Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture

2008 - 20th Annual Conference Proceedings

STATISTICAL ISSUES IN THE NORMALIZATIONOF MULTI-SPECIES MICROARRAY DATA

John R. Stevens

Balasubramanian Ganesan

Prerak Desai

Sweta Rajan

Bart C. Weimer

See next page for additional authors

Recommended Citation

Stevens, John R.; Ganesan, Balasubramanian; Desai, Prerak; Rajan, Sweta; and Weimer, Bart C. (2008). "STATISTICAL ISSUES IN THE NORMALIZATIONOF MULTI-SPECIES MICROARRAY DATA," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1096

Author Information

John R. Stevens, Balasubramanian Ganesan, Prerak Desai, Sweta Rajan, and Bart C. Weimer

Statistical Issues in the Normalization of Multi-Species Microarray Data

John R. Stevens[1,3], Balasubramanian Ganesan[2,3], Prerak Desai[2,3],
Sweta Rajan[2,3], and Bart C. Weimer[2,3]

[1] Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill, Logan, UT 84322-3900 USA; [2] Department of Nutrition and Food Sciences, and Center for Microbe Detection and Physiology, Utah State University, 8700 Old Main Hill, Logan, UT 84322-8700 USA; [3] Center for Integrated BioSystems, Utah State University, 4700 Old Main Hill, Logan, UT 84322-4700 USA

Abstract

Several species of bacteria are involved in the production of cheese, including *Lactobacillus brevis* and *Lactococcus lactis*. A custom-designed Affymetrix microarray was recently developed to study gene expression in three organisms on a single chip. This array contains only perfect match features for the coding and non-coding regions in the genomes of all three sequences. The multi-species nature of this array version raises interesting questions regarding the preprocessing or normalization strategies for the analysis of gene expression data. We present and evaluate several possible strategies using both cDNA dilution data and experimental expression data from a repeated measures design. The statistical protocols highlighted in this work are applicable to other multi-species microarrays.

Keywords: microarray, normalization, multi-species, preprocessing

1 Introduction

During the production of cheese, a variety of microbial organisms contribute to specific attributes of the final product. For example, *Lactococcus lactis* is a bacteria used for fermentation and flavor in cheese. The SK11 strain of *L. lactis* was identified as a key organism to produce beneficial flavor compounds. Consequently, this genome was sequenced and studied using genomic and proteomic approaches to better understand the basis of SK11's effect on cheese flavor (Makarova et al. 2006). Another factor in cheese flavor is the enzyme glutamate dehydrogenase (GDH). The fact that both SK11 and GDH are involved in cheese flavor naturally leads to the biological question of how the gene expression profile of SK11 differs in the presence vs. the absence of GDH.

Microarray technology (Lockhart et al. 1996; Craig et al. 2003) can be used to monitor the expression values of thousands of genes simultaneously in a given biological sample, and the literature is rich with applications of this technology. This paper assumes that the reader is somewhat familiar with the technology, especially the GeneChip microarray from Affymetrix (`www.affymetrix.com`), with its associated vocabulary of multiple probes per gene (or probeset) on a single array. For the current application, a custom chip (LACTICREa520200F) was designed representing the following three bacterial genomes: strains SK11 and IL1403 of *Lactococcus lactis*, and *Lactobacillus brevis.* The design of the custom chip also included all known lactococcal bacteriophages (viruses that infect lactococcal bacteria) as well as all known unique lactococcal plasmids. Plasmids are extrachromosomal DNA molecules that can be shared between bacteria and code for important traits such as drug resistance and toxin production. In cheese production two of the most important traits are on plasmids – lactose use and protein degradation. The custom chip included only perfect match features, with no mismatch features on the array; in other words, each spot or feature on the array contains a short sequence of nucleotides that perfectly matches a known sequence from either one of the three species' genomes, a lactococcal bacteriophage, or a lactococcal plasmid. The number of probesets (per strain) on the LACTICREa520200F is summarized in Table 1.

Figure 1 represents the physical locations of the probesets for the three strains on the LACTICREa520200F array, with no obvious spatial pattern. Of the 4,448 SK11 probesets on the array, only 2,623 correspond to coding genes; the others correspond to intergenic DNA, sometimes called "junk" DNA. The noncoding regions of the SK11 genome are represented on the array because their involvement in gene regulation is of interest. However, only the 2,623 coding genes are expected to detect true signal (gene expression) when an SK11 sample is hybridized to the array. The physical locations of the probesets for these 2,623 SK11 coding genes are also represented in Figure 1, also revealing no obvious spatial pattern.

To address the biological question of how the gene expression profile of SK11 differs in the presence vs. the absence of GDH, a repeated measures experiment was conducted. Table 2 summarizes the design. Six batches of SK11 were cultured, with three in the absence (-) of the GDH enzyme, and three in the presence (+) of GDH. At each of three time points (0, 7, or 14 days), a cheese slurry was drawn from each batch and the mRNA was extracted and prepared for hybridization to an array. In all, 18 slurries were drawn, resulting in 18 arrays. The repeated measures model for each gene individually is

$$Y_{ijk} \quad = \quad \mu + B_{i(j)} + G_j + T_k + GT_{jk} + \epsilon_{ijk}, \tag{1}$$

where Y_{ijk} is the log-scale expression value of the gene for batch i under GDH status j at time k. The traditional assumptions are made for a two-factor experiment with repeated measures on one factor (Neter et al. 1996), namely that μ is a constant, the $B_{i(j)}$ are independent $N(0, \sigma^2_B)$, the G_j are constants subject to $\sum_j G_j = 0$, the T_k are constants subject to $\sum_k T_k = 0$, the GT_{jk} are constants subject to $\sum_j GT_{jk} = 0$ for all k and $\sum_k GT_{jk} = 0$ for all j, the ϵ_{ijk} are independent $N(0, \sigma^2_\epsilon)$, and the $B_{i(j)}$ and ϵ_{ijk} are independent. In this model, batch is random, while GDH status and time are fixed. The test for the GDH main effect (Neter et al. 1996) is

$$H_0 \quad : \quad G_1 = G_2. \tag{2}$$

This test focuses on identifying genes that exhibit differential expression between the GDH- and GDH+ conditions, averaging over all time points. (Other tests are possible, including tests to identify genes exhibiting interaction between GDH condition and time.) In order to fit this model, the log-scale expression value Y_{ijk} must first be estimated.

2 Methods

The methods of data analysis to take probe-level intensities on a group of arrays and arrive at expression estimates for each gene on each array are generally referred to as preprocessing (Gentleman et al. 2005). Preprocessing methods typically involve three components: background correction, normalization, and summarization The purpose of background correction is to remove noise and small local artifacts on each array. The purpose of normalization is to standardize intensities on different arrays so that comparisons across arrays are more interpretable. Summarization involves combining the intensities of a gene's probes within each array to arrive at a single expression estimate for the gene on each array. Because these steps prepare and standardize the raw intensity data for subsequent analysis (such as testing for differential expression), they are sometimes collectively referred to as normalization. A variety of preprocessing methods have been proposed, with MAS5 (Affymetrix 2001; Affymetrix 2002), dChip (Li and Wong 2001; Li and Wong 2003), RMA (Irizarry et al. 2003), GCRMA (Wu et al. 2004), and PLIER (Affymetrix 2005; Therneau and Ballman 2005) among the most commonly used. Of these, the RMA method draws attention for the current application for two reasons. First, the LACTICREa520200F array has no mismatch features, and RMA makes no use of mismatch intensities. Second, the RMA background correction model is well suited for the multi-species nature of this array.

2.1 RMA Preprocessing

The RMA method assumes that each perfect match (PM) intensity is the sum of background (bg) and signal (s). Background may result from cross-hybridization or another technical source of variation, while signal results from actual gene expression. With the array-specific assumptions

$$\begin{aligned} PM &= bg + s, \\ bg &\sim N(\mu, \sigma^2), \text{ and} \\ s &\sim Exp(\alpha), \end{aligned} \tag{3}$$

the RMA background convolution model obtains a closed-form transformation

$$\begin{aligned} PM' &= E[s|PM] \\ &= f(PM, \hat{\mu}, \hat{\sigma}, \hat{\alpha}). \end{aligned} \tag{4}$$

This transformation (details in Bolstad 2004 and McGee and Chen 2006) is array-specific, and can be made once the estimates for μ, σ, and α are obtained. The default estimates from the RMA method are

$$\hat{\mu} = \text{mode of (PM less than the array mode),}$$

$$\begin{aligned}\hat{\sigma} &= \sqrt{2} \times [\text{SD of (PM less than the array mode)}], \text{ and} \\ \hat{\alpha} &= 1/[\text{mode of (PM greater than the array mode)}]. \end{aligned} \tag{5}$$

These ad-hoc estimates are used because maximum likelihood or EM algorithm approaches are unstable and slow for this type of data (Bolstad 2004; McGee and Chen 2006). The resulting gene expression summaries have been shown to perform well in practice (Irizarry et al. 2003; Bolstad 2004), although they can be improved (McGee and Chen 2006). After estimating these parameters on each array as in Equation 5, the RMA background-correcting transformation in Equation 4 is performed on each array. Then the background-corrected intensities are quantile-normalized across all arrays in the experiment (Irizarry et al. 2003).

Figure 2 demonstrates the effect of RMA's background correction and quantile normalization steps, using data from the 18-slurry SK11 experiment. RMA background correction essentially stretches the distribution of intensities toward zero, and forces or exaggerates bimodality. Quantile normalization forces a similar distribution of intensities on all arrays.

After RMA preprocessing, the distribution of expression values is slightly bimodal on the LACTICREa52002F array. This bimodality has a biological origin and is not an artifact of RMA preprocessing. Of the 10,751 probesets on the array (recall Table 1), only 2,623 correspond to SK11 coding genes and so are capable of detecting true signal (expression). Figure 3 illustrates how these 2,623 signal-possible genes' intensities are distinctly bimodal. This suggests that some of the SK11 signal-possible features on the array are detecting true signal (the upper peak) while others are detecting only background (the lower peak), since the distribution of the 8,128 features other than the 2,623 coincides with the lower peak of the signal-possible features' distribution.

2.2 RMA-MS Preprocessing

As noted at the beginning of Section 2, the RMA background correction model is well suited for the multi-species nature of this LACTICREa520200F array. Specifically, the multi-species nature of this array contributes to the bimodal pattern noted in Figure 3. This pattern can be exploited to make more informed estimates of the background and signal parameters in Equation 3 than the default ad-hoc estimates in Equation 5. The background parameters μ and σ can be estimated using only the "background-only" perfect match intensities (the 8,128 non-SK11-coding genes in this case), and the signal parameter α can be estimated using only the "signal-possible" perfect match intensities (the 2,623 SK11 coding genes in this case), as follows:

$$\begin{aligned}\hat{\mu} &= \text{median of background-only PM} \\ \hat{\sigma} &= \text{median absolute deviation of background-only PM} \\ \hat{\alpha} &= 1/[\text{mean of (signal-possible PM greater than } \hat{\mu})] \end{aligned} \tag{6}$$

As in the traditional RMA method, after estimating these parameters on each array, the background-correcting transformation in Equation 4 is performed on each array. Then the background-corrected intensities of the signal-possible features only are quantile-normalized across all arrays in the experiment; the background-only features are ignored at quantile normalization. Because this proposed preprocessing method is based on the RMA model and differs only in how the multi-species nature of the array affects estimation of model parameters and which features are quantile-normalized, it is called the RMA-MS method ("MS" for "multi-species").

The bimodal distribution of log-scale perfect match intensities is observed in other multi-species experiments. Figure 4 represents the distributions of intensities for three arrays from experiments other than the 18-slurry SK11 experiment. In one experiment using the LACTICREa520200F array, a single sample of SK11 cDNA was diluted at different concentrations, and each array in this experiment exhibited the same pattern suggesting an abundance of background-only features, with a mixture of signal-possible. Using a second custom array (STYLMONOa520430F) with multiple microbial species, two other experiments hybridizing Listeria and Salmonella cDNA to the array (designed by the Center for Integrated BioSystems at Utah State University) demonstrated clear bimodality, again suggesting a mixture of signal-possible and background-only features on multi-species arrays. Such clear bimodality is not seen in single-species arrays, and appears to arise in multi-species arrays for biological reasons – the presence of both background-only and signal-possible features.

2.3 Preprocessing Strategies

The following four preprocessing strategies are applied to the 18-slurry SK11 experiment data, as well as to the SK11 cDNA dilution experiment:

1. Full: RMA background-correction followed by quantile-normalization of all features (traditional RMA)

2. MS: RMA-MS background-correction followed by quantile-normalization of only signal-possible features

3. MS2: RMA-MS background-correction followed by quantile-normalization of all features

4. SP: RMA background-correction using only signal possible features, followed by quantile-normalization of only signal-possible features (traditional RMA using only signal-possible features; "SP" for "signal-possible")

In every strategy, after background-correction and quantile-normalization, the perfect match intensities for each gene on each array are summarized on the log2 scale using the median polish (Tukey 1977; Irizarry et al. 2003).

3 Results

Figure 5 illustrates the comparison of RMA and RMA-MS background-correction. The upper (signal) peaks coincide, while after RMA-MS background-correction the lower (background) peak is not stretched as far towards zero as is the background peak after RMA background-correction. RMA background-correction appears to produce a greater separation between background and signal peaks. However, RMA background-correction also appears to force signal where there should be none, as evidenced by the slight bimodality in the background-only distribution.

Figure 6 illustrates the distributions of gene expression estimates of the 2,623 signal-possible SK11 genes, in all 18 slurries, after each of the four preprocessing strategies. The two strategies that used RMA background-correction (Full, SP) produced noticeably more variability between background and signal peaks. The two strategies that used only the signal-possible features for quantile normalization (MS, SP) produced slightly less variability within peaks.

Using the gene expression estimates from each of the four preprocessing strategies on the 18-slurry data, the repeated measures model (Equation 1) was fit for each of the SK11 signal-possible genes. The test of differential expression was performed (corresponding to the null hypothesis in Equation 2), producing 2,623 p-values. To adjust for multiple comparisons and to control the false discovery rate, these p-values were converted to q-values (Storey 2003). Figure 7 summarizes the results. There is general agreement between the results of the Full and MS2 strategies, which both used all features for quantile normalization, and which both failed to find any gene close to significant at $q = 0.05$. There is also general agreement between the results of the MS and SP strategies, which both used only signal-possible features for quantile normalization. The MS strategy systematically produced slightly smaller q-values than the SP strategy.

In the SK11 cDNA dilution experiment (Figure 4), a single sample of SK11 cDNA was diluted at three concentrations: 250 ng, 500 ng, and 1000 ng, with two replicate arrays at each concentration. Each of the four preprocessing strategies was applied to these data, with quantile normalization only done within concentration level to preserve systematic shifts in intensities between levels (Figure 8). Again (Figures 8a-c) the same pattern noted in Figure 6 is observed, with greater variability between peaks for the two strategies that used RMA background-correction (Full, SP). Figures 8d-f show the distribution of log fold-changes between different concentration levels. All four strategies appear to underestimate the true log fold change. The log fold-changes from the MS and MS2 strategies have a distinct mode at 0, suggesting that they identify genes that were not present in the initial SK11 sample, while the Full and SP strategies suggest that nearly every gene was present.

The 575 genes whose log fold-changes fall in the lower peak of Figures 8d-f for the MS and MS2 strategies can be characterized. Over 90% of those genes were either bacteriophages or transposases (genes that jump locations between chromosomes, between locations in the same chromosome, and sometimes duplicate in the same chromosome). The raw intensities of these genes in both the SK11 dilution and 18-slurry experiments have the same distribution as the background-only genes in those experiments (Figure 9). This suggests that although these 575 genes were signal-possible (i.e., coding) genes, they were apparently not expressed in either the SK11 dilution or the 18-slurry experiments. The two preprocessing strategies that used RMA background-correction (Full, SP) seemed to force signal differences (Figure 8d-f) for these 575 genes where there were none.

4 Summary

As gene expression technology is applied to more (and smaller) organisms, multi-species microarrays will become more common, because researchers can put the entire genomes of multiple species on a single chip. When the genomes of multiple species are present on a chip, a subset of the features on the array can be identified as signal-possible when they are known to be coding genes for the species for which the gene expression study is conducted. The RMA-MS (or MS for "multi-species") preprocessing strategy is based on the RMA method, but uses RMA-MS background-correction followed by quantile-normalization of only signal-possible features, and finally median polish summarization.

With multi-species microarrays, the distribution of intensities (and resulting expression summaries) is typically bimodal, with background and signal peaks. The RMA-MS method produces gene expression summaries with reduced variability within the signal peak, compared to the traditional RMA method. This reduced variability within peak allows easier detection of significant differential expression, as shown by Figure 7.

A subset of 575 SK11 coding genes were shown to be not expressed in either the 18-slurry or the SK11 dilution experiment (Figure 9). Despite this lack of expression, these genes exhibited non-zero log fold-changes in the SK11 dilution experiment when preprocessing used traditional RMA background-correction (Full and SP in Figure 8). However, when preprocessing used RMA-MS background correction, these non-expressed genes exhibited fold-changes close to their expected log fold-change of zero (MS and MS2 in Figure 8).

The RMA-MS approach makes use of features on a multi-species microarray to perform background correction in a biologically meaningful manner. It uses background-only features to estimate background parameters, and signal-possible features to estimate signal parameters. At the quantile normalization stage, the RMA-MS approach restricts attention to signal-possible features, because they are the only features of biological interest (after background has been accounted for). This results in less variability within signal peak, facilitating significance in tests for differential expression. The RMA-MS approach has been implemented in R code (R Development Core Team 2007) with syntax familiar to Bioconductor users (Gentleman et al. 2005). All necessary code appears in the Appendix of this paper.

References

Affymetrix (2001). *Affymetrix Microarray Suite User's Guide Version 5.0.* Affymetrix, Santa Clara, CA.

Affymetrix (2002). *Statistical Algorithms Description Document.* Affymetrix, Santa Clara, CA. www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.

Affymetrix (2005). *Technical Note: Guide to Probe Logarithmic Intensity (PLIER) Estimation.* Affymetrix, Santa Clara, CA. www.affymetrix.com/support/technical/technotes/plier_technote.pdf.

Bolstad, B. M. (2004). *Low-Level Analysis of High-Density Oligonucleotide Array Data: Background, Normalization, and Summarization*. Ph. D. thesis, University of California, Berkeley, California.

Craig, B. A., M. A. Black, and R. W. Doerge (2003). Gene expression data: The technology and statistical analysis. *Journal of Agricultural, Biological, and Environmental Statistics 8*(1).

Gentleman, R., V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, NY.

Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research 31*(4), e14.

Li, C. and W. H. Wong (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science 98*(1), 31–36.

Li, C. and W. H. Wong (2003). DNA-chip analyzer (dChip). In G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger (Eds.), *The Analysis of Gene Expression Data: Methods and Software*, Chapter 5. Springer, NY.

Lockhart, D. J., H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology 14*.

Makarova, K., A. Slesarev, Y. Wolf, A. Sorokin, B. Mirkin, E. Koonin, A. Pavlov, N. Pavlova, V. Karamychev, N. Polouchine, V. Shakhova, I. Grigoriev, Y. Lou, D. Rohksar, S. Lucas, K. Huang, D. M. Goodstein, T. Hawkins, V. Plengvidhya, D. Welker, J. Hughes, Y. Goh, A. Benson, K. Baldwin, J. H. Lee, I. Diaz-Muniz, B. Dosti, V. Smeianov, W. Wechter, R. Barabote, G. Lorca, E. Altermann, R. Barrangou, B. Ganesan, Y. Xie, H. Rawsthorne, D. Tamir, C. Parker, F. Breidt, J. Broadbent, R. Hutkins, D. O'Sullivan, J. Steele, G. Unlu, M. Saier, T. Klaenhammer, P. Richardson, S. Kozyavkin, B. C. Weimer, and D. A. Mills (2006). Comparative genomics of the lactic acid bacteria. *Proceedings of the National Academy of Sciences 103*(42), 15611–15616.

McGee, M. and Z. Chen (2006). Parameter estimation for the exponential-normal convolution model for background-correction of affymetrix genechip data. *Statistical Applications in Genetics and Molecular Biology 5*(1), 24.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman (1996). *Applied Linear Statistical Models*. McGraw-Hill, Boston, MA.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of Statistics 31*, 2013–2035.

Therneau, T. M. and K. V. Ballman (2005). *What does PLIER really do?* Technical Report Series No. 75, Department of Health Science Research, Mayo Clinic, Rochester, Minnesota. `http://cancercenter.mayo.edu/mayo/research/biostat/upload/75.pdf`.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Wu, Z., R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association 99*(468), 909–919.

Table 1: Numbers of probesets per bacterial strain on the LACTICREa520200F custom array.

Strain	Probesets
IL1403	3,399
L. brevis	2,342
SK11	4,448
L. lactis	460
other	102
	10,751

Table 2: Repeated measures design used for the gene expression study. Eighteen slurries were prepared and hybridized to arrays.

		GDH Status					
		-	-	-	+	+	+
Time:	0	S1	S2	S3	S10	S11	S12
(Days)	7	S4	S5	S6	S13	S14	S15
	14	S7	S8	S9	S16	S17	S18
Batch:		1	2	3	4	5	6

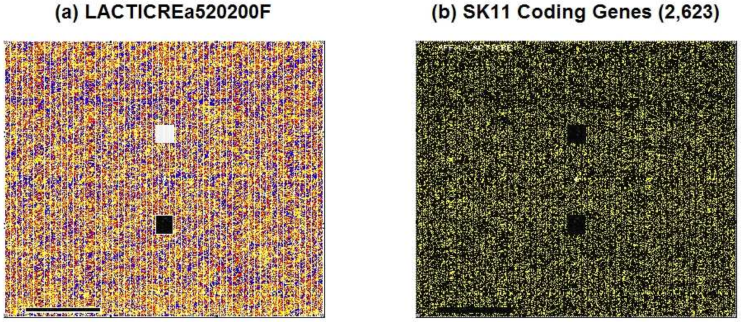

Figure 1: (a) Physical locations of probesets on the LACTICREa520200F custom array, colored by bacterial strain. IL1403 features are red, L. brevis features are blue, and SK11 features are yellow. (b) Physical locations of probesets corresponding to only the 2,623 SK11 coding genes are shown, in yellow. In both images, control and image-alignment features are visible as black, white, or grey squares.

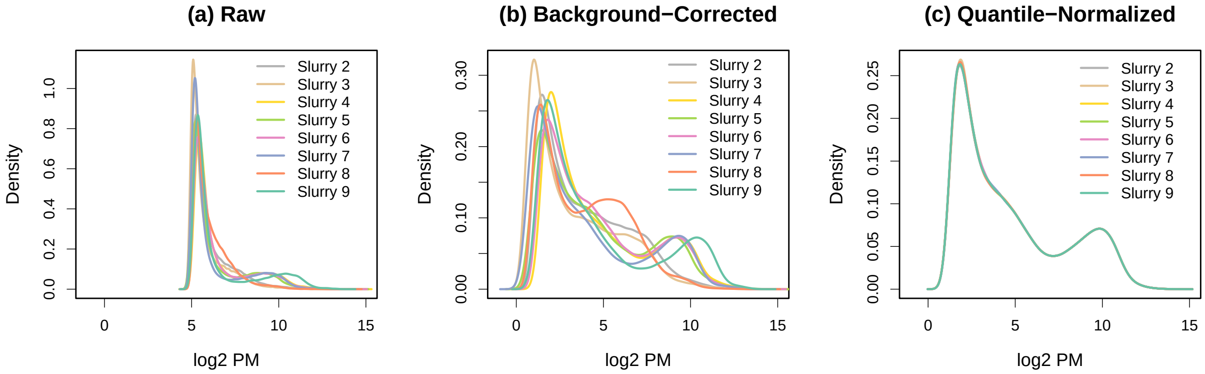

Figure 2: (a) Smoothed histograms of log2-scale perfect match intensities, without any preprocessing. (b) Smoothed histograms after RMA background-correction. (c) Smoothed histograms after RMA background-correction and then quantile normalization. All eighteen arrays (slurries) were preprocessed using RMA, but only eight are shown here for ease of visualization.

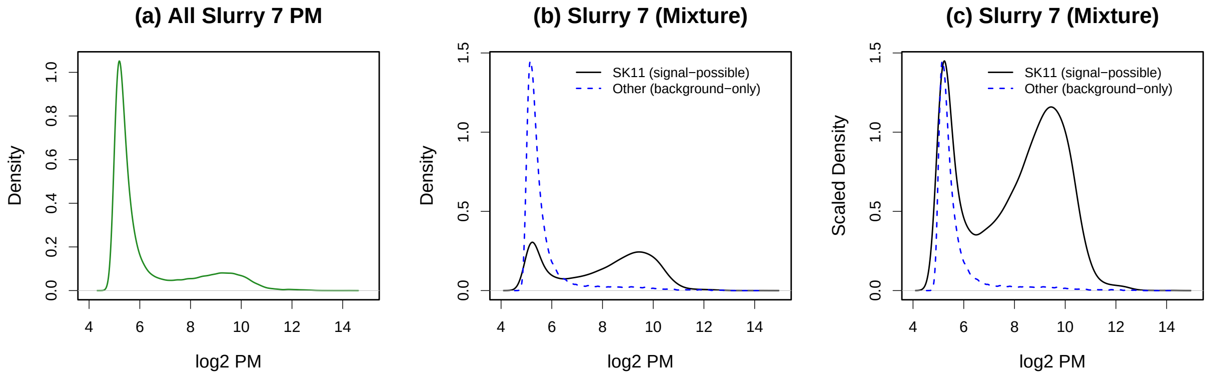

Figure 3: (a) Smoothed histograms of log2-scale perfect match intensities for one of the eighteen slurries, before any preprocessing. (b) The bimodal pattern in (a) can be explained as a mixture of SK11 (signal-possible) features and other (background-only) features. (c) The signal-possible distribution is emphasized by re-scaling the density of both distributions to have a shared maximum.

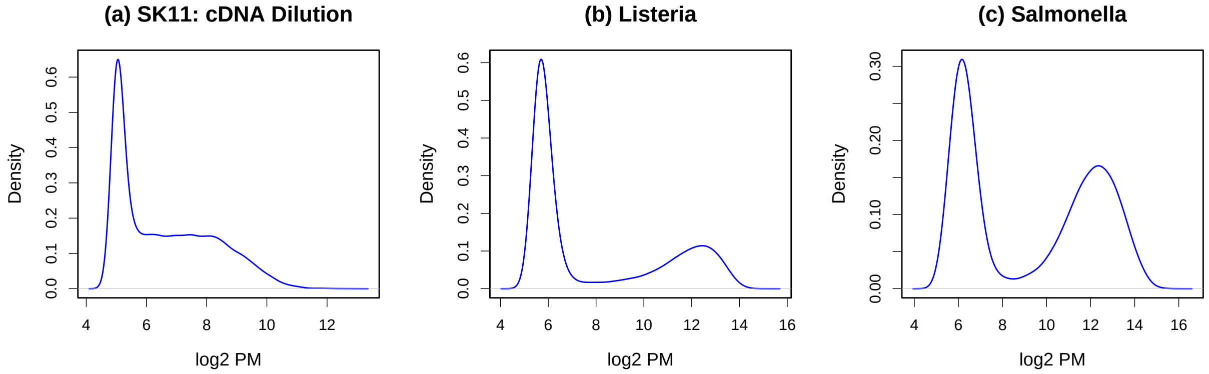

Figure 4: Smoothed histograms of log2-scale perfect match intensities for three arrays from experiments other than the 18-slurry SK11 experiment. Each exhibits bimodality, suggesting the mixture of signal-possible and background-only features.

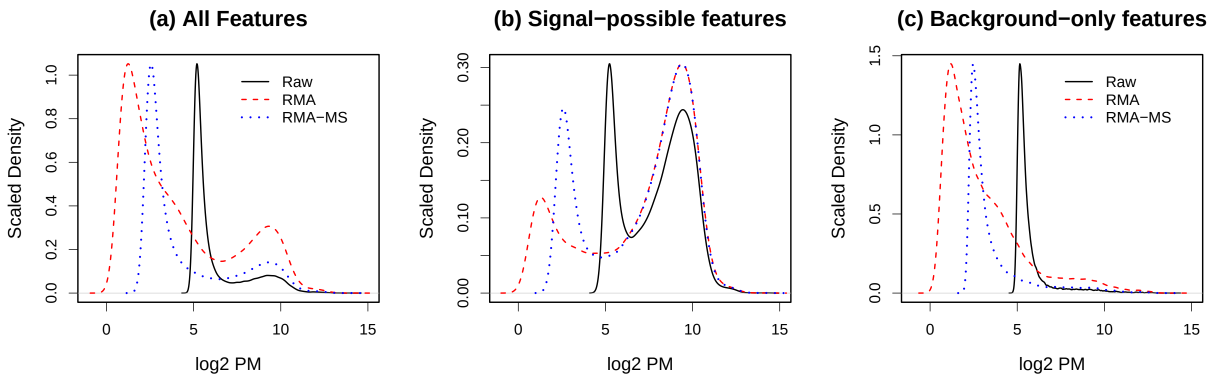

Figure 5: Smoothed histograms of log2-scale perfect match intensities for one of the eighteen slurries, after one of three background-correction methods: none ("Raw"), RMA, and RMA-MS. (a) All 10,751 features on the array are represented. (b) Only the 2,623 signal-possible features are represented. (c) Only the 8,128 background-only features are represented.

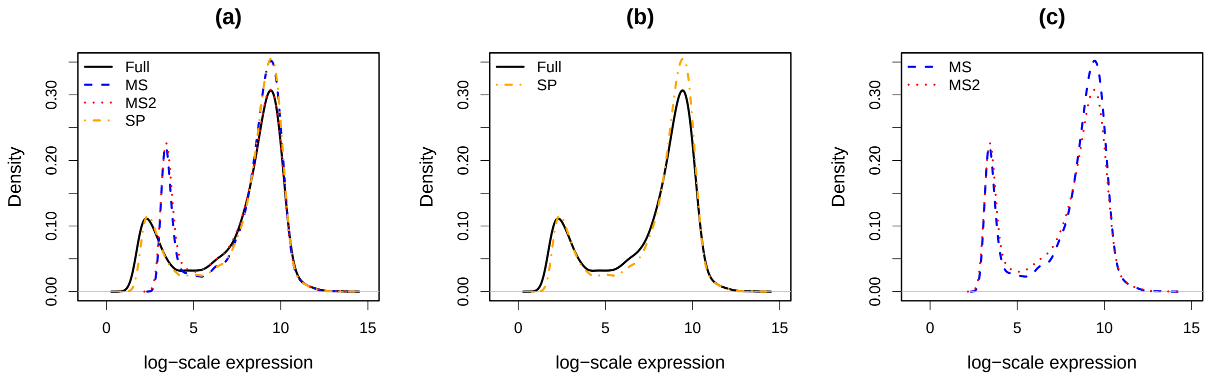

Figure 6: Smoothed histograms of final gene expression estimates of the 2,623 signal-possible genes, on all eighteen slurries, after one of four preprocessing strategies: Full, MS, MS2, and SP. (a) All four methods are simultaneously represented. (b) The two strategies that used RMA background-correction are represented. (c) The two strategies that used RMA-MS background-correction are represented.

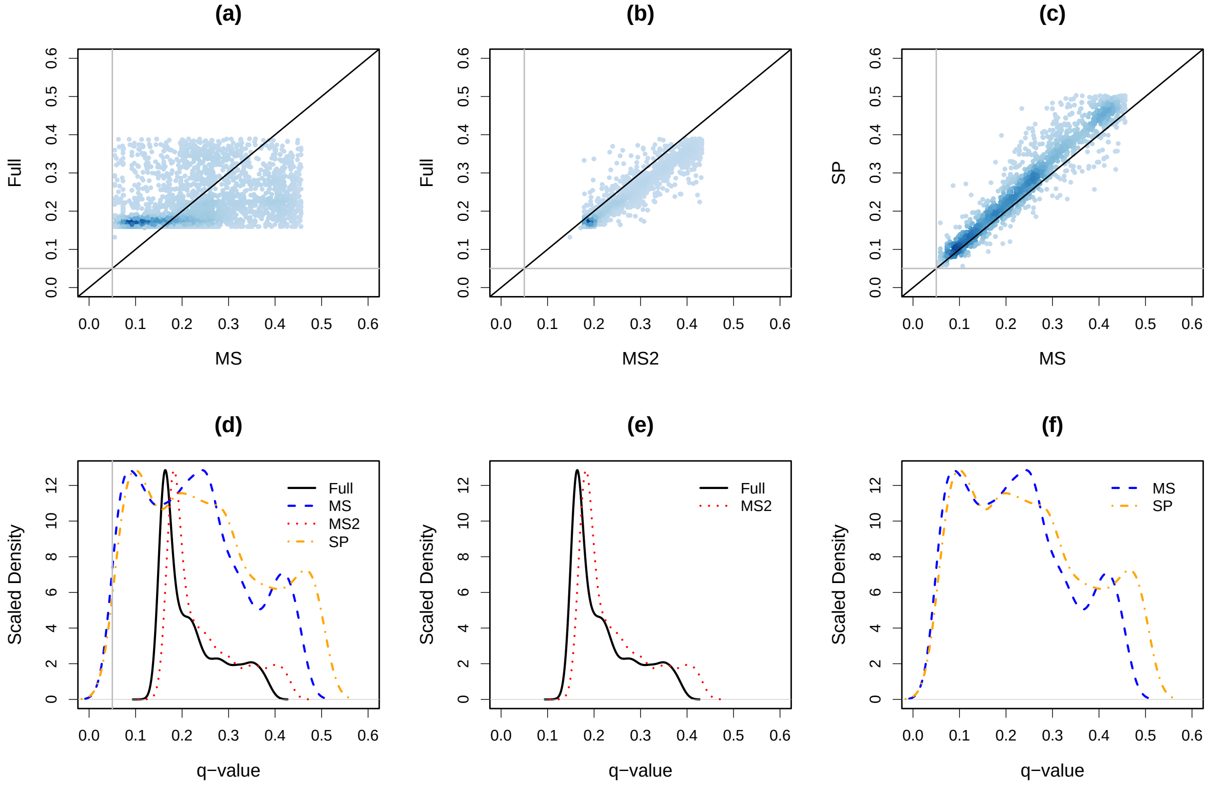

Figure 7: Plots comparing results of the test for differential expression based on four preprocessing strategies. (a-c) Scatterplots comparing q-values for the 2,623 signal-possible genes under different pairs of strategies, with darker color corresponding to greater density of points. (d-f) Smoothed histograms of q-values under different strategies, with vertical axis scaled so that all four densities share a maximum. In (d), all four strategies are represented. Figure (e) represents the two strategies that used all features for quantile normalization, while (f) represents the two strategies that used only signal-possible features for quantile normalization.

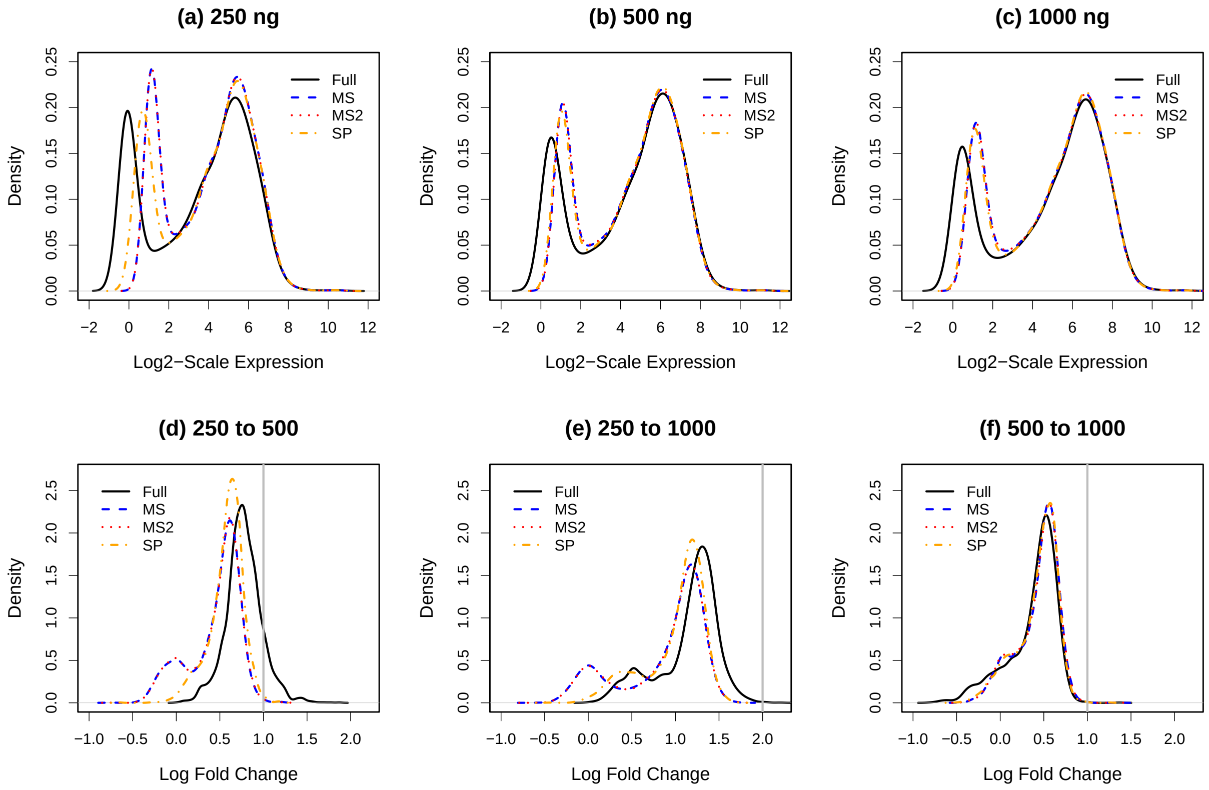

Figure 8: Summary of the four preprocessing strategies applied to the SK11 cDNA dilution experiment. Only the 2,623 SK11 coding genes are represented here. (a-c) Smoothed histograms of the final expression values by dilution concentration level, after each of the four strategies. (d-f) Smoothed histograms of log (base 2) fold changes between pairs of dilution concentration levels, after each of the four strategies. The vertical reference lines indicate the “true” log fold-change between concentration levels.

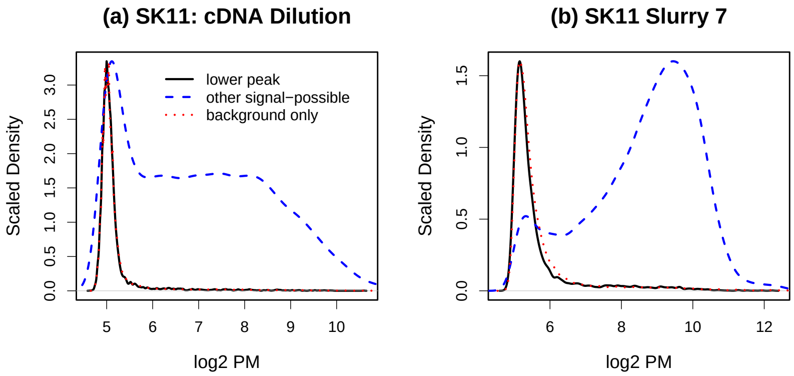

Figure 9: Distribution of log-scale perfect match intensities on two arrays from the (a) SK11 dilution and (b) 18-slurry experiments. In both figures, three distributions are overlaid. The background-only distribution represents the 8,128 non-SK11-coding genes, while the "lower peak" distribution represents the 575 SK11 coding genes that fell in the lower peak for both MS and MS2 in Figure 8d-f. The remaining 2,048 SK11 coding genes are represented by the "other signal-possible" distribution.

Appendix: R Code for RMA-MS

```
# Sample call:
#    library(affy)
#    data <- ReadAffy(...)  # AffyBatch object to be preprocessed
#    gn.sp <- ...  # vector of signal-possible geneNames (probeset ids)
#    source(...) # source in this code from a separate file
#    eset <- rma.ms(data,gn.sp)  # returns an ExpressionSet object

# RMA-MS: abatch is AffyBatch object, gn.sig is a vector of signal-possible geneNames
rma.ms <- function(abatch,gn.sig)
{ rma.abatch <- bg.correct(abatch, method="ms", gn.sig=gn.sig)
  rma.abatch.norm <- qnorm.subset(rma.abatch,gn.sig)
  out.eset <- expresso(rma.abatch.norm,bg.correct=FALSE,normalize=FALSE,
    pmcorrect.method="pmonly",summary.method="medianpolish",summary.subset=gn.sig)
  return(out.eset)
 }

# RMA-MS background correction; based on bg.correct.rma
bg.correct.ms <- function (object, gn.sig, ...)
{ pn <- probeNames(object);  t.sig <- is.element(pn,gn.sig)
  pm.mat <- pm(object);      pm.sig <- apply(pm.mat, 2, bg.adjust.ms, t.sig=t.sig)
  pm.mat <- pm.sig;          pm(object) <- pm.mat
  return(object)
 }
bg.adjust.ms <- function (pm, n.pts = 2^14, t.sig, ...)
{ param <- bg.parameters.ms(pm, n.pts, t.sig)
  b <- param$sigma;    pm <- pm - param$mu - param$alpha * b^2
  pm + b * ((1/sqrt(2 * pi)) * exp((-1/2) * ((pm/b)^2)))/pnorm(pm/b)
 }
bg.parameters.ms <- function (pm, n.pts = 2^14, t.sig)
{ pm.bg <- pm[!t.sig];       pm.sig <- pm[t.sig]
  mu.bg <- median(pm.bg);    sd.bg <- mad(pm.bg)
  alpha.sig <- mean(pm.sig[pm.sig>mu.bg])
  list(alpha = 1/alpha.sig, mu = mu.bg, sigma = sd.bg) # Note 1/alpha.sig
 }
bgcorrect.methods <- c(bgcorrect.methods, "ms")

# Quantile normalize a subset of genes; based on normalize.AffyBatch.quantiles
qnorm.subset <- function(abatch,subset)
{ pms <- unlist(pmindex(abatch,subset))
  noNA <- rowSums(is.na(intensity(abatch)[pms, , drop = FALSE])) == 0
  pms <- pms[noNA]
  intensity(abatch)[pms, ] <- normalize.quantiles(intensity(abatch)[pms, , drop = FALSE],
    copy = FALSE)
  return(abatch)
 }
```