fitting the simulated data to their expectation of it makes their conclusions completely circular.

Response:

We agree that the simulated data was generated with differing fold changes in mind. We also had several other things in mind, such as emulating the distribution of fold change that is seen in the MAQC data set and in other data sets (thus the three different simulation contexts), simulating distributions of error in replicates seen in the MAQC data sets and other data sets, and considering the relationship in the variance of replicates between different sites and different platforms. Therefore, we feel that our simulations are based on distributions and error models that do in fact emulate real microarray data. It is somewhat puzzling to consider how one would simulate real microarray data and not consider/control one of the most important aspects of the experiment, which is the nature and distribution of fold change. However, the simulated microarray data were created with no *a priori* expectation related to the DEG reproducibility of any particular method. In fact, many of us were quite surprised with the results initially, so it is incongruous to characterize the simulations as part of an analysis that fits preconceived expectations or provides results based on circular reasoning.

Actions taken:

(1) The Methods section is modified to describe in more detail the various factors considered that make the simulated microarray data emulate real microarray data.

5. My other concerns have to do with the exact comparisons performed, but these concerns are minor compared to the two above. In fact, from my perspective the two problems above render main conclusions of this paper unsupported.

Response:

We would be happy to address any additional, specific concerns the reviewer may have on our work.

Begin of original comments:

Reviewer #2(Remarks to the Author):

This manuscript, produced by the MAQC project, addresses a very important question of reproducibility of groups of markers identified from microarray studies. Although the question addressed is a critical one, the manuscript falls far short of addressing it to any extent due to serious flaws in the methodology used.

The manuscript's main claim is that fold change, not p-values, should be used to order differentially expressed genes (and p-values used to evaluate sens/spec). This claim is weak to begin with due to numerous statistical and practical arguments, and has been previously published by the authors in proceedings of Second Annual MidSouth Computational Biology and Bioinformatics Society Conference. This new dataset, while impressive, does not support their claim further due to flaws in their comparison methodology.

First and foremost, the authors use 2-tailed t-test as the statistical test to compare against. As the authors correctly note, this is indeed not the right test to use for small datasets (because normality assumption doesn't hold), and in general this is not a state-of-the-art statistical method for this problem. What about trying a rank-based test or one of the newer methods such as SAM etc?

Second, their simulated data are generated with fold change in mind. There is a strong basis to generation of sensible microarray data simulations in the literature (e.g. David Rocke's group work and others). If the authors want to use simulated data, they should simulate it based on distributions and error models that have been shown to simulate real microarray data. Fitting the simulated data to their expectation of it makes their conclusions completely circular.

My other concerns have to do with the exact comparisons performed, but these concerns are minor compared to the two above. In fact, from my perspective the two problems above render main conclusions of this paper unsupported.

End of original comments

Point-by-Point Response to Peer Reviewer #3' Comments

Manuscript: [MAQC MS-6]
Title: The reproducibility of lists of differentially expressed genes in microarray studies
Corresponding author: Leming Shi (leming.shi@fda.hhs.gov)
Date: July 13, 2006

General Response to Peer Reviewer #3

New statistical methods for the identification of differentially expressed genes (DEGs) continue to appear in the scientific literature. In fact, the variety of existing and emerging methods has caused some confusion in the research community. These methods are typically promoted in terms of improved sensitivity (power) under various assumptions or conditions while retaining nominal rates of specificity. Reproducibility is a fundamental requirement in scientific experiments and clinical contexts, but is seldom emphasized in microarray literature. Reproducibility is a critical third dimension that is distinct from specificity and sensitivity. It is equally if not more important than sensitivity and specificity in certain experimental and clinical contexts. Until recently reproducibility has not adequately been used as an essential criterion for evaluating the pros and cons of statistical methods for identifying DEGs. Demonstrating reproducible performance is critical to the acceptance of microarray-based data in clinical and regulatory environments. We anticipate that the editors of [*the journal*] will consider the potential positive impact on the scientific community in considering this work for publication.

We would like to emphasize the following:

1. The focus of our work is the reproducibility of lists of putatively differentially expressed genes (DEGs) in microarray studies.
2. The apparent lack of reproducibility of such DEGs has been used as scientific evidence to criticize microarray technology.
3. Despite the availability of numerous statistical methods for the identification of DEGs, the simple t-statistic (and slight variations) is arguably still the most widely used test statistic, and many of the various methods that exist to create lists of DEGs primarily improve upon the inference from this basic test statistic. This includes the simple unmodified two-sample t-test, Bonferroni and step-up/step-down procedures applied to the t-test, and others. We also note that a ranking criterion based on the t-statistic or the *P*-value derived from it is equivalent.
4. Statistical significance (*P*) derived from the simple two-group t-test has historically been widely used as the only criterion to identify DEGs, often with disappointing results related to reproducibility when it has been measured.
5. Our work was not intended to serve as a comprehensive performance survey of different statistical procedures. Such a survey is not within the scope of our work and by itself constitutes a separate large study.
6. We are NOT claiming that a concurrent use of FC ranking combined with a *P* threshold is the ultimate and best way of identifying DEGs in all circumstances. Instead, it appears to be a reasonable, straightforward (baseline) analysis procedure that can be used to enhance the

reproducibility of DEG lists, especially if the microarray-based experiment is to be reviewed in a clinical or regulatory environment.

A good understanding of these factors is critical for the peer reviewers, editors, and readers to better appreciate the urgency of the issue being addressed in this work and its important contribution to the microarray field.

To clarify the overall goals of this paper and of the MAQC study as a whole, we have made some modifications throughout the manuscript to emphasize the focus on the reproducibility of lists of differentially expressed genes. We also now provide a self-contained description of the design of the MAQC study. For example, we have modified the Abstract to read:

Abstract:
Reproducibility is a fundamental requirement in scientific experiments and clinical contexts. Recent publications raise concerns about the reliability of microarray technology because of the apparent lack of agreement between lists of differentially expressed genes (DEGs). In this study we demonstrate that (1) such discordance may stem from ranking and selecting DEGs solely by statistical significance (*P*) derived from widely used simple *t*-tests; (2) when fold change (FC) is used as the ranking criterion, the lists become much more reproducible, especially when fewer genes are selected; and (3) the instability of short DEG lists based on *P* cutoffs is an expected mathematical consequence of the high variability of the *t*-values. We recommend the use of FC ranking plus a non-stringent *P* cutoff as a baseline practice in order to generate more reproducible DEG lists. The FC criterion enhances reproducibility while the *P* criterion balances sensitivity and specificity.

Additionally, we added a paragraph and modified a sentence in the Introduction that clearly states that:

"The MAQC study was specifically designed to address these previously identified sources of variability in DEG lists. Two very different RNA samples, Stratagene Universal Human Reference RNA and Ambion Human Brain Reference RNA, with thousands of differentially expressed genes, were prepared in sufficient quantities and distributed to three different laboratories for each of the five different commercial whole genome microarray platforms participating in the study. For each platform, each sample was analyzed using five technical replicates with standardized procedures for sample processing, hybridization, scanning, data acquisition, data preprocessing, and data normalization at each site. The probe sequence information was used to generate a stringent mapping of genes across the different platforms and 906 genes were further analyzed with TaqMan® assays using the same RNA samples.

A careful analysis of these MAQC data sets, along with numerical simulations and mathematical arguments, demonstrates that the reported lack of reproducibility of DEG lists can be attributed in large part to identifying DEGs from simple *t*-tests without consideration of FC when sample numbers are small. The finding holds for intra-laboratory, inter-laboratory, and cross-platform comparisons independent of sample pairs and normalization methods, and is increasingly apparent with decreasing number of genes selected."

Point-by-Point Response to Reviewer #3

Note: The reviewer's original comments/questions are in Arial font and the authors' response is in Roman font. The reviewer's comments are numbered for convenience in authors' response. Text changes to the manuscript are indicated in blue fonts.

1. The manuscript begins with a brief and somewhat self-serving literature review pointing out problems observed in trying to find concordance across microarray studies but completely ignoring more recent publications, including three independent publications that appeared in Nature Methods in 2005 (Bammler et al. 2005; Irizarry et al. 2005; Larkin et al. 2005), a paper that appeared in BMC Genomics earlier this year (Wang et al. 2006), as well as many others (e.g., Carter et al. 2005; Ulrich et al. 2004), that reached very different conclusions regarding the quality of microarray experimental results and helped to define methods to assure concordance between the results.

Response:

The reviewer appears to be concerned about the bias in the selection of references in our paper. For any scientific manuscript other than a review article, the authors are always faced with the challenge of selecting a limited number of references for citation from a much larger number of relevant ones. We acknowledge that this was not an easy task. We did survey the literature prior to selecting a representative set. Unfortunately this may have excluded work that this reviewer finds particularly appealing based on his/her experience or involvement in this area. Our work was not intended to serve as a comprehensive literature review of the publications related to the topic emphasized in our work – the reproducibility of lists of differentially expressed genes in microarray studies. We believe that we have selected the cited references appropriately.

We do not understand how and why the reviewer reached the conclusion that we were "***completely ignoring*** *more recent publications*". In fact, out of the six references mentioned by the reviewer, two of them were cited in our manuscript:

Our ref #39 = "*Irizarry et al. 2005*"
Our ref #24 = "*Wang et al. 2006*"

In addition, our ref #15 (= Mecham *et al*. 2004) came from the same group (Szallasi Z) that published the reference mentioned by the review as "*Carter et al. 2005*". These two references discussed the same concept of sequence-based matching for improving cross-platform consistency.

We note that we were very familiar with the "*Bammler et al. 2005*" and "*Larkin et al. 2005*" publications from the same issue of *Nature Methods* as our ref #39, and the "*Ulrich et al. 2004*" publications from the HESI-led study in *Environ Health Perspect*. We agree that many factors could result in non-reproducible microarray results in terms of the lists of differentially expressed genes. However, none of these publications focused on the impact that selection of data reduction and analysis methods has on the reproducibility of microarray results in terms of lists of differentially expressed genes.

2. The manuscript then presents, in an almost unreadable format populated with sentences that appear to composed almost entirely of acronyms, an analysis of artificial datasets constructed for the MAQC project and simulations to demonstrate that t-tests are not the best method for the analysis of microarray data. There are a number of problems with this analysis, not the least of which is the fact that simulations require a thorough understanding of the nature of the data, the relationships between the entities being simulated, and the nature of the structure of the variance in the observations. None of this is really known for microarrays and gene expression profiles and so a simulation can be constructed based on assumptions of what one wants to find that will easily verify the underlying assumptions The MAQC dataset itself is also extremely problematic since it does not represent "real" gene expression data and consequently does not represent the full spectrum of inter-related patterns that appear in microarray profiles.

Response:

"***Acronyms***": This is a good point. We recognize that the use of acronyms impacts the readability of any manuscript. The authors debated intensively regarding whether we should allow limited use of acronyms in the manuscript. The word count for the manuscript was a very practical consideration that led us to a compromise on this point. In order to meet space limitations without compromising readability we used a limited number of acronyms in the manuscript and made sure that they were clearly defined twice. The number of acronyms used in the manuscript is limited (mainly DEG and POG) and we clearly define the acronyms at the first usage after the Abstract and again adding the definition in the parentheses right after each acronym's first occurrence in the text.

"***Simulations***": We did not have *a priori* expectations for the simulation. The results of the simulation were surprising to many in the project and reviewed by all (even at the source code level). If the results were surprising, then it does not logically follow that the results of the simulation were predetermined based on "*assumptions of what one wants to find*". We have added text within MS-6 to describe the assumptions and patterns, error and FC distributions created for the simulations. We have studied a variety of conditions that appear to greatly influence reproducibility in real microarray experiments: CV within group, amount of differential expression, size of gene list, differences between sites and platforms. Our simulated distributions of error in replicates were based on what we observed in the MAQC data sets and other "*real*" data sets (e.g., toxicogenomics studies) by considering the relationship in the variance of replicates between different sites and different platforms. Therefore, we feel that our simulations are based on distributions and error models that emulate "*real*" microarray data.

"***The MAQC dataset, …, extremely problematic***": We agree that the MAQC dataset does not examine a biological problem. However, the dataset was generated using biological reference RNA samples and the dataset is useful for understanding conditions where platforms and/or laboratories agree and disagree, even in this somewhat complex case. However, in the simulations within this paper, and in a much more realistic toxicogenomics data set based on an actual experiment that was used to "validate" the MAQC study and submitted to *Nature Biotechnology* as an accompanying paper, we have examined a variety of scenarios that use more extreme data (large amounts of expression with higher magnitude of FC) and more biologically realistic data. We have seen that our simulations appear to closely emulate real experimental results related to reproducibility. So we have covered extreme cases, and biologically realistic cases in this paper. As we have more confidence in the simulations covering both normal and extreme situations, we can use the simulations to examine important trade-offs between reproducibility, sensitivity, and specificity, something that we cannot do in most microarray experiments as we cannot be sure of absolute truth with real biological specimens. We also should recognize that there exists no data set that represents "***the full spectrum*** *of inter-related patterns that appear in microarray profiles.*" However, we feel that the data sets, real and simulated, related to the MAQC project are very informative and thought-provoking regarding reproducibility. In addition, we agree that there are important issues such as "interrelatedness of genes" that need to be carefully considered, but that is beyond the scope of this work.

3. Nevertheless, the manuscript reaches a conclusion that anyone working in the field has known for years - that simple t-tests are not the best method for the analysis of microarray data, particularly for small sample sizes and particularly for genes expressed at low levels where the signal approaches the noise. This, indeed, was the justification for the development of SAM, an algorithm that uses pooled variance for genes binned by expression level to correct for this effect - and not surprisingly the authors find that SAM is superior to a simple t-test. The other approach the authors find to be superior to the t-test? Genes selected based on volcano plots. Again, this is not something new but a technique that has been used for quite some time.

Response:

We do not simply conclude, as the reviewer claims in both paragraph 2 and paragraph 3 of the review that "*t-tests are not the best method for analysis of microarray data*". This represents a misreading of our manuscript and misses our significant conclusion that use of the t-test alone to generate lists of differentially expressed genes causes the lack of reproducibility of short gene lists. Starting with the Abstract we clearly state that:

"To generate more reproducible DEG lists across a variety of biological, laboratory, and platform scenarios, the concurrent use of FC ranking and *P* cutoff is recommended. An FC criterion explicitly incorporates the measured quantity to ensure reproducibility, whereas a *P* criterion incorporates control of sensitivity and specificity." (original manuscript)

In the revised manuscript, we have modified these sentences to make our message clearer: "We recommend the use of FC ranking plus a non-stringent *P* cutoff as a baseline practice in order to generate more reproducible DEG lists. The FC criterion enhances reproducibility while the *P* criterion balances sensitivity and specificity." (revised manuscript)

We do NOT conclude that the simple t-tests should not be used for the analysis of microarray data. In fact, we use t-test for the analysis of microarray data in our manuscript in order to generate a significance measure for the fold-change of each gene. **We do not think t-test itself is wrong in microarray data analysis.** Instead, the problem of the reported lack of reproducibility of gene lists stems from **the use of t-statistic (*P*) as the <u>ranking</u> criterion** for the identification of differentially expressed genes, with or without a FC threshold.

We note further that **SAM does not necessarily produce more reproducible lists compared to FC**, although it may produce more highly specific lists than t-tests. Our technique of using the combination of significance and fold change is not new, but **fold change ranking with a significance threshold** is more specific than simply saying “using volcano plots” without clearly identifying the more important factor, FC. The ramifications of this technique are ideally suited for reproducibility, which is a new concept.

We agree with the reviewer that, in a variety of contexts, DEGs may be identified using numerous different statistical tests including rank tests (*e.g.*, Wilcoxon rank-sum test) and shrunken t-tests (*e.g.*, SAM). These methods are typically promoted in terms of improved sensitivity (power) while retaining nominal rates of specificity. Reproducibility is a fundamental requirement in scientific experiments and clinical contexts, but is seldom emphasized in microarray literature. However, our work was not intended to serve as a comprehensive performance survey of different statistical procedures. Although valuable, such a survey is out of the scope of this work and would be a different large study.

It should also be emphasized that despite the publication of numerous new statistical methods for the identification of DEGs, the **simple t-test is arguably still the most widely used approach by the general microarray community**.

Actions taken:

(1) A few extra sentences justifying our choice of using simple t-test have been added to the Abstract, Introduction, and Conclusion.

(2) In a sincere attempt to satisfy the reviewer, we created a POG graph (Figure 5 of the revised manuscript) by including SAM and Wilcoxon rank-sum test, which was requested by another reviewer, using AFX site-site comparison as an example. As can be easily seen from Figure 5, the SAM POG (pink line), although greatly improved over that of simple t-test (purple line), approached, but did not exceed, the level of POG based on FC ranking (green line). In addition, to address the reviewer’s concerns on the use of “artificial data set”, we created a POG graph (Figure R3-A) from a real rat toxicogenomics data set (Guo *et al.*) using FC and SAM ranking for identifying differentially expressed genes. **Compared to fold change ranking, SAM reduced inter-site concordance in this case, a finding that is consistent with what was observed from the MAQC data sets.**

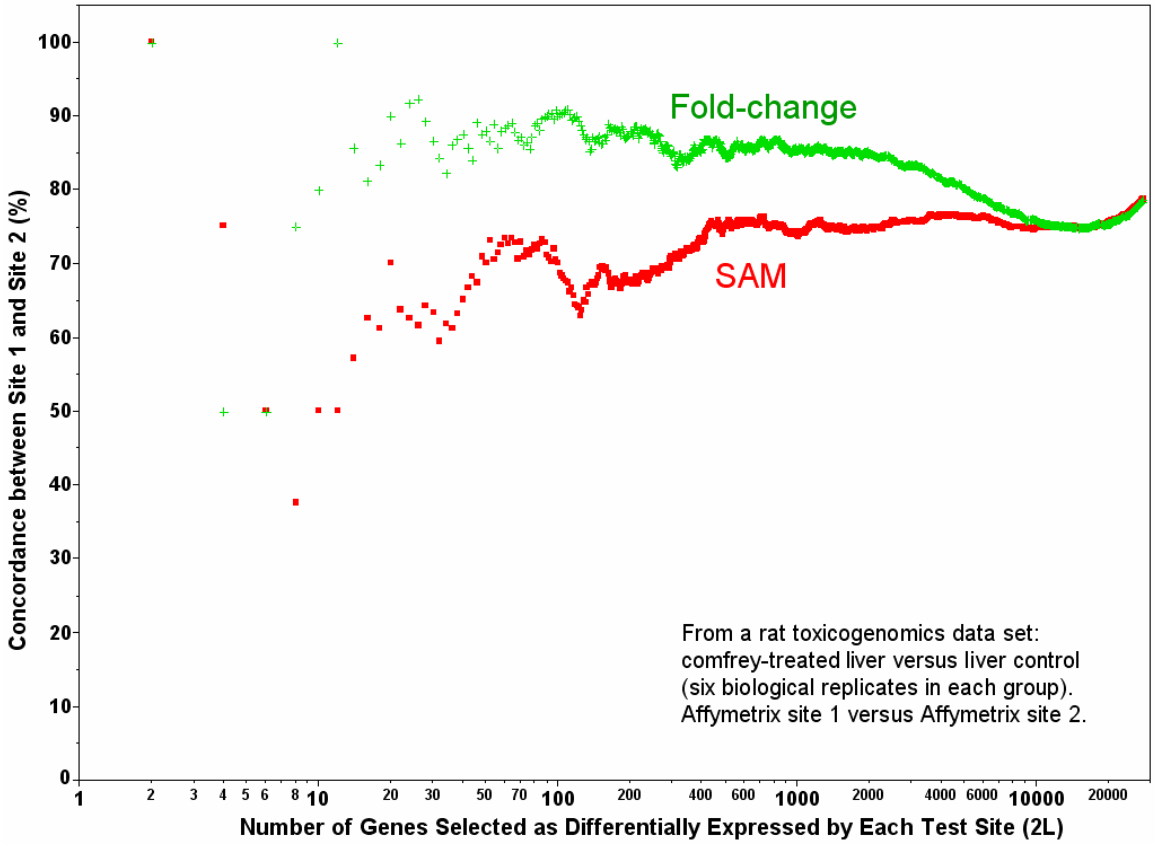


Figure R3-A: SAM reduces inter-site concordance in a real toxicogenomics data set.

(3) Scatterplots of SAM d values and FCs (Figure R3-B) using AFX site-site comparison were created as an example. Again, inter-site consistency of SAM d values did not out-perform that of log2 FC (Table R3). In summary, SAM did not appear to make microarray data (in terms of DEG lists or SAM d values) more reproducible across laboratories. Interestingly, in one case (the AFX data) the stabilization factor used in the denominator of SAM became large. One consequence of this is that the denominators in the SAM test statistic become much more homogeneous for all genes. Therefore, ranking order by SAM approaches that by FC. In fact the mathematical analysis of the statistical properties of the non-central t-statistic in the Inset Box in the manuscript should provide the reader with valuable insight into why the SAM statistic and other shrunken t-statistics can provide improved reproducibility of DEG lists.

Table R3: Correlation matrix of log2 FC and SAM d in inter-site comparison.

	log2 FC (site 1)	**log2 FC (site 2)**	**log2 FC (site 3)**	**SAM_d (site 1)**	**SAM_d (site 2)**	**SAM_d (site 3)**
log2 FC (site 1)	**1.0000**	**0.9895**	**0.9934**	**0.9989**	**0.9885**	**0.9931**
log2 FC (site 2)	**0.9895**	**1.0000**	**0.9944**	**0.9889**	**0.9987**	**0.9947**
log2 FC (site 3)	**0.9934**	**0.9944**	**1.0000**	**0.9919**	**0.9925**	**0.9993**
SAM_d (site 1)	**0.9989**	**0.9889**	**0.9919**	**1.0000**	**0.9890**	**0.9924**
SAM_d (site 2)	**0.9885**	**0.9987**	**0.9925**	**0.9890**	**1.0000**	**0.9938**
SAM_d (site 3)	**0.9931**	**0.9947**	**0.9993**	**0.9924**	**0.9938**	**1.0000**

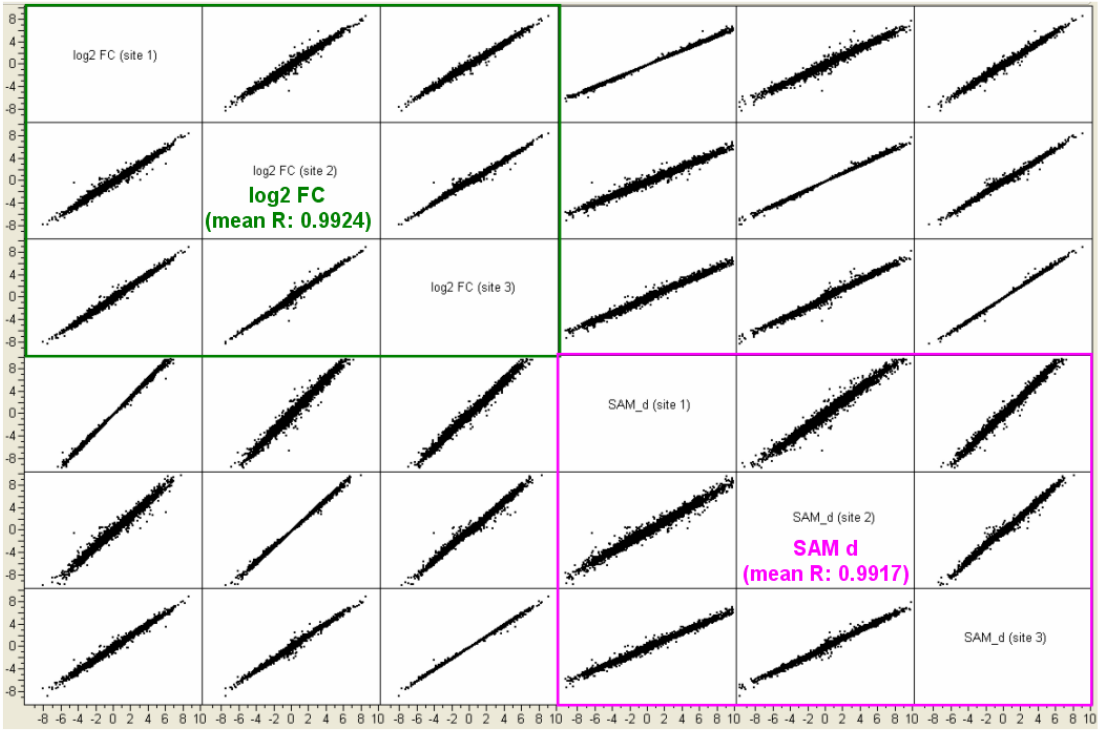


Figure R3-B: Scatter plots of SAM d values and log2FC in inter-site comparison. Affymetrix data on samples A and B from three test sites for the "12,091" commonly mapped genes were used. No flagged ("Absent") genes were excluded in the analysis.

4. So in the end, the manuscript creates a problem that has largely been resolved and arrives at solutions that have been well known for some time.

Response:

We feel that this problem of reproducibility is yet to be properly understood and recognized related to microarray analysis. While the use of FC and significance combined is not new, which we acknowledge, our recommendations for its particular use of **FC ranking** with a non-stringent significance threshold related to performance with respect to reproducibility are new and are important to the microarray field. It represents a dramatic change to the common practice of "statistical" analysis of microarray data where statistical significance measures (*e.g.*, t-statistic) have been widely used for ranking and selecting differentially expressed genes.

Begin of original comments:

Reviewer #3(Remarks to the Author):

Shi and colleagues present

The manuscript begins with a brief and somewhat self-serving literature review pointing out problems observed in trying to find concordance across microarray studies but completely ignoring more recent publications, including three independent publications that appeared in Nature Methods in 2005 (Bammler et al. 2005; Irizarry et al. 2005; Larkin et al. 2005), a paper that appeared in BMC Genomics earlier this year (Wang et al. 2006), as well as many others (e.g., Carter et al. 2005; Ulrich et al. 2004), that reached very different conclusions regarding the quality of microarray experimental results and helped to define methods to assure concordance between the results.

The manuscript then presents, in an almost unreadable format populated with sentences that appear to composed almost entirely of acronyms, an analysis of artificial datasets constructed for the MAQC project and simulations to demonstrate that t-tests are not the best method for the analysis of microarray data. There are a number of problems with this analysis, not the least of which is the fact that simulations require a thorough understanding of the nature of the data, the relationships between the entities being simulated, and the nature of the structure of the variance in the observations. None of this is really known for microarrays and gene expression profiles and so a simulation can be constructed based on assumptions of what one wants to find that will easily verify the underlying assumptions The MAQC dataset itself is also extremely problematic since it does not represent "real" gene expression data and consequently does not represent the full spectrum of inter-related patterns that appear in microarray profiles.

Nevertheless, the manuscript reaches a conclusion that anyone working in the field has known for years - that simple t-tests are not the best method for the analysis of microarray data, particularly for small sample sizes and particularly for genes expressed at low levels where the signal approaches the noise. This, indeed, was the justification for the development of SAM, an algorithm that uses pooled variance for genes binned by expression level to correct for this effect - and not surprisingly the authors find that SAM is superior to a simple t-test. The other approach the authors find to be superior to the t-test? Genes selected based on volcano plots. Again, this is not something new but a technique that has been used for quite some time.

So in the end, the manuscript creates a problem that has largely been resolved and arrives at solutions that have been well known for some time.

Bammler, T., R.P. Beyer, S. Bhattacharya, G.A. Boorman, A. Boyles, B.U. Bradford, R.E. Bumgarner, P.R. Bushel, K. Chaturvedi, D. Choi, M.L. Cunningham, S. Deng, H.K. Dressman, R.D. Fannin, F.M. Farin, J.H. Freedman, R.C. Fry, A. Harper, M.C. Humble, P. Hurban, T.J. Kavanagh, W.K. Kaufmann, K.F. Kerr, L. Jing, J.A. Lapidus, M.R. Lasarev, J. Li, Y.J. Li, E.K. Lobenhofer, X. Lu, R.L. Malek, S. Milton, S.R. Nagalla, P. O'Malley J, V.S. Palmer, P. Pattee, R.S. Paules, C.M. Perou, K.

Phillips, L.X. Qin, Y. Qiu, S.D. Quigley, M. Rodland, I. Rusyn, L.D. Samson, D.A. Schwartz, Y. Shi, J.L. Shin, S.O. Sieber, S. Slifer, M.C. Speer, P.S. Spencer, D.I. Sproles, J.A. Swenberg, W.A. Suk, R.C. Sullivan, R. Tian, R.W. Tennant, S.A. Todd, C.J. Tucker, B. Van Houten, B.K. Weis, S. Xuan, and H. Zarbl. 2005. Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods 2: 351-356.

Carter, S.L., A.C. Eklund, B.H. Mecham, I.S. Kohane, and Z. Szallasi. 2005. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. BMC Bioinformatics 6: 107.

Irizarry, R.A., D. Warren, F. Spencer, I.F. Kim, S. Biswal, B.C. Frank, E. Gabrielson, J.G. Garcia, J. Geoghegan, G. Germino, C. Griffin, S.C. Hilmer, E. Hoffman, A.E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S.Q. Ye, and W. Yu. 2005. Multiple-laboratory comparison of microarray platforms. Nat Methods 2: 345-350.

Larkin, J.E., B.C. Frank, H. Gavras, R. Sultana, and J. Quackenbush. 2005. Independence and reproducibility across microarray platforms. Nat Methods 2: 337-344.

Ulrich, R.G., J.C. Rockett, G.G. Gibson, and S.D. Pettit. 2004. Overview of an interlaboratory collaboration on evaluating the effects of model hepatotoxicants on hepatic gene expression. Environ Health Perspect 112: 423-427.

Wang, Y., C. Barbacioru, F. Hyland, W. Xiao, K.L. Hunkapiller, J. Blake, F. Chan, C. Gonzalez, L. Zhang, and R.R. Samaha. 2006. Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. BMC Genomics 7: 59.

End of original comments