11-30-2004

Multiple Lab Comparison of Microarray Platforms

Rafael A. Irizarry et al.
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, rafa@jhu.edu

Multiple Lab Comparison of Microarray Platforms

Rafael A. Irizarry[1],Daniel Warren[2],Forrest Spencer[2 3],Shyam Biswal[4], Bryan C. Frank[5],Edward Gabrielson[6],Joe G.N. Garcia[7], Joel Geoghegan[8], Gregory Germino[9], Constance Griffin[10], Sara C. Hilmer[11] Eric Hoffman[11], Anne E. Jedlicka[12], Ernest Kawasaki[8], Irene F. Kim[9] Laura Morsberger[10], Hannah Lee[4], David Petersen[8], John Quackenbush[5], Alan Scott[13], Michael Wilson[14], Yanqin Yang[2], Shui Qing Ye[7], Wayne Yu[6]

November 30, 2004

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Publich Health, Baltimore, MD

[2]Department of Molecular Biology and Genetics, Johns Hopkins University, Baltimore, MD

[3]JHMI microarray Core, Johns Hopkins University, Baltimore, MD

[4]Deptartment of Environmental Health Sciences, Johns Hopkins Bloomberg School of Public Health, Baltimore 21205,MD

[5]The Institute for Genomic Research, 9712 Medical Center Dr. Rockville, MD

[6]Oncology Microarray Facility, Johns Hopkins University, Baltimore, MD

[7]Division of Pulmonary and Critical Care Medicine, Johns Hopkins University School of Medicine, Mason F. Lord Bldg., Center Tower #665, Baltimore MD

[8]NCI's Microarray Core Facility, Advanced Technology Center, Gaithersburg, MD

[9]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

[10]Department of Pathology, Johns Hopkins University, School of Medicine,Baltimore, MD

[11]Research Center for Genetic Medicine, Children's National Medical Center, George Washington University, Washington, D.C

[12]W. Harry Feinstone Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

[13]Molecular Microbiology and Immunology and Gene Array Core Facility, Malaria Research Institute,Bloomberg School of Public Health,Johns Hopkins University,Baltimore, MD

[14]Microarray Research Facility,Research Technologies Branch, DIR, National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA

Abstract

Microarray technology is a powerful tool able to measure RNA expression for thousands of genes at once. Various studies have been published comparing competing platforms with mixed results: some find agreement, others do not. As the number of researchers starting to use microarrays and the number of cross-platform meta-analysis studies rapidly increase, appropriate platform assessments become more important.

Here we present results from a comparison study that offers important improvements over those previously described in the literature. In particular, we notice that none of the previously published papers consider differences between labs. For this paper, a consortium of ten labs from the Washington DC/Baltimore (USA) area was formed to compare three heavily used platforms using identical RNA samples: Appropriate statistical analysis demonstrates that relatively large differences exist between labs using the same platform, but that the results from the best performing labs agree rather well.

1 Introduction

Microarray technology has become an important tool in medical science and basic biology research. The number of publications based on microarray data this year is well over 1000. A first time user will find many platform options and little guidance on the most appropriate for their application. Various comparison studies have been published presenting contradictory results. Some find agreement [1, 2, 3, 4, 5, 6], others do not [7, 8, 9, 10]. In this paper we demonstrate that the disagreement found by some studies may be due to disputable statistical analyses. In particular, none of these studies consider the lab to lab variability. This *lab effect* is well known to exist and has been observed in all scientific fields[11]. For this reason it is essential to assess the lab effect before drawing conclusions about platform performances.

In this paper we describe a comparison study that permits an appropriate assessment of the competing platforms. A consortium of ten labs from the Washington DC/Baltimore (USA) area was formed to compare three of the leading platforms. Each lab was given identical RNA samples which were processed according to what each lab considered best practice. Five of the labs used Affymetrix GeneChips, three used two-color spotted cDNA arrays, and two used two-color long oligo arrays. We describe features of our experiment that are necessary for such studies to be informative and a set of simple *assessment measures* useful for summarizing and interpreting the observed data. Some of these assessment measures are also useful for quantifying the level of agreement across technologies, an important task given the advent of cross-platform meta-analysis studies [12, 13, 14]. Such meta-analysis add substantial value to experimental datasets collected for one purpose, but that are later recognized as having value for another.

To decide among various strategies for measuring the same quantity one looks to optimize accuracy and precision. Because in many situations precision can be improved at the cost of accuracy, and vice-versa, one tries to find the strategy providing the "best" balance. The definition of best will depend on the application. Thus, it is important to consider precision and accuracy in the context of a realistic applied problem. For the comparisons presented in all the cited publications the most common application of microarray technology, screening for a few candidate genes that appear to be differentially expressed among thousands of *null genes* that are not, was mimicked. The candidate genes are typically validated with an established technology such as RT-PCR. In such experiments, the measurement of interest is relative expression across the samples being compared, which is quantified with log fold change (the log provides symmetry between genes that are

up and down regulated). A preferred strategy will better distinguish the log fold change of differentially expressed genes from the log fold change of null genes. A fundamental problem with many of the cited papers[1, 8, 9, 5, 6] is that conclusions are established on accuracy assessments based on a very small subset of the genes, without considering precision in the context of the "real-life" application.

An appropriate comparison experiment in the context of detecting differentially expressed genes, requires at least the following three features: 1) To appropriately assess precision we should have an a-priori expectation of no fold change for most genes. 2) To appropriately assess accuracy, an a-priori expectation of non-zero log fold change of a few genes is needed. 3) To be able to distinguish between platform effect and lab effect, at least two labs should provide data from each platform. We have designed the first platform comparison experiment that includes all of these features. To incorporate feature 1) into our experiment each lab hybridized the same RNA to two separate arrays, referred to as *technical replicates*, which permitted us to measure the technical variation for each lab. Feature 2) can be accomplished by the use of *spike-in experiments*[15]. However, the development of spike-in experiments is costly and arguably unrealistic in the context of platform comparisons. We developed an alternative strategy based on mixtures from four human cell lines, each deficient for a gene required for peroxisome biogenesis. Specifically, we created two samples for which we expect a few genes to be differentially expressed. For four specific genes we have an a-priori expectation of fold changes: 2:1, 3:2, 2:3, and 1:2. We refer to these genes as the *altered genes*. For each of these samples we created an exact copy, or technical replicate, for a total of four samples. Exact copies of these four samples where hybridized by the ten labs. Details are given in Section 4. The raw data from each lab was then pre-processed to form two replicates of relative expression measurements, log (base 2) fold change, comparing the two samples for each gene. These were then summarized with our assessment measures, which were used to assess performance and agreement.

2 Results

We quantified relative expression between our two samples with log (base 2) fold change. For each lab we had technical replicates, thus we obtained two replicate log fold change measurements for each gene. These measurements were then summarized with the assessment measures and plots used to assess platform perfomance. Table 1 and Figure 1 show the assessment measures and plots, described

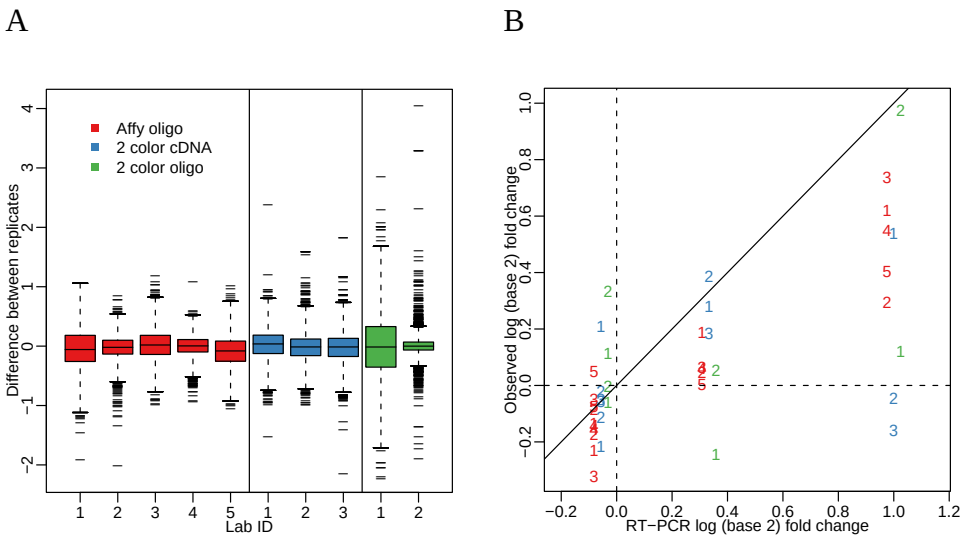

Figure 1: A) Box-plot of the difference in log fold change between replicate measurements from each of the 10 labs. B) Observed log fold change versus nominal (from RT-PCR) log fold change for the four altered genes. The results for each of the 10 labs are represented with colors for the different platorms, as in A), and numbers for lab IDs assigned within each platforms. The solid line is the identity function and reprensents perfect accuracy.

in detail in Section 2.1, for all ten labs. In general, better accuracy and precision was achieved by the Affymetrix labs. However, overall, the best performing lab was two-color oligo lab 2 (each lab was assigned an arbitrary identification number). Furthermore, two-color cDNA lab 1 outperformed most Affy oligo labs in many categories. The worst performance was observed from two-color oligo lab 1. Notice the best and worst overall performance came from labs using the same platform. This fact underscores the importance of considering the lab effect. In general, we found lab had a larger effect on, for example, precision than platform and that the results from the best performing labs agreed rather well. Detailed results are described in the remainder of this Section.

2.1 Assessment Measures and Plots

To summarize **precision** we use two simple measures: correlation across replicate log fold change measurements and standard deviation (SD) of the difference between replicate log (base 2) fold change measurements. The latter measure can

Table 1: Assessment measures for the 10 labs.

Platform	Lab ID	Correlation	SD	Signal (SE)	Proportion of Agreement 25	50	100
Affy oligo	1	0.48	0.32	0.63 (0.19)	0.72	0.56	0.54
Affy oligo	2	0.76	0.17	0.29 (0.08)	0.80	0.70	0.70
Affy oligo	3	0.67	0.24	0.70 (0.14)	0.68	0.66	0.60
Affy oligo	4	0.79	0.15	0.53 (0.08)	0.80	0.70	0.65
Affy oligo	5	0.59	0.25	0.36 (0.09)	0.64	0.68	0.55
two-color cDNA	1	0.65	0.23	0.59 (0.13)	0.68	0.64	0.65
two-color cDNA	2	0.68	0.21	0.08 (0.15)	0.28	0.30	0.38
two-color cDNA	3	0.46	0.23	-0.08 (0.13)	0.72	0.68	0.50
two-color oligo	1	0.68	0.51	0.03 (0.16)	0.40	0.36	0.33
two-color oligo	2	0.90	0.10	0.87 (0.17)	0.44	0.72	0.81

be interpreted as the typical percentage increase in log-fold change when looking at replicate data. For example, a platform with standard deviation value of 0.14 will typically produce measurements (which in theory should be equal) that are 10% [$\log_2(0.14) = 1.10$] different. These assessment measures can also be used to quantify the similarity between measurements made by different platforms. Throughout the paper we will refer to these two assessment measures as *correlation* and *SD* (columns 3 and 4 in Table 1). Box-plots of the differences used to compute the SD are summary plots that provide more information, such as the size of the largest difference. Figure 1A shows a box-plot for each lab. Both Table 1 and Figure 1A demonstrate that precision is comparable across platforms. With the exception of two-color oligo lab 1, all the labs performed similarly, and it is clear that the lab effect is stronger than the platform effect. Notice in particular that if we order the SD values for the Affy labs in increasing order of lab technician experience, the SDs are 0.32, 0.25, 0.24, 0.17, 0.14, i.e. the more experience the more precise. A figure plotting SD against years of experience is given in the supplemental materials.

To assess **accuracy** we regressed the observed log (base 2) fold changes of the altered genes against nominal log (base 2) fold changes obtained using RT-PCR. Ideally, if the nominal fold change doubles, so should the observed fold change. Thus on the $\log_2$ scale, observed fold change should be linear in nominal fold change with a slope of 1. Throughout the paper we will refer to this assessment measure as *signal* (column 5 of Table 1). Scatter-plots of the observed versus nominal log fold changes are a graphical way to summarize the same informa-

tion while providing more details. Figure 1B shows the scatter plot for all labs with an identity line added to demonstrate perfect accuracy. All the labs appear to give attenuated log fold change estimates which is consistent with previous observations[15]. In general, the Affy labs appear to have better accuracy than the two-color platforms, although the best signal measure is attained by two-color oligo lab 2. However, most of the observed differences are not statistically significant.

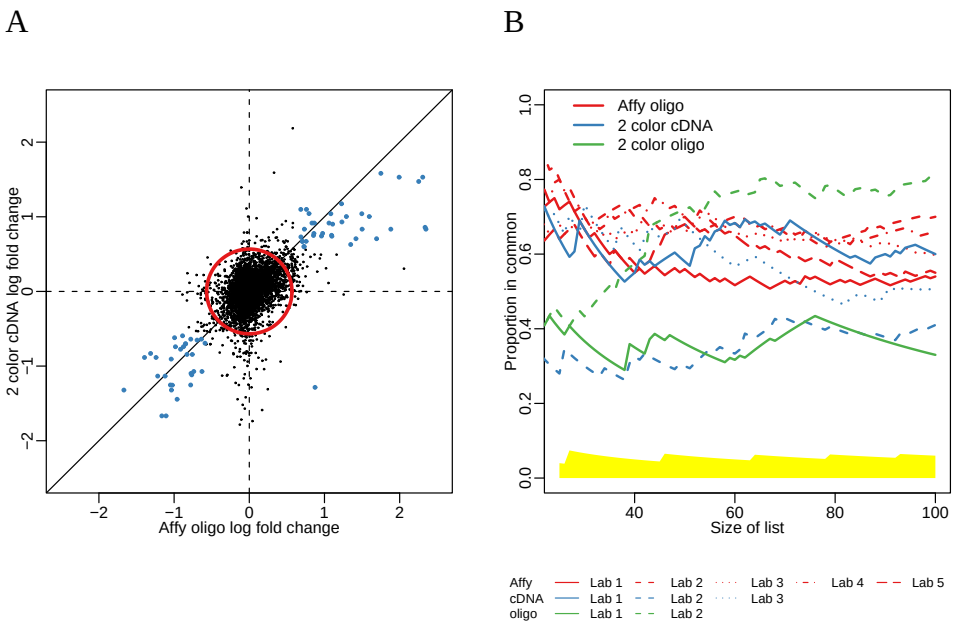

Figure 2: A) Scatter-plot of observed log fold change from two-color cDNA lab 1 and Affy oligo lab 4. Points inside red circle represent genes that do not appear to be differentially expressed. Blue points are genes that appear to be differentially expressed. B) CAT plot showing curves of log fold change agreement between lab replicates for list sized ranging from 25 to 100. The different colors represent the different platforms. The different line types represent the different labs within each platform, thus each lab is represented by a unique color/line type pair. The yellow strip represents critical values for rejecting the null hypothesis of no agreement at the 0.001 level.

Our final assessment measure/plot is particularly useful for summarizing across

technology agreement. In most experiments, we expect only some genes to be differentially expressed, which attenuates correlations across technologies toward 0. To see this, in Figure 2A we show a scatter-plot of the log fold changes obtained for the best performing Affy oligo and two-color cDNA labs. Notice that about 95% of genes are within the blue circle where there appears to be no correlation. These genes have fold-changes close to 0 and are probably not differentially expressed. Because we are measuring 0 log fold change plus random measurement error for these genes, one does not expect across platform measurements to be correlated. However, for the few genes outside the blue circle, which are likely to be differentially expressed, there is good agreement. In practice we typically screen a small subset of genes that appear to be differentially expressed. Therefore, it is more important to assess agreement for genes that are likely to pass this screen. are more important for this subset. To account for this fact we introduce a new descriptive plot: the *correspondence at the top* (CAT) plot. To create the CAT plot we rank genes by log fold change using two different measurements. We then create lists of candidate genes for a range of list sizes. CAT curves show the proportion of genes in common plotted against the size of the lists. Figure 2B shows an example of CAT curves created by comparing lists generated by replicate measurements from each lab. This plot shows, for example, that the agreement of lists of the top 100 genes created from replicate fold-change measurements ranged from 33% to 81% percent. For the CAT plots used in this study, we stop at a list size of 100 because we do not expect more than 100 genes to be differentially expressed, thus correspondence of larger lists are not of interest. As assessment measures, we reported the value of these curves for list sizes 25, 50, and 100. We will refer to these assessment measures as the *Proportion of Agreement* (columns 6, 7, and 8 in Table 1). Figure 2B also shows a yellow band representing critical values, at the 0.001 level, against the null hypothesis of no agreement at all.

We used CAT plots to assess across platform agreement. Figure 3 shows CAT curves where the agreement of the results from all labs are compared to the best performing affy oligo lab (Figure 3A), the best performing two-color cDNA lab (Figure 3B), and the best performing two-color oligo lab (Figure 3C). Figure 3 demonstrates that the Affy oligo labs are consistently similar to the best performing labs. Figure 3 also demonstrates that, apart from 2 labs, there appears to be good agreement regardless of platform. For example, the agreement of Affy oligo lab 4 with the four other Affy labs was 52%, 58%, 52%, and 60%, and with the best performing two-color lab the agreement was 42%. Notice that the agreement of the replicate measurements from Affy oligo lab 4 was 65%. In Figure 3 we also see that the agreement of Affy oligo labs 1, 2, 3, and 5 with Affy oligo lab

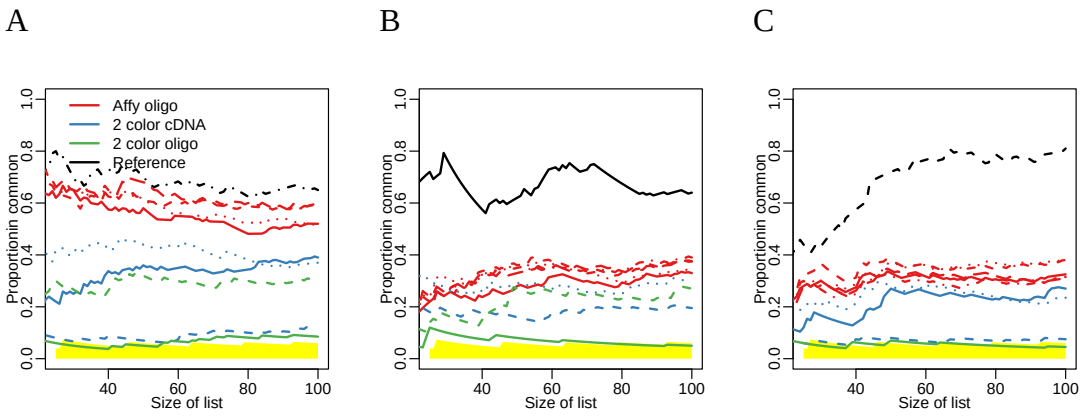

Figure 3: A) CAT plots with curves for each lab compared to the best performing Affy oligo lab. The line types represent the different labs as in Figure 2. The different color represent the different platforms. The black curve is the CAT curve comparing the reference lab to itself. B) As A) but for the best performing two-color cDNA lab. C) As A) but for the best performing two-color oligo lab.

4 is almost the same as the agreement of lab 4 with itself (comparing replicates). The same is not true for the other technologies. This suggest that the Affymetrix platform is by far the most consistent across labs. The supplemental material provides tables, showing all pairwise correlations and agreement proportions (in a list of size 100) between the different labs, that support this result.

Throughout the paper and supplemental material we do not discuss the statistical significance of differences observed in the precision measures. This is because these are based on thousands of data points and any difference shown in our tables is statistically significant. The measure of accuracy is based on only 8 data points, thus we include a measure of uncertainty in Table 1. Although the accuracy results are not nearly as reliable as the precision results, the across platform CAT plots provide a measure of *relative accuracy* based on enough data points to provide reliable conclusions.

2.2 Pre-processing

In all microarray technologies a fair amount of further pre-processing occurs following image processing. Background corrections and normalization are typical steps. Various groups have shown that these steps can have a significant impact on downstream analysis[15, 16]. Most array manufacturers provide analytic pre-

processing software requiring very little input from the user. We found that within and across platform performance can be greatly improved by the use of alternative pre-processing algorithms.

2.3 Annotation

For our analysis, probe level data from Affymetrix oligo arrays were pre-processed with the robust multi-array analysis (RMA)[15]. Print-tip normalization with no background correction was used to pre-process probe level data from the two-color technologies[16]. Because algorithms implementing these methodologies are available from the Bioconductor project[17], we will refer to them as the *Bioconductor procedures*. We compared the results obtained with this approach to those obtained with what we consider to be the default approaches: Affymetrix's MAS 5.0 algorithms for Affymetrix oligo arrays and median adjustment normalization with background correction for the two-color technologies. Although in general the default procedures had slightly better accuracy (not statistically significant), the gains in precision given by the Bioconductor procedures were dramatic. For example, the SD assessment for Affy oligo lab 4 improved from 0.46 to 0.15. Furthermore, the agreement across technologies was much improved. For example the correlation between Affy oligo lab 4 and two-color cDNA lab 1 improved from 0.13 to 0.50. More detailed results are available in the supplemental material. Because of the great improvement provided by the Bioconductor procedures, we use them for all the results presented in this paper.

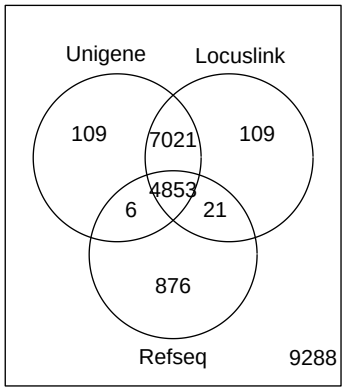

Figure 4: For each mapping (UNIGENE, LOCUSLINK, and REFSEQ) we obtain a different set of genes that have identifiers in each platform. This Venn diagram shows the agreement between these three different lists.

Each platform assigns a unique identifier to each of its features. However,

these identifiers do not match across platforms. To match features across platforms we use mappings matching features to genomic entities that are available from various public data-bases. Resourcerer[18] provides mappings linking features to UNIGENE, LOCUSLINK, and REFSEQ for all the platforms used by the labs participating in our experiment. Resourcerer also provides its own annotation called EGO. Unfortunately, none of these mappings are one to one: Not all the features in the arrays are annotated and some are annotated with more than one genomic identifier. Therefore, given a particular annotation only a subset of the array features will have an entry for each platform. Furthermore, these subsets differ depending on which annotation one uses. Figure 4 shows a Venn diagram representing the agreement between these subsets for UNIGENE, LOCUSLINK and REFSEQ.

The annotation used to match had an effect on the across platform agreement. For example, the correlation between measurements from Affy oligo lab 4 and two-color cDNA lab 1 were 0.39 to 0.44 when using UNIGENE and EGO respectively. More detailed results are available from the supplemental material. We found that using the intersection of all four annotation mappings provided the best agreement. All the analyses in this paper used the subset of genes obtained from this intersection.

3 Discussion

We defined a series of assessment measures and plots used to compare three leading microarray platforms. These were justified by questions to assess agreement for subsets of gene that are likely to pass this screenof to assess agreement for subsets of gene that are likely to pass this screenbiological to assess agreement for subsets of gene that are likely to pass this screeninterest and have practical interpretations. The signal measure represents the expected log (base 2) fold change of a gene that should be differentially expressed with nominal fold change of two and the SD measure gives us the expected log (base 2) fold change of a null gene. This combination gives us a clear idea of the signal to noise ratio. For example, perfect accuracy, signal=1, with lack of precision, say SD=1, is not useful in practice because distinguishing true positives, with large fold changes, from the large number of false positives reaching fold changes greater than two would be impossible. Notice that with SD=1 we would expect about 30% of the null genes to have fold changes greater than 2. Similarly, perfect precision, SD=0, combined with no ability to detect signal, signal=0, is even less useful since all fold changes,

regardless of true biological signal, would be identical.

Our assessment measures and plots demonstrated that, in general, the Affymetrix platform performed better than the two-color counterparts. However, we illustrated that labs using any of the two-color technologies can perform as well or better than the best performing Affymetrix lab. The main difference was that the Affymetrix platform was less susceptible to the lab effect. However, it is important to keep in mind that Affymetrix GeneChip arrays are substantially more expensive than the alternatives.

We also demonstrated that relatively good agreement is achieved between the Affymetrix labs and the best performing two-color labs. These results contradict some previously published papers that find disagreement across platforms[7, 8, 9, 10]. The conclusion reached by these studies are likely due to three misconceptions. The first misconception is that absolute measurements of expression can be used to assess across platforms. Notice that both studies using absolute measures found disagreement[7, 10]. Results established on absolute measures are misleading because they are adversely affected by *platform-dependent-probe-effects* which can be removed by considering relative measurements of expression instead. The statistical model used to motivate our assessment measures, described in Section 4, can be used to demonstrate this point (see Section 4 for more details). Notice that in all studies interested in differential expression, relative expression is the quantity of interest, thus this type of measurement is always available. The second misconception is that pre-processing has no significant effect on final results. With one exception[4] all previous studies used the default algorithms which have been shown to be inferior to alternatives developed by the academic community[16, 15]. Finally, the third misconception is that platform performance is not affected by lab. The the existence of the sizeable lab effect was ignored in all previously published comparison studies. This permits the possibility that studies using, for example, an experienced technicians may find agreement and studies using a less experienced technicians may find disagreement.

Although we found relatively good across platform agreement it is quite far from being perfect. In Figure 2A we see that for many of the genes that are differentially expressed there is agreement. However, there is a small group of genes that had relatively large fold changes for one technology but not for the other. Notice in particular the group of points with large negative vertical scale values and almost 0 horizontal scale values. In a recent report[19] (included as supplemental material) it was found that "in the small number of cases where there were discrepancies in the microarray profiles, qRT-PCR generally failed to confirm either result, suggesting that sequence-specific effects may make expression estimates

difficult for any technique." We conjecture that some genes are not correctly measured, not because the technologies are not performing adequately, but because transcript information and annotation can still be improved.

Our results provide a useful assessment of three leading technologies, it also demonstrates the need for continued cross-platform comparisons. We hope that our study serves as a starting point for larger, more comprehensive comparisons. More importantly, we expect our study to motivate the creation of some standards that can be used to assess performance of microarray labs. We believe this is essential for the success of microarray technology as a general measurement tool.

4 Methods

4.1 Sample Preparation

Each lab was given two identical aliquots from each of two different samples containing variable amounts of four individual PEX mutant cell lines. We will refer to the two large pooled samples as A and B respectively and to the aliquots as A_1 and A_2 or B_1 and B_2. For this experiment A_1 and A_2 are identical to each other while and B_1 and B_2 are also identical to each other but different from A_1 and A_2. RNA samples were provided in duplicate for the performance of technical replicate labeling and hybridization at each expert lab. The four samples will be denoted with A_1, A_2, B_1, and B_2.

RNA samples A and B were actually mixtures of four human cell lines, each deficient for a gene required for peroxisome biogenesis. The cell lines were PBD002, deficient for PEX1[20]; PBD106, deficient for PEX6[21]; PBD072, deficient for PEX7[22]; and PBD006, deficient for PEX12[23]. Nonsense mutations identified in the four cell lines used in this analysis are predicted to dramatically reduce mRNA levels for each gene through nonsense mediated RNA decay. The four human skin fibroblast cell lines had been previously transformed with SV40 large tumor antigen (T antigen) and have variable levels of aneuploidy. Master solutions of A and B total RNA at 1.3μg/μl were generated by mixing individual RNA preparations obtained from the 4 initial cell lines. The RNA was isolated according to manufacturer's instructions for TRIZOL (Invitrogen, Carlsbad CA) and was subsequently "cleaned up" using RNeasy (Qiagen. Valencia CA) . Sample A was a mixture composed of 1/2 PBD002, 2/5 PBD106, and 1/10 PBD072 and sample B was a mixture of 1/2 PBD106, 2/5 PBD072, and 1/10 PBD006. This mixture strategy was designed as though the contribution of mRNA for a given

PEX gene to sample A or B from a cell line carrying a mutation in that PEX gene was negligible and that mutations in one of these genes does not affect expression on any of the other three. Under this assumption, the expected ratios for the four genes are: 2:1, 3:2, 2:3 and 1:2 for PEX1, PEX6, PEX7, and PEX12 respectively. RNA integrity was verified by size distribution on gel electrophoresis, as well as cDNA synthesis efficiency (yield and size) by one of the participating labs. The samples were then aliquoted and distributed. The actual fold changes generated by mixing will deviate from ideal for at least two reasons. First, mRNA for the test PEX genes in mutant cells will not be precisely zero, although the level is expected to be low. Furthermore, aneuploidy in the SV40-transformed cells can contribute noise due to alterations in the amount of total RNA/average gene copy. During the RNA preparation step, sib cultures from each cell line were analyzed for karyotype. As it is common in such lines, aneuploid states were noted in each. To obtain a realistic expectation of fold changes we used RT-PCR as described in Section 4.2. The resulting fold-changes were 1.26, 1.50, 1/1.04, and 1/1.04 for PEX1, PEX6, PEX7, and PEX12 respectively.

4.2 RT-PCR

Real-time polymerase chain reaction (real-time RT-PCR) confirmation studies were done, for the four altered (PEX) genes, using total RNA of the same RNA population used for the microarray study. Fluorogenic LUX^{TM} primers (Invitrogen Life Technologies, Carlsbad, CA) were used to validate relative changes in mRNA levels by multiplex RT-PCR, using an ABI Prism 7700 Sequence Detection System (Applied Biosystems, Foster City, CA). To remove genomic DNA contamination, total RNA was treated with DNase I (Invitrogen Life Technologies, Carlsbad, CA) for 15 minutes at room temperature prior to cDNA synthesis using the SuperScript^{TM} First-Strand Synthesis System for RT-PCR (Invitrogen Life Technologies, Carlsbad, CA). First strand cDNA was synthesized from 1μg of total RNA. Multiplex PCR reactions were performed in triplicate using Platinum® Quantitative PCR SuperMix-UDG and ROX reference dye (Invitrogen Life Technologies, Carlsbad, CA). Reactions were performed in 25 μl reaction volume under the following conditions: $50°$C for 2 minutes, $95°$C for 2 minutes, and 45 cycles of $95°$, 15 seconds, $55°$C, 30 seconds, $72°$C, 30 seconds. Results were analyzed by the comparative Ct method (User Bulletin Number 2, ABI Prism Sequence Detector 7700, P/N 4303859). The resulting fold changes are reported in the previous Section and in Table 6 of the supplemental material.

4.3 Microarray Hybridization

Each participating lab used their standard operating procedure to label, hybridize and scan the samples provided. For Affymetrix labs, each sample was analyzed on a separate HGU133A GeneChip. Two-color cDNA labs 1 and 2 used the Human 20K clone set from Research Genetics to create spotted arrays, and lab 3 used custom spotted arrays containing 32,448 elements. The two-color oligo labs both used Human Genome Oligo Set Version 2.0. Each of the two-color labs adopted a co-hybridization and dye-swap scheme of their choice. Details about experimental design and hybridization procedures for each lab are given in the supplemental material.

4.4 Data Analysis

The assessment measures and Figures are motivated by a simple statistical model. This model also motivates the use of relative, as opposed to absolute, measurements of expression. A general model that describes the data well is

$$Y_{i,j,k} = \theta_i + \phi_{i,j} + \varepsilon_{i,j,k}, \tag{1}$$

where Y_{ijk} represents measurement k of log scale expression on gene i by platform j. Here, θ represents absolute gene expression in the log-scale. The platform specific probe or spot effect is denoted with ϕ_{ij}. Measurement error is represented with ε. Similar models have been described by various statistical groups[16, 24, 25]. To illustrate how an incorrect statistical assessment can lead to an incorrect conclusion and to motivate appropriate statistical assessments we will consider each of the effects in model (1) to be random. We will assume that they are mutually independent with variances σ^2_θ, σ^2_ϕ, and σ^2_ε for the expression level, probe effect and the measurement error respectively.

Many researchers have observed that the probe effect variance σ^2_ϕ is large[26]. In fact, for Affymetrix arrays it appears there is more variation due to probe effects than to different expression levels, i.e. $\sigma^2_\phi > \sigma^2_\theta$. This fact will result in large correlations when comparing measurements from the same technology. However, these high correlations are not to be interpreted as a positive aspect as they are driven by the probe effects. To see this notice that within platform correlation is given by:

$$\text{corr}(Y_{i,j,1}, Y_{i,j,2}) = \frac{\sigma^2_\theta + \sigma^2_\phi}{\sigma^2_\theta + \sigma^2_\phi + \sigma^2_\varepsilon}. \tag{2}$$

Table 2: Correlation and SD measures computed for absolute and relative measurements of expression. Affy oligo lab 4 and two-color cDNA lab 1 are used for this comparison.

	Correlation		SD	
	Absolute	Relative	Absolute	Relative
Affy oligo versus Affy oligo	0.98	0.79	0.16	0.15
two-color cDNA versus two-color cDNA	0.91	0.65	0.29	0.23
Affy oligo versus two-color cDNA	0.40	0.44	0.91	0.25

This correlation is typically close to 1, but only because σ_ϕ^2 is usually much larger than σ_θ^2 and σ_ε^2. Empirical results confirm this, see Table 2. If we compare across platforms the correlation will not be as large, but only because the probe effect is not common to the two platforms, for example Affymetrix probes are quite different from the typical spot on a two-color cDNA array, and therefore does not affect the correlation. In the across platform case, the correlation is given by:

$$\text{corr}(Y_{i,j,1}, Y_{i,j,2}) = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\phi^2 + \sigma_\varepsilon^2}. \tag{3}$$

Notice that the large σ_ϕ^2 term is no longer in the numerator. A smaller correlation is observed empirically, see Table 2.

A simple solution to the probe effect problem is to consider relative expression instead of absolute expression. Most experiments compare between different samples, for example treatment versus control, diseased versus healthy, etc... thus in general this type of measure is readily available.

Following model (1), we define the observed relative expression between samples A and B as

$$M_{i,j,k} = Y_{i,j,k}^A - Y_{i,j,k}^B = d_i + b_{i,j} + \delta_{i,j} \tag{4}$$

with $Y_{i,j,1}^A$ and $Y_{i,j,k}^B$ the absolute log-scale expression measure for samples A and B respectively. Because the Ys are log scale measurements the difference M is simply the log ratio of the absolute expression levels. The advantage of using this quantity is that, in theory, the platform dependent probe effect is canceled out. Here d_i represents the true amount of differential expression (log fold change), and $\delta_{i,j} = \varepsilon_{i,j,1} - \varepsilon_{i,j,2}$ represents measurement error. In practice, the probe-effect is not removed completely and thus we include the $b_{i,j}$ term to represent a

platform-dependent bias. As before the within platform correlations are given by:

$$\text{cor}(M_{i,j,1}, M_{i,j,2}) = \frac{\sigma_d^2 + \sigma_b^2}{\sigma_d^2 + \sigma_b^2 + \sigma_\delta^2} \tag{5}$$

and the across lab correlations by

$$\text{cor}(M_{i,1,k}, Y_{i,2,k}) = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_b^2 + \sigma_\delta^2}. \tag{6}$$

The correlation of the relative expression measurements can be seen in Table 2. Notice that the within platform correlations are substantially smaller and the across lab correlation is a bit larger. In theory $b_{i,j}$ should be 0, but given the different ways transcripts are defined and the arrays constructed it is no surprise that there is some positive variance σ_b^2 which explains why the correlations in column 2 of Table 2 are not all the same. However, the data confirms that σ_b^2 is much smaller than σ_ϕ^2.

The above model calculations demonstrate that results based on absolute expression measurements are misleading because they are adversely affected by the platform specific probe effect. We propose that only assessments based on relative expression are useful. All the results presented in this paper deal with relative expression M as defined by equation (4).

The code and data necessary to reproduce the results presented in this paper are available from `http://www.biostat.jhsph.edu/~ririzarr/techcomp`.

References

1. Kane, M., Jatkoe, T., Stumpf, C., Lu, J., Thomas, J., and Madore, S. Assessment of the sensitivity and specificity of oligonucleo-tide (50mer) microarrays. *Nucleic Acid Research* **28**(22), 4552 (2000).

2. Hughes, T., Mao, M., Jones, A., Burchard, J., Marton, M., Shannon, K., Lefkowitz, S., Ziman, M., Schelter, J., Meyer, M., Kobayashi, S., Davis, C., Dai, H., He, Y., Stephaniants, S., Cavet, G., Walker, W., West, A., Coffey, E., Shoemaker, D., Stoughton, R., Blanchard, A., Friend, S., and Linsley, P. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology* **19**(4), 342–347 (2001).

3. Yuen, T., Wurmbach, E., Pfeffer, R. L., Ebersole, B. J., and .Sealfon, S. C. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Research* **30**(10), e48 (2002).

4. Barczak, A., Rodriguez, M. W., Hanspers, K., Koth, L. L., Tai, Y. C., Bolstad, B. M., Speed, T. P., and Erle, D. J. Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Research* **13**(7), 1775–1785 (2003).

5. Carter, M., Hamatani, T., Sharov, A., Carmack, C., Qian, Y., Aiba, K., Ko, N., Dudekula, D., Brzoska, P., Hwang, S., and Ko, M. In situ-synthesized novel microarray optimized for mouse stem cell and early developmental expression profiling. *Genome Research* **13**(3), 1011–21 (2003).

6. Wang, H., Malek, R., Kwitek, A., Greene, A., Luu, T., Behbahani, B., Frank, B., Quackenbush, J., and Lee, N. Assessing unmodified 70-mer oligonucleotide performance on glass-slide microarrays. *Genome Biology* **4**(1), R5 (2003).

7. Kuo, W., Jenssen, T., Butte, A., Ohno-Machado, L., and Kohane, I. Analysis of mrna measurements from two different microarray technologies. *Bioinformatics* **18**(3), 405–412 (2002).

8. Kothapalli, R., Yoder, S., Mane, S., and Jr., L. T. Microarray results: how accurate are they? *BMC Bioinformatics* **3**(1), 22 (2002).

9. Li, J., Pankratz, M., and Johnson, J. Differential gene expression patterns revealed by oligo-nucleotide versus long cDNA arrays. *Toxicological Sciences* **69**(2), 383–390 (2003).

10. Tan, P., Downey, T., Spitznagel, E. J., Xu, P., Fu, D., Dimitrov, D., Lempicki, R., Raaka, B., and Cam, M. Evaluation of gene expression measurements from commercial platforms. *Nucleic Acids Research* **31**(19), 5676–5684 (2003).

11. Youden, W. Enduring values. *Technometrics* **14**, 1–11 (1972).

12. Parmigiani, G., Garrett-Mayer, E., Anbazhagan, R., and Gabrielson, E. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clinical Cancer Research* **10**(9), 2922–2927 (2004).

13. Lee, J., Bussey, K., Gwadry, F., Reinhold, W., Riddick, G., Pelletier, S., Nishizuka, S., Szakacs, G., Annereau, J., Shankavaram, U., Lababidi, S., Smith, L., Gottesman, M., and Weinstein, J. Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the nci-60 cancer cells. *Genome Biology* **4**(12), R82 (2003).

14. Moreau, Y., Aerts, S., De Moor, B., De Strooper, B., and Dabrowski, M. Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends in Genetics* **19**(10), 570–577 (2003).

15. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–64 (2003).

16. Dudoit, S., Yang, Y. H., Luu, P., Lin, D. M., Peng, V., Ngai, J., , and Speed, T. P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**(4), e15 (2002).

17. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Li, F. L. C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80 (2004).

18. Tsai, J., Sultana, R., Lee, Y., Pertea, G., Karamycheva, S., Antonescu, V., Cho, J., Parvizi, B., Cheung, F., and Quackenbush, J. RESOURCERER: a database for annotating and linking microarray resources within and across species. *Genome Biology* **2**, software0002.1–0002.4 (2001).

19. Larkin, J. E., Frank, B. C., Gavras, H., and Quackenbush, J. Independence and reproducibility across microarray platforms. *Submitted* **0**, 0–0 (2004).

20. Collins, C. and Gould, S. Identification of a common pex1 mutation in zellweger syndrome. *Human Mutatation* **14**, 45–54 (1999).

21. Yahraus, T., Braverman, N., Dodt, G., Kalish, J., Morrell, J., Moser, H., Vall, e. D., and Gould, S. The peroxisome biogenesis disorder group 4 gene, px-

aaa1, encodes a cytoplasmic atpase required for stability of the pts1 receptor. *EMBO Journal* **15**, 2914–2923 (1996).

22. Braverman, N., Chen, L., Lin, P., Obie, C., Steel, G., Douglas, P., Chakraborty, P. K., Clarke, J. T., Boneh, A., Moser, A., Moser, H., and Valle, D. Mutation analysis of pex7 in 60 probands with rhizomelic chondrodysplasia punctata and functional correlations of genotype with phenotype. *American Journal of Human Genetics* **20**, 284–297 (2002).

23. Chang, C. and Gould, S. Phenotype-genotype relationships in complementation group 3 of the peroxisome-biogenesis disorders. *American Journal of Human Genetics* **63**, 1294–1306 (1998).

24. Chu, T., Weir, B., and Wolfinger, R. A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences* **176**(1), 35–51 (2002).

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

26. Li, C. and Wong, W. H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci.* **98**, 31–36 (2001).