Correspondence

Feature-level exploration of a published Affymetrix GeneChip control dataset

Rafael A Irizarry*, Leslie M Cope† and Zhijin Wu‡

A comment on **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset** by SE Choe, M Boutros, AM Michelson, GM Church and MS Halfon. *Genome Biology* 2005, **6:**R16.

Addresses: *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205-2179, USA. †Department of Oncology, The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, 550 N. Broadway, Suite 1131 Baltimore, MD 21205, USA. ‡Center for Statistical Sciences, Department of Community Health, Brown University, 167 Angell Street, Providence, RI 02912, USA.

Correspondence: Rafael A Irizarry. Email: rafa@jhu.edu

Published: 1 September 2005

Genome ***Biology*** 2006, **7:**404 (doi:10.1186/gb-2006-7-8-404)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/8/404

In a recent *Genome Biology* article, Choe *et al.* [1] describe a spike-in experiment that they use to compare expression measures for Affymetrix GeneChip technology. In this work, two sets of triplicates were created to represent control (C) and experimental (S) samples. We describe here some properties of the Choe *et al.* [1] control dataset one should consider before using it to assess GeneChip expression measures. In [2] and [3] we describe a benchmark for such measures based on experiments developed by Affymetrix and GeneLogic. These datasets are described in detail in [2]. A web-based implementation of the benchmark, is available at [4]. The experiment described in [1] is a worthy contribution to the field as it permits assessments with data that is likely to better emulate the nonspecific binding (NSB) and cross-hybridization seen in typical experiments. However, there are various inconsistencies between the conclusions reached by [1] and [3] that we do not believe are due to NSB and cross-hybridization effects. In this Correspondence we describe certain characteristics of the feature-level data produced by [1] which we believe explain these inconsistencies. These can be divided into characteristics induced by the experimental design and an artifact.

Experimental design

There are three characteristics of the experimental design described by [1] that one should consider before using it for assessments like those carried out by Affycomp. We enumerate them below and explain how they may lead to unfair assessments. Other considerations are described by Dabney and Storey [5].

First, the spike-in concentrations are unrealistically high. In [3] we demonstrate that background noise makes it harder to detect differentially expression for genes that are present at low concentrations. We point out that in the Affymetrix spike-in experiments [2,3] the concentrations for spiked-in features result in artificially high intensities but that a large range of the nominal concentrations are actually in a usable range (Figure 1a of this Correspondence). Figure 1b demonstrates that in a typical experiment [6], features related to differentially expressed genes show intensities with a similar range as the rest of the genes - in particular, that less than 10% of genes, including the differentially expressed genes, are above intensities of 10. Figure ADF5-3 in the Additional data files for [1] shows that less than 20% of their spiked-in gene intensities are below 10. Additional data file 5 of [1] also contains a reanalysis using only the lower-intensity genes, which provide results that agree a bit better with results from Affycomp. A problem is that for the Affycomp assessment one needs to decide *a priori* which genes to include in the analysis, for example, setting a cutoff based on nominal spike-in concentration. In the analysis described in Additional data file 5 of [1] one needs to choose genes *a posteriori*, that is, based on observed intensities. The latter approach can easily lead to problems such as favoring the inclusion of probesets exhibiting low intensities as a result of defective probes. Furthermore, our Figure 1c shows that, despite the use of an experimental design that should induce about 72% of absent genes, we observe intensities for which the higher percentiles (75-95%) are twice as large as what we observe in typical experiments. This suggests that the spike-in concentrations were high enough to make this experiment produce atypical data. We do not expect a preprocessing algorithm that performs well on this data to necessarily perform well in general, and vice versa.

Second, a large percentage of the genes (about 10%) are spiked-in to be differentially expressed and all of these are expected to be upregulated. This design makes this spike-in data very different from that produced by many experiments where at least one of the following assumptions is expected to hold: a small percentage of genes are differentially expressed, and there is a balance between up- and downregulation. Many preprocessing algorithms (for example, loess normalization, variance stabilizing normalization (VSN), rank-invariant) implement normalization routines motivated by one or both of these assumptions; thus we should not expect many of the existing expression measure methodologies to perform well with the Choe *et al.* [1] data.

Third, a careful look at Table 1 in [1] shows that nominal concentrations and fold-change sizes are confounded. This problem will slightly cloud the distinction between ability to detect small fold changes from the ability to detect differential expression when concentration is low. Why this distinction is important is shown in [3]. However, Figure ADF5-1 in Additional data file 1 of Choe *et al.* [1] demonstrates that this difference in nominal concentrations does not appear to translated into observed intensities. This could, however, be an indication of saturation, which is a common problem when high intensities are observed (see the first point of this argument above). One case of the confounding is seen: genes with nominal fold-changes larger than 1 result in intensities that, on average, are about three times larger than genes with nominal fold-changes of 1.

The artifact

Figure 1a-c of this Correspondence is based on raw feature-level data. No preprocessing or normalization was performed. We randomly selected 100 pairs of arrays from experiments stored in the Gene Expression Omnibus (GEO) and without exception they produced MA-plots similar to those seen in Figure 1a,b (MA-plots are log expression in treatment minus (M) log expression in control versus average (A) log expression plots). These plots have most of the points in the lower range of concentrations and an exponential tapering as concentration increases [7]. However, the Choe *et al.* [1] data show a second cluster centered at a high concentration and a negative log ratio. Not one of the MA-plots from GEO looked like this. Figure 2 in this Correspondence reveals that the feature intensities for genes spiked-in to be at 1:1 ratios behave very

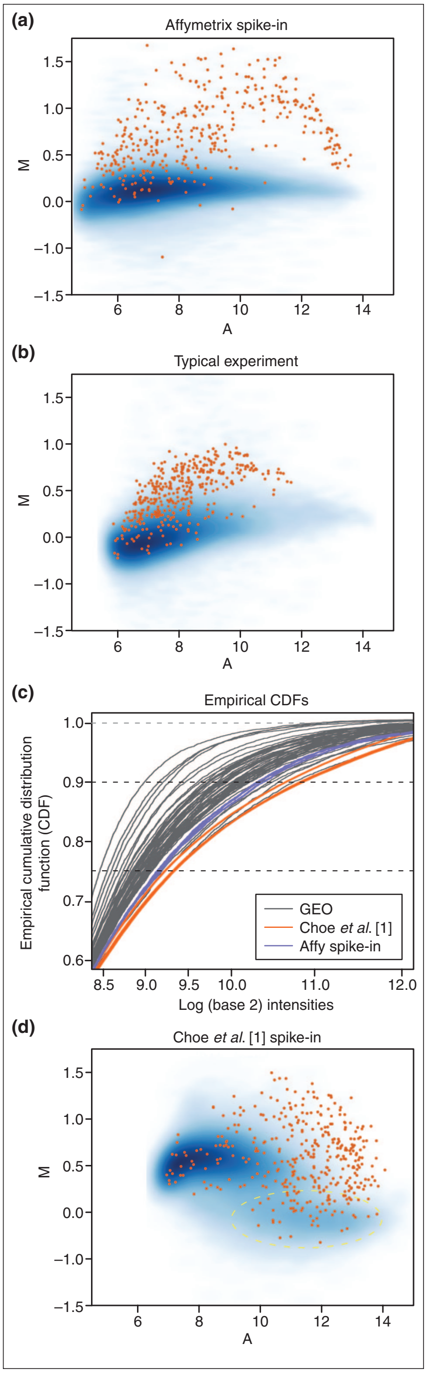

Figure 1
MA and cumulative distribution function (CDF) plots. MA-plots are log expression in treatment minus (M) log expression in control versus average (A) log expression plots. **(a)** For two sets of triplicates from the Affymetrix HGU133A spike-in experiment [2,3] we calculated the average log ratio across the three comparisons (M) and the average log intensity (A) across all six arrays for each feature. The figure shows M plotted against A. However, because there are hundreds of thousands of features, instead of plotting each point, we use shades of blue to denote the amount of points in each region of the plot. About 90% of the data is contained in the dark-blue regions. Orange points are the 405 features from the 36 genes with nominal fold changes of 2. **(b)** As in (a) but using two sets of biological triplicates from a study comparing three trisomic human brains to three normal human brains. The orange dots are 385 features representing 35 genes on chromosome 21 for which we expect fold changes of 1.5. **(c)** Empirical cumulative density functions for the median corrected log (base 2) intensities of 50 randomly chosen arrays from the Gene Expression Omnibus (GEO), three randomly selected arrays from Affymetrix HGU133A spike-in experiment, and the three S samples from Choe *et al.* [1] facilitate the comparison; the intensities were made to have the same median. The dashed black horizontal lines show the 75% and 95% percentiles. **(d)** As (a) but showing the two sets of triplicates described by Choe *et al.* [1]. The orange dots are 375 features randomly sampled from those that were spiked-in to have fold changes greater than 1. The yellow ellipse is used to illustrate an artifact: among the data with nominal fold changes of 1, there appear to be two clusters having different overall observed log ratios.

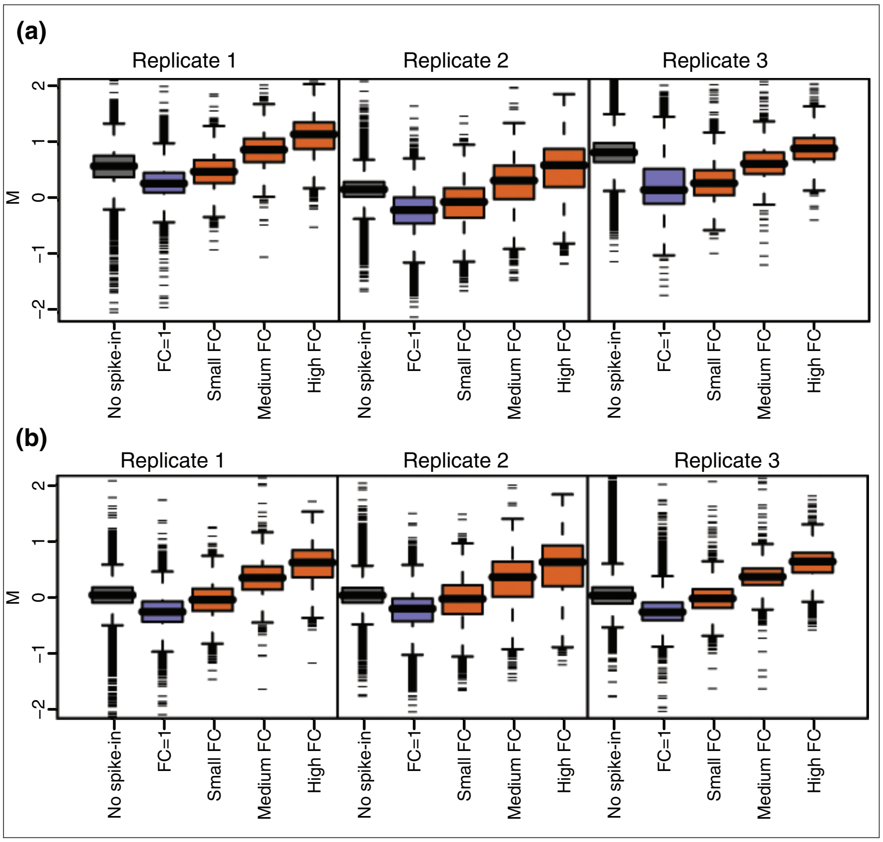

Figure 2

Log-ratio box-plots. **(a)** For the raw probe-level data in [1] we computed log fold changes comparing the control and spike-in arrays for each of the three replicates. The C and S arrays were paired according to their filenames: C1-S1, C2-S2, and C3-S3. Box-plots are shown for five groups of probes: not spiked-in (gray), spiked-in at equal concentrations (purple), spiked-in with nominal fold-changes between 1 and 2, 2 and 3, and 3 and 4 (orange). **(b)** As (a) but after quantile normalizing the probes.

differently from the features from non-spiked-in genes which, in a typical experiment, exhibit, on average, log fold changes of 0 (in practice there are shifts, some nonlinear, but standard normalization procedures correct this).

This problem implies that, unless an *ad hoc* correction is applied, what Choe *et al.* [1] define as false positive might in fact be true positives. Figure 2 shows that this problem persists even after quantile normalization [8]. In Choe *et al.* [1] a normalization scheme based on knowledge of which genes have fold-changes of 1 is used to correct this problem. However, preprocessing algorithms are not designed to work with data that has been manipulated in this way, which makes this dataset particularly difficult to use in assessment tools such as Affycomp. Furthermore, Figure 1c,d of this Correspondence shows that the data produced by [1] is quite different from data from typical experiments for which most preprocessing algorithms were developed.

Currently, experiments where the normalization assumptions do not hold seem to be a small minority. However, our experience is that they are becoming more common. For this type of experiment we will need new preprocessing algorithms, and the Choe *et al.* [1] data may be useful for the development of these new methods.

Additional data files

Additional data file 1 contains MA plots for 100 randomly chosen pairs of arrays from the Gene Expression Omnibus (GEO) is available online with this Correspondence.

Acknowledgements

The work of R.A.I. is partially funded by the National Institutes of Health Specialized Centers of Clinically Oriented Research (SCCOR) translational research funds (212-2492 and 212-2496).

Sung E Choe, Michael Boutros, Alan M Michelson, George M Church and Marc S Halfon respond:

Irizarry *et al.* raise a number of interesting points in their Correspondence that highlight the continued need for carefully designed control microarray experiments. They posit that "the spike-in concentrations are unrealistically high" in our experimental design. Although we have estimated that the average per-gene concentration is similar to that in a typical experiment [1], we do not know individual RNA concentrations and so cannot verify or deny this assertion. Since the majority of probesets in our dataset correspond to non-spiked-in genes, and therefore have a signal range consistent with absent genes, we think it seems reasonable that the spiked-in genes have higher signal than the rest of the chip. Regardless of this, in Additional Data File 5 of [1], we repeated the receiver-operator characteristics (ROC) analysis using as the "known differentially expressed" probe sets only the subset with low signal levels. The results we obtained for gcrma (robust mutli-array average using sequence information) [9] were very similar to the conclusions in [3] and [10]; in addition, the performance of MAS5 [11] was similar between [1] and [10]. The inconsistencies between the different studies may therefore be less extreme than they seem. In particular, we think that a large source of the disagreement between [1] and [3] is simply the different choice of metric for the ROC curves.

There is no question that our analysis of low-signal-intensity probesets as well as the specific selection of non-differentially expressed genes to use for normalization purposes required prior knowledge of the composition of the dataset. This, of course, is one of the great strengths of a wholly-defined dataset such as that from [1] - we can choose idealized conditions for assessing the performance of different aspects of the analysis. Unfortunately, as Irizarry *et al.* correctly point out, it also makes it difficult to use for certain other types of assessment, such as those provided by Affycomp [3].

A more critical consideration lies in the point raised by Irizarry *et al.* that our dataset violates two main assumptions of most normalization methods: that a small fraction of genes should be differentially expressed; and that there should be roughly equal numbers of up- and down regulated genes. It is important to note that these two assumptions are just that - assumptions - and ones that are extremely difficult to prove or disprove in any given microarray experiment. Thus there is an inherent circularity in the design of analysis algorithms that explicitly rely on these assumptions: they perform well on data assumed to have the properties based on which they are designed to perform well. This is an issue all too often overlooked in the microarray field. The violation of these two core assumptions seen in our dataset may be more common than generally appreciated; certainly we can conceive of many situations in which they are unlikely to hold (for example, when comparing different tissue types, in certain developmental time courses, or in cases of immune challenge). Developing assumption-free normalization methods, and diagnostics to assess the efficacy of the normalization used for a given dataset (see [12] for an example), should thus be important research priorities.

This discussion underscores the need for more control datasets that specifically address matters of RNA concentration, fractions of differentially expressed genes, direction of changes in gene regulation, and the like. Only then can we truly devise and assess the performance of analysis methods for the large variety of possible scenarios encountered in the course of conducting microarray experiments focused on real biological problems.

Correspondence should be sent to Marc S Halfon: Department of Biochemistry and Center of Excellence in Bioinformatics and the Life Sciences, State University of New York at Buffalo, Buffalo, NY 14214, USA. Email: mshalfon@buffalo.edu

References

1. Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS: **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset.** *Genome Biol* 2005, **6:**R16.
2. Cope L, Irizarry R, Jaffee H, Wu Z, Speed T: **A benchmark for Affymetrix GeneChip expression measures.** *Bioinformatics* 2004, **20:**323–331.
3. Irizarry R, Wu Z, Jaffee H: **Comparison of Affymetrix GeneChip expression measures.** *Bioinformatics* 2006, **22:**789-794.
4. **Affycomp II**: **A benchmark for Affymetrix GeneChip expression measures** [http:affycomp.biostat.jhsph.edu]
5. Dabney A, Storey J: **A reanalysis of a published Affymetrix GeneChip control dataset.** *Genome Biol* 2006, **7:**401.
6. Saran NG, Pletcher MT, Natale JE, Cheng Y, Reeves RH: **Global disruption of the cerebellar transcriptome in a Down syndrome mouse model.** *Hum Mol Genet* 2003, **12:**2013-2019.
7. **One hundred MA plots from GEO** [http://www.biostat.jhsph.edu/~ririzarr/papers/hundredMAs.pdf]
8. Bolstad B, Irizarry R, Åstrand M, Speed T: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19:**185-193.
9. Wu Z, Irizarry R, Gentleman RC, Martinez-Murillo F, Spencer F: **A model-based background adjustment for oligonucleotide expression arrays.** *Journal of the American Statistical Association* 2004, **99:**909-917.
10. Qin LX, Beyer RP, Hudson FN, Linford NJ, Morris DE, Kerr KF: **Evaluation of methods for oligonucleotide array data via quantitative real-time PCR.** *BMC Bioinformatics* 2006, **7:**23.
11. **GeneChip Expression Analysis: data analysis fundamentals** [http://www.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf]
12. Gaile DP, Miecznikowski JC, Choe SE, Halfon MS: **Putative null distributions corresponding to tests of differential expression in the Golden Spike dataset are intensity dependent. Technical report 06-01.** Buffalo, N.Y.: Department of Biostatistics, State University of New York; 2006, [http://sphhp.buffalo.edu/biostat/research/techreports/UB_Biostatistics_TR0601.pdf].