BIOINFORMATICS

Vol. 00 no. 00 2004
Pages 1–7

Supplement to "A learned comparative expression measure for Affymetrix GeneChip DNA microarrays"

Will Sheffler [a] *Eli Upfal* [b] *John Sedivy* [c] *William Stafford Noble* [a,d]

[a]*Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, WA, 98195, USA,* [b]*Department of Computer Science, Brown University, Providence, RI, USA,* [c]*Department of Molecular and Cell Biology and Biochemistry, Brown University, Providence, RI, USA,* [d]*Department of Computer Science and Engineering, University of Washington, Seattle, WA, 98195, USA*

ABSTRACT

- Figure 1 shows the consistency of upward and downward measurements using LCEM.
- Figures 2 and 3 show histograms of fold changes in the three data sets used in our work. Examples from the Affymetrix and GeneLogic spike-in data sets contain examples of large fold changes, while the examples we construct from the GeneLogic dilution study contain more challenging examples.
- Figure 4 shows ROC plots for the Affymetrix latin square examples with and without the set of 56 outliers shown in Table 1. Performance with outliers included shows that LCEM is more robust to these outliers than MAS5 and RMA.
- Figures 5 and Figure 6 show comparative expression measure plotted against average expression. Figure 5 shows unchanged examples in the Affymetrix latin square data set, and Figure 6 shows changed examples from the dilution data set. These figures show that the MAS5 log ratio statistic has a strong bias towards false positives on low expressors, whereas LCEM shows no such bias.
- Figure 7 shows four data patterns on which RMA and LCEM disagree. Two are from changed examples and two are from unchanged examples.

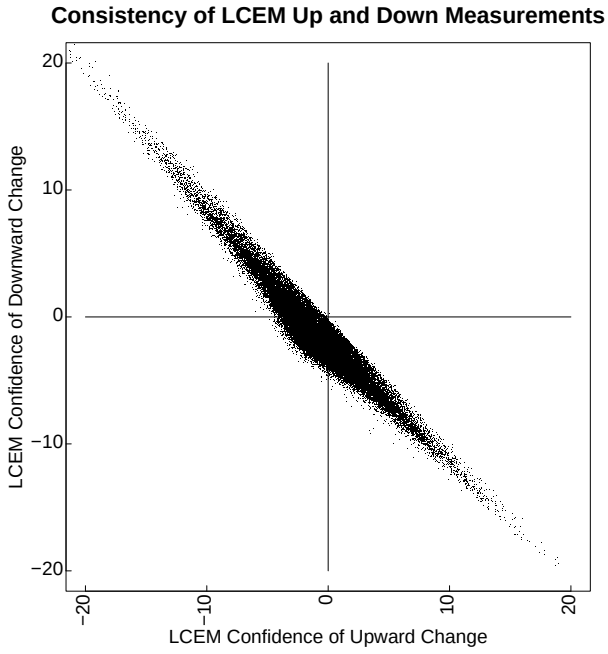

Fig. 1. Consistency of LCEM and LCEM Reverse Shown is a scatterplot of LCEM values for reversed data vectors versus unchanged data vectors from an independent data set (O'Connell et al., 2003). It is theoretically possible that separate measurement for upward and downward expression changes could cause inconsistency. However, this figure shows no inconsistencies, because positive measurements of upward and downward change do not occur simultaneously. In some cases, negative measurements occur for both upward and downward expression change, but this does not lead to inconsistency because those genes will be classified as unchanged.

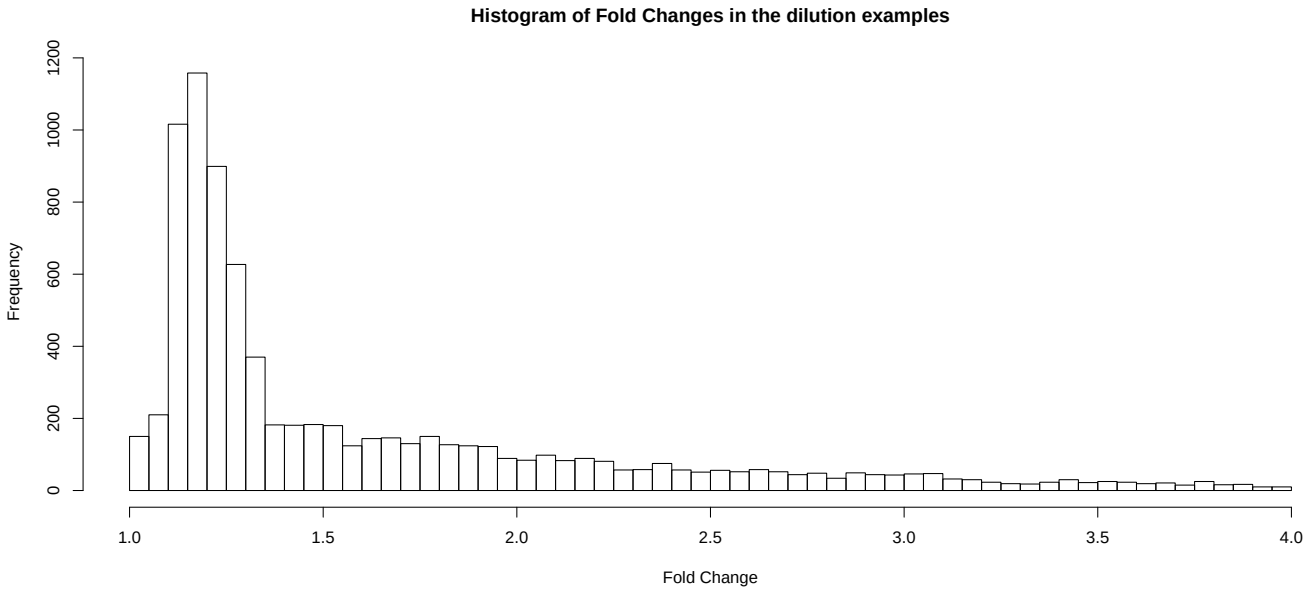

Fig. 2. Fold changes below 4 in the dilution examples Fold changes shown in the figure are estimates based on average RMA expression levels of the 100% liver and 100% CNS samples and adjusted to reflect mixture of the samples. Changes below 2-fold make up approximetly 71% of the examples and about half are below 1.5 fold. Most commonly represented are differences of about 20%.

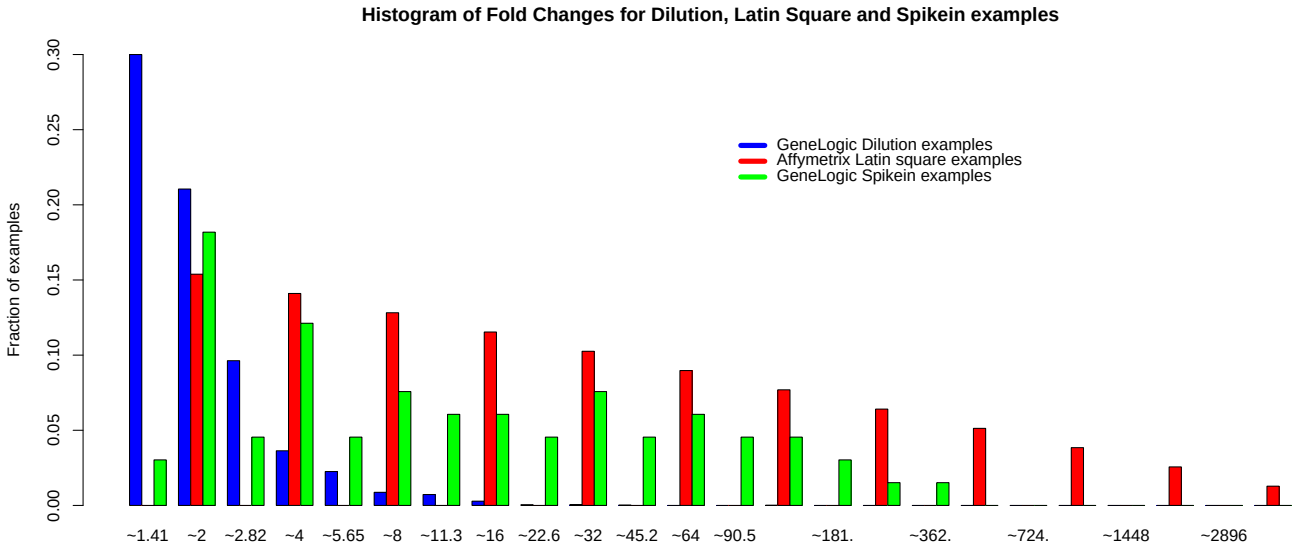

Fig. 3. Fold changes in all three example sets The first bar for the dilution examples has been truncated from height 0.61 to height 0.3 for clarity. The histogram is shown on a log scale with dilution examples in blue, GeneLogic spike-in examples in green, and Affymetrix latin square examples in red. The difference between the data sets is clear. The latin square design of the Affymetrix data set leads to a linearly decreasing number of examples as fold changes double. The less structured GeneLogic data set contains fewer large fold changes and a few below 2-fold, but the great majority of examples are still quite of quite large fold changes. The GeneLogic dilution examples are nearly all below 16 fold with the majority of examples below 2-fold.

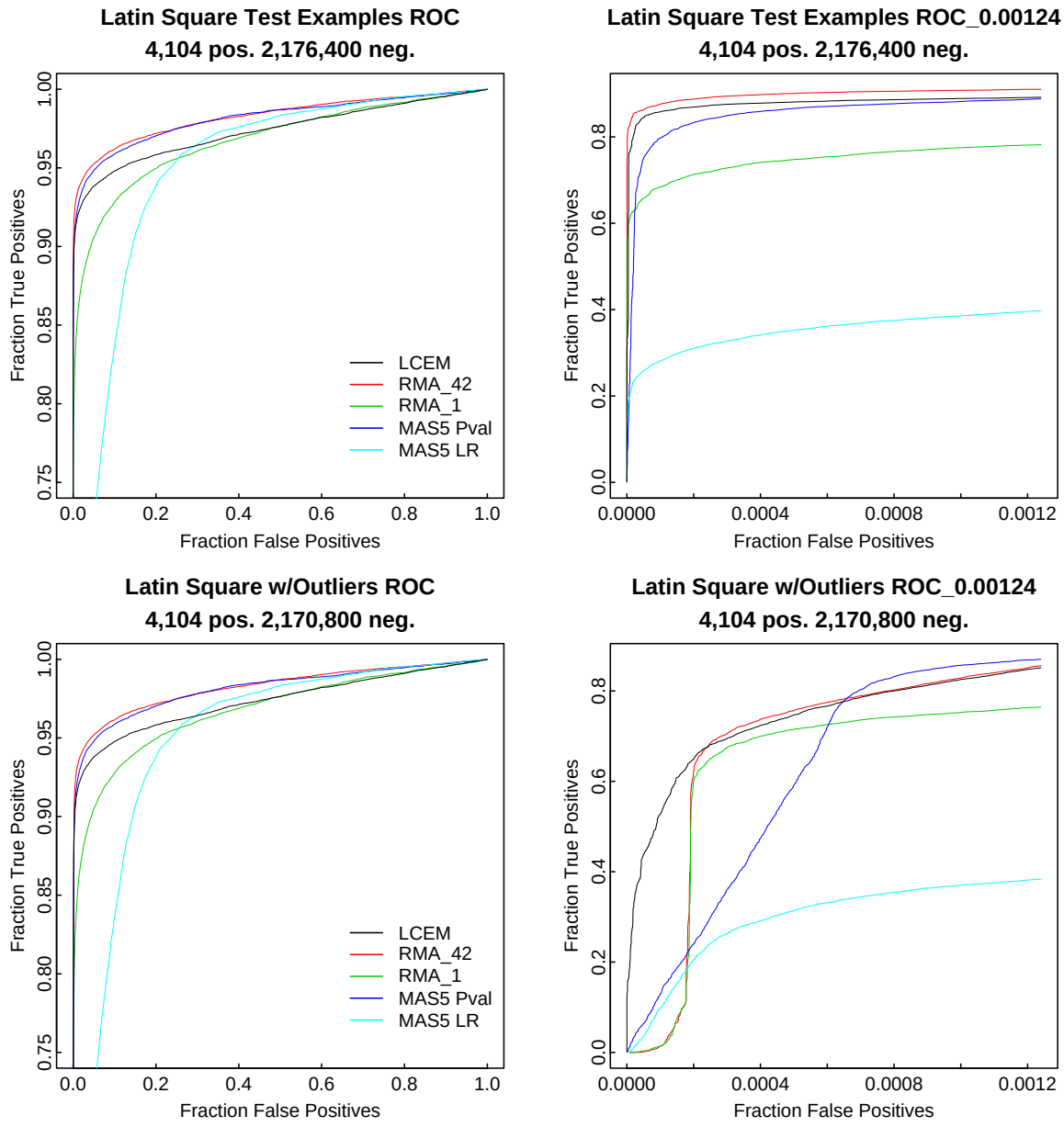

Fig. 4. Effect of outlier genes on latin square ROC curves Shown are ROC curves for the Affymetrix latin square examples with and without the 56 outlier genes, listed in Table 1. The outliers have a lower impact on LCEM than on the other four expression measure, suggesting that LCEM is robust to unusual data patterns.

Table 1. Affymetrix IDs of 56 outlier Ids. The 56 Affymetrix IDs listed were removed from the Affymetrix latin square dataset. While not among the 42 spiked in transcripts of the study, these IDs showed considereable expression differences with all three expression measures, MAS5, LCEM, and RMA. The effect of taking these IDs as negative examples is shown in the ROC curves of Figure 4.

200665_s_at	203173_s_at	203471_s_at	203508_at
204205_at	204417_at	204430_s_at	204513_s_at
204563_at	204836_at	204890_s_at	204891_s_at
204912_at	204951_at	204959_at	205267_at
205291_at	205398_s_at	205569_at	205692_s_at
205790_at	206060_s_at	207160_at	207540_s_at
207641_at	207655_s_at	207777_s_at	207968_s_at
208010_s_at	209354_at	209374_s_at	209606_at
209734_at	209795_at	212827_at	213060_s_at
AFFX-r2-Bs-dap-3_at	AFFX-r2-Bs-dap-5_at	AFFX-r2-Bs-dap-M_at	AFFX-r2-Bs-lys-3_at
AFFX-r2-Bs-lys-5_at	AFFX-r2-Bs-lys-M_at	AFFX-r2-Bs-phe-3_at	AFFX-r2-Bs-phe-5_at
AFFX-r2-Bs-phe-M_at	AFFX-r2-Bs-thr-3_s_at	AFFX-r2-Bs-thr-5_s_at	AFFX-r2-Bs-thr-M_s_at
AFFX-r2-TagA_at	AFFX-r2-TagB_at	AFFX-r2-TagC_at	AFFX-r2-TagD_at
AFFX-r2-TagE_at	AFFX-r2-TagF_at	AFFX-r2-TagG_at	AFFX-r2-TagH_at

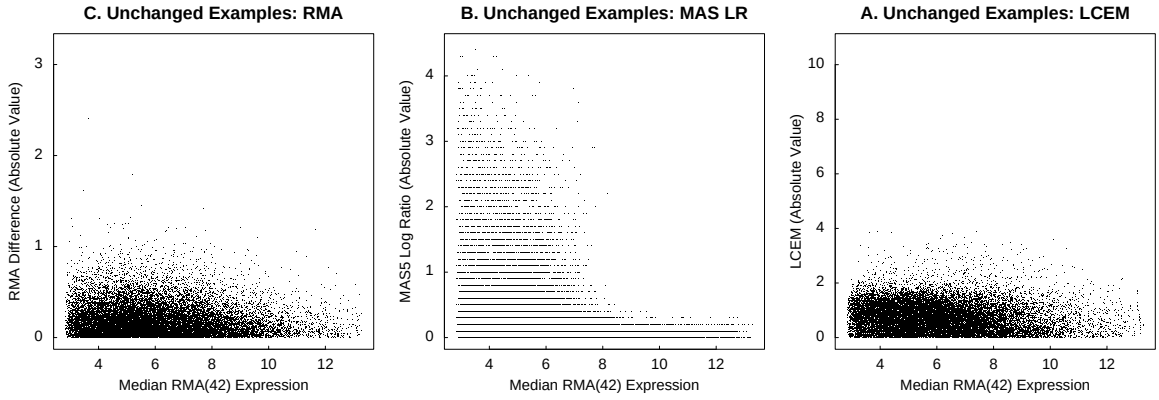

Fig. 5. Comparative expression versus average expression of unchanged examples. This plot shows comparative expression measures versus gene expression for unchanged genes in the Affymetrix latin square data set. Because the genes are unchanged, the LCEM, MAS5 LR and RMA Difference should ideally be 0. MAS5 shows a large bias towards high expression change measurements for genes expressed at low levels. RMA and LCEM do not show this bias.

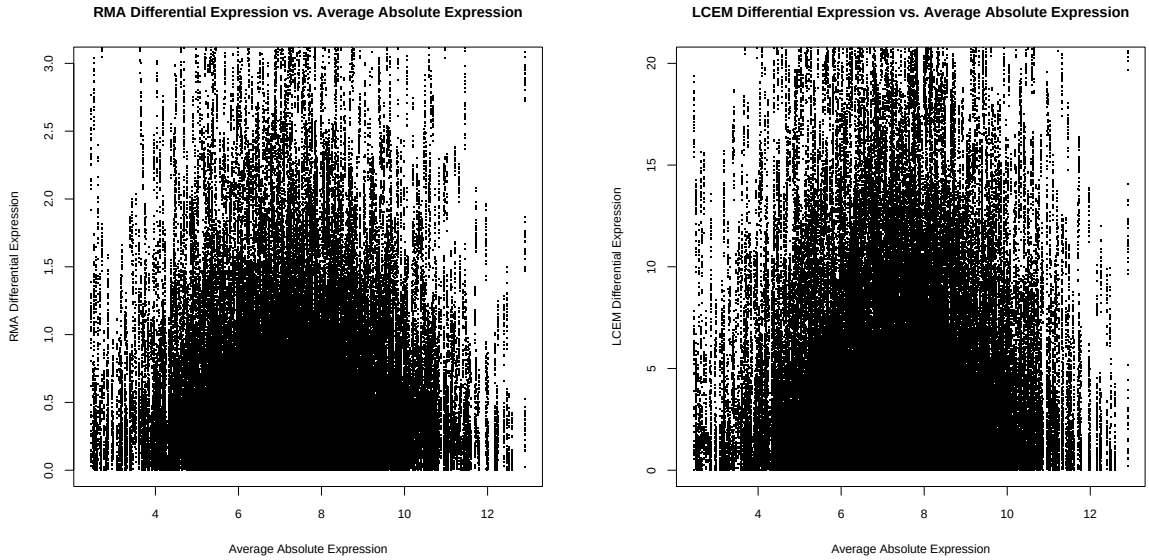

Fig. 6. Comparative expression versus average expression for RMA and LCEM on changed examples This figure shows that LCEM has no bias against low expressors as compared to RMA.

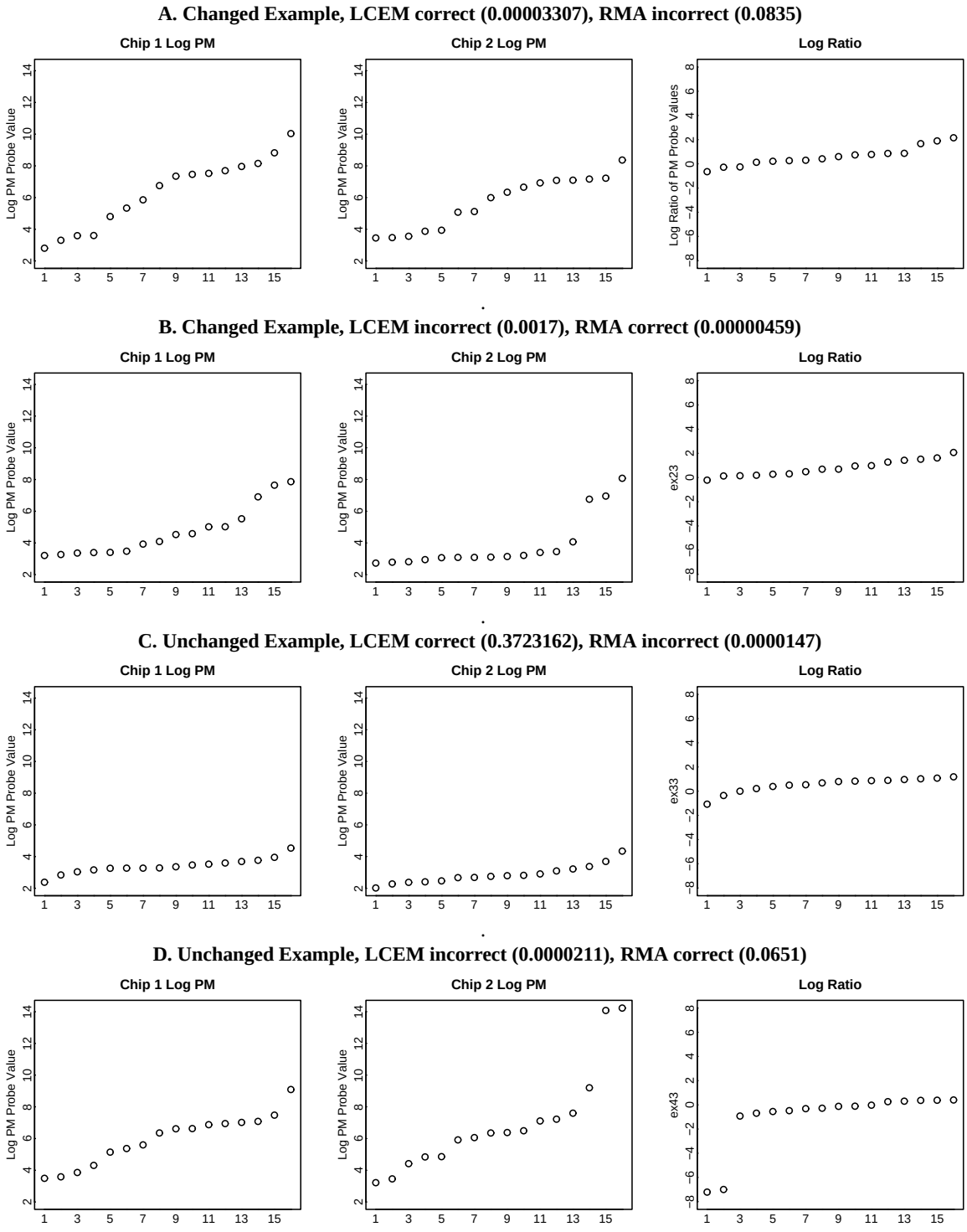

Fig. 7. Examples Shown are examples of raw data for the dilution data set in four cases for which RMA_{25} and LCEM disagree. The left and middle boxes show order statistics of perfect match log values for the two chips involved in the comparison. The right boxes show order statistics of log ratios between corresponding perfect match probes on the two chips. Quantities in parentheses are the fraction of examples which the respective method ranks above the given example. (A) In this case, LCEM correctly identified the example as having changed expression, while RMA places about 8% of all examples above it. (B) In this case, RMA correctly identifies the example as changed expression, while LCEM places 0.17 % of examples above it. (C) In this case, LCEM correctly classifies the example as unchanged, while RMA places the example above almost all the examples. (D) In this case, RMA correctly identifies the example as unchanged and LCEM places the example above nearly all the examples.

REFERENCES

O'Connell, B., A. F. Cheung, C. P. Simkevich, W. Tam, X. Ren, M. K. Mateyak, and J. M. Sedivy (2003). A large scale genetic analysis of c-Myc-regulated gene expression patterns. *J Biol Chem. 278*(14), 12563–73.