3-17-2006

FEATURE-LEVEL EXPLORATION OF THE CHOE ET AL. AFFYMETRIX GENECHIP CONTROL DATASET

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

Leslie Cope
Johns Hopkins University School of Medicine, Oncology Center, Lcope1@jhmi.edu

Zhijin Wu
Brown University, Center for Statistical Science, Zhijin_Wu@brown.edu

Feature-level exploration of the Choe et al. Affymetrix GeneChip control dataset

Rafael A. Irizarry, Leslie Cope, and Zhijin Wu

Headline

We describe why the Choe et al. control dataset should not be used to assess GeneChip expression measures.

Introduction

In [1] a spike-in experiment is described which the authors use to compare expression measures for Affymetrix GeneChip technology. Two sets of triplicates were created to represent control (C) and experimental (S) samples. In [2] and [3] we describe a benchmark for such measures based on experiments developed by Affymetrix and a GeneLogic. These datasets are described in detail in [2]. A web-based implementation of the benchmark, is available at `affycomp.biostat.jhsph.edu`. There are various inconsistencies between the conclusions reached by [1] and [3]. In this letter we describe certain characteristics of the feature-level data produced by [1] which we believe explain these inconsistencies. These can be divided into 1) induced by the experimental design and 2) an artifact.

Experimental design

There are three characteristics of the experimental design described by [1] make the resulting data inappropriate for assessment. Below we enumerate these problems and explain how they lead to unfair assessments. Other problems with the experimental design are described by [4].

1. 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 in low concentrations. In [3] we point out that in the Affymetrix spike-in experiments the concentrations for spiked-in features are artificially high but that a large number of these are actually in a usable range (See Figure 1A). Figure 1B demonstrates that in a typical experiment, features related to differentially expressed genes show intensities with a similar range as the rest of the genes. However, Figures 1C-D suggest that none of the genes spiked-in by [1] are in a usable range since less than 1% of the data would reach the intensity levels seen for the spiked-in genes. This implies that expression measure assessments based on this dataset only apply to unlikely situations where we expect differentially expressed genes to be in the top 1% of overall expression.

2. A large percentage of the genes (about 10%) are spiked-in to be differentially expressed and all of these are expected to be up-regulated. This design makes this spike-in data very different from those produced by typical experiments where at least one of the following assumptions is expected to hold: 1) a small percentage of genes are differentially expressed, 2) there is a balance between up and down regulation, and 3) the gene expression distribution across arrays is roughly the same. Most preprocessing algorithms implement normalization routines motivated by one or more of these assumptions, thus we should not expect existing expression measure methodology to perform well with the Choe et al. data.

3. A careful look at Table 1 in [1] shows that nominal concentrations and fold change sizes are confounded. This is better demonstrated by a graphical representation (Figure 2). This problem will not permit us to distinguish ability to detect small fold changes from the ability to detect differential expression when concentration is low. [3] show why this distinction is important.

The artifact

Figure 1 are based on raw feature-level data. No preprocessing or normalization was performed. A typical expression data will produce MA-plots with most of the points in the lower-range of concentration and an exponential tapering as concentration grows. However, the Choe et. al data shows a second cluster centered at a high concentration and a negative log ratio. Figure 3 reveals that the features spiked-in to be at 1:1 ratios behave very differently from the features from non-spiked-in genes which should also have a nominal log fold change of 0. Without more sample preparation information it is impossible to give an explanation

for this artifact. This problem implies that what [1] define as false positive might in fact be true positives. Figure 3 shows that this problem persists even after quantile normalization [5]. This artifact makes this dataset particularly difficult to use in assessment because the distinction between true and false positives is not clear.

Abbreviations

- MA-plot - Log expression in treatment minus (M) log expression in control versus average (A) log expression plot.

Acknowledgments

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).

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**(2):R16.

[2] Cope L, Irizarry R, Jaffee H, Wu Z, Speed T: **A benchmark for Affymetrix GeneChip expression measures**. *Bioinformatics* 2004, **20**(3):323–331.

[3] Irizarry R, Wu Z, Jaffee H: **Comparison of Affymetrix GeneChip Expression Measures**. Dept. of biostatistics working papers, Johns Hopkins University 2005. [`www.bepress.com/jhubiostat/paper86`].

[4] Dabney A, Storey J: **A re-analysis of the Choe et al. Affymetrix GeneChip control dataset**. *Submitted* 2005.

[5] Bolstad B, Irizarry R, Åstrand M, Speed T: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinfromatics* 2003, **19**(2):185–193.

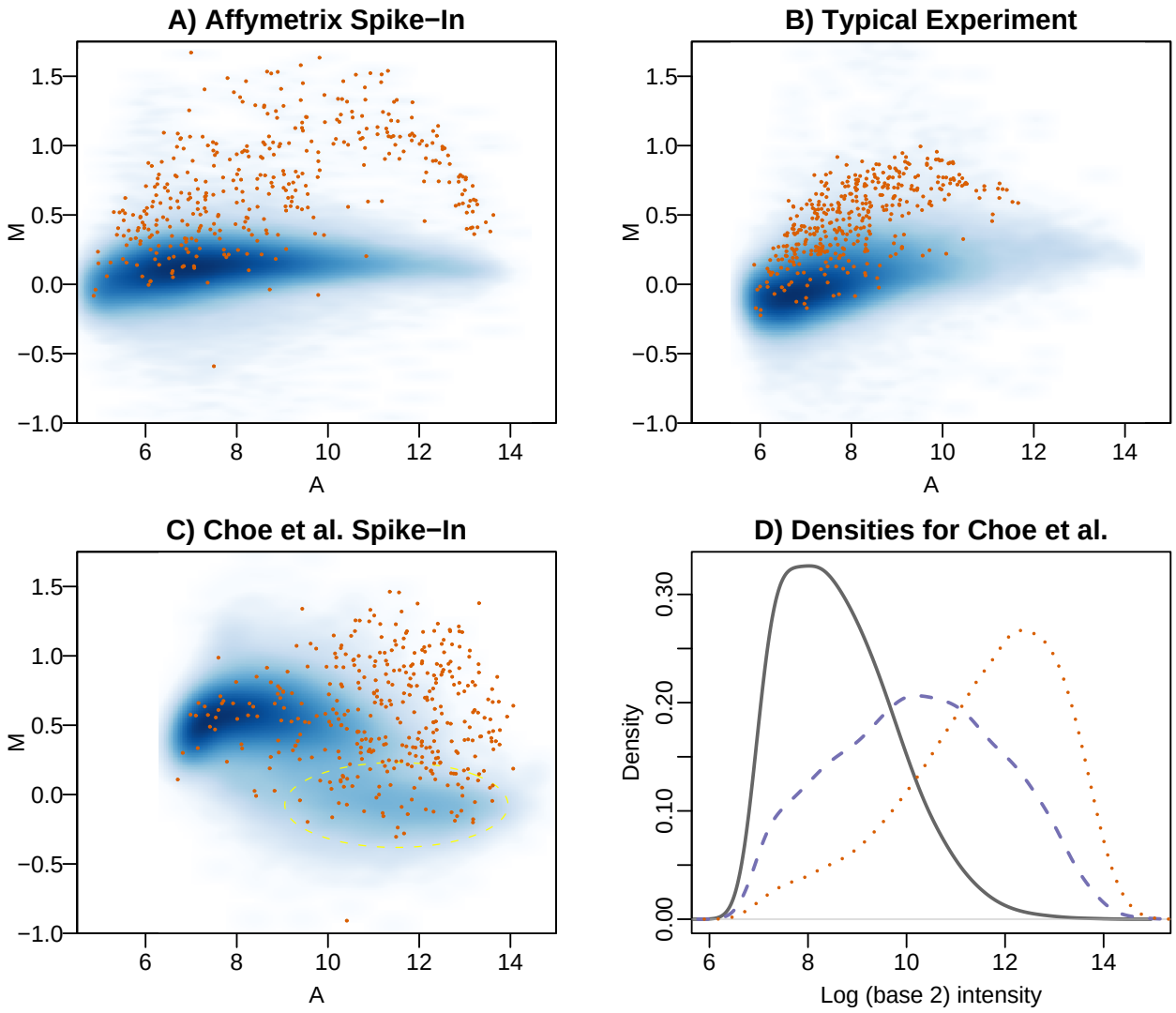

Figure 1: A) For two sets of triplicates from the Affymetrix HGU133A spike-in experiment we calculated the average log ratio across the three comparisons (M) and the average log intensity (A) across all six arrays for each feature. This figure shows M plotted against A. However, because there are hundreds of thousands of feature, 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. 405 features from the 36 genes with nominal fold-changes of 2 are shown as orange points. B) As Figure 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 3/2. C) As A) but showing the two sets of triplicates described by Choe et al. 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 appears to be two clusters having different overall observed log ratios. D) Empirical density estimators for the log intensity data, in one of S samples in the Choe el al data, for non-spiked in features (solid/gray), features spiked-in but to be at a 1:1 ratio (dashed/purple), and features spiked-in to have nominal fold changes greater than 1 (dotted/orange).

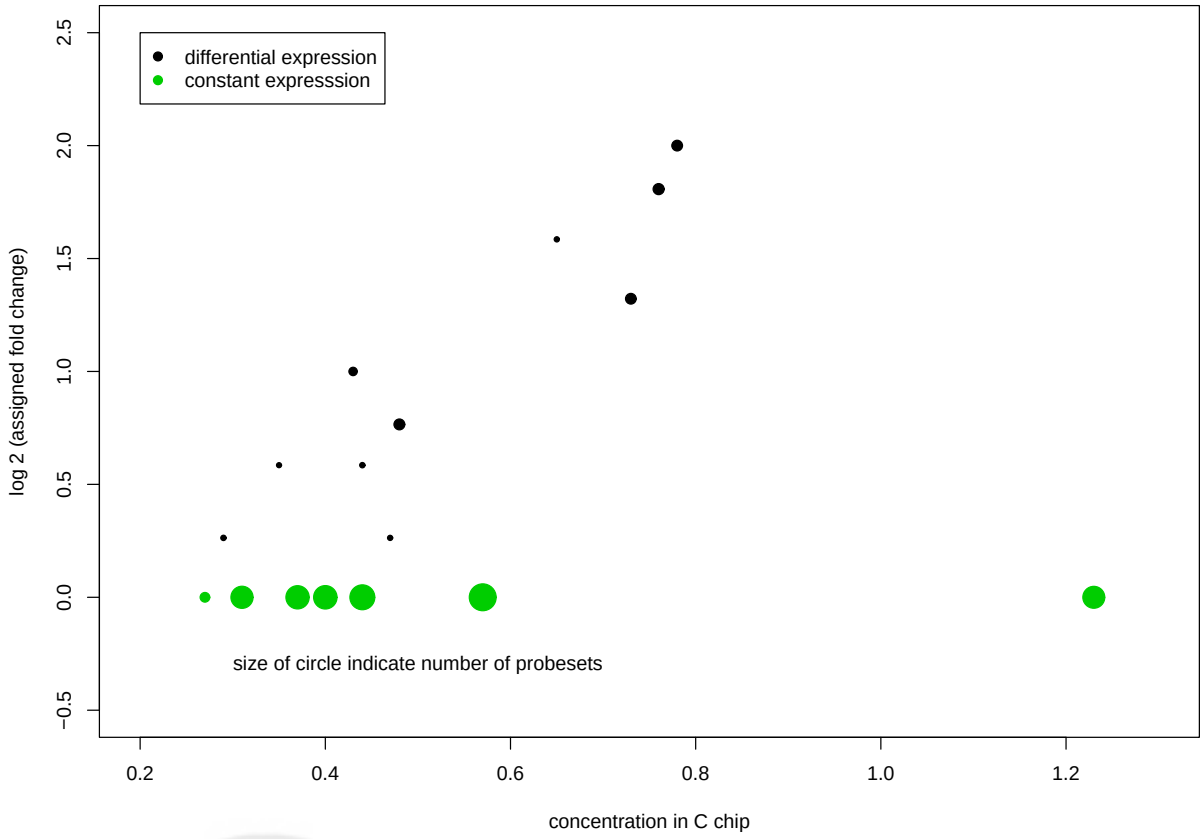

Figure 2: Graphical representation of Table 1 in [1]. The coordinates of the points represent nominal fold changes and concentrations that were used. The size of the points are proportional to the number of genes having those values.

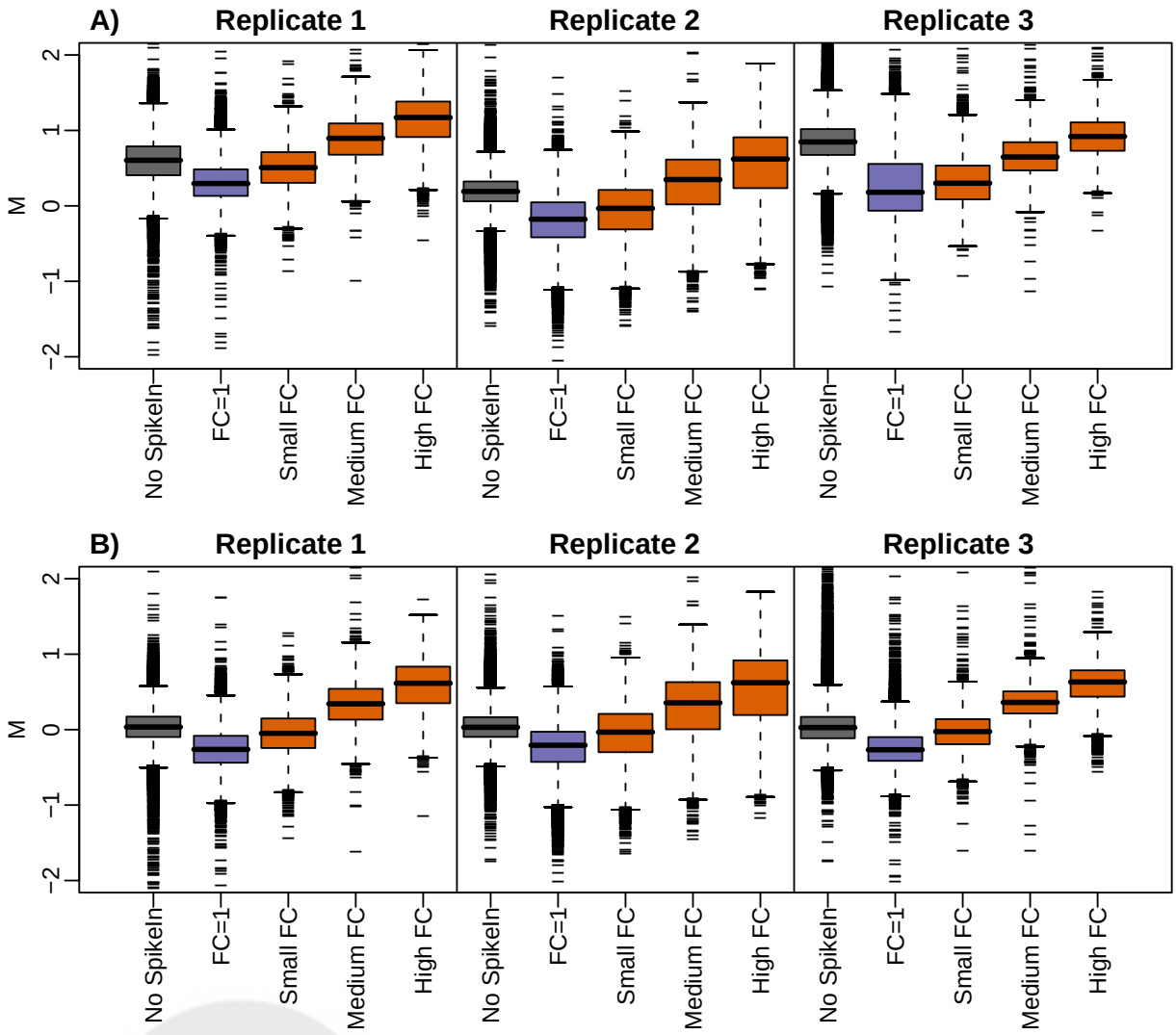

Figure 3: A) For the raw probe-level data in [1] we computed log fold changes comparing the control and spike-in arrays for each of the three replicates. The C and S arrays were paired according to their filenames: C1-S1, C2-S2, and C3-S3. Boxplots are shown for five groups of probes: not spiked-in (gray), 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.