FACTORS CONTRIBUTING TO VARIABILITY IN DNA MICROARRAY RESULTS: THE ABRF MICROARRAY RESEARCH GROUP 2002 STUDY

K. L. Knudtson[1], C. Griffin[2], A. I. Brooks[3], D. A. Iacobas[4], K. Johnson[5], G. Khitrov[6], K. Lilley[7], A. Massimi[4], A. Viale[8], W. Zhang[4], Y. Bao[9], G. Grills[10], H. Thaler[8], D. Peterson[3]

[1]University of Iowa, [2]University of California San Francisco, [3]University of Rochester Medical Center, [4]Albert Einstein College of Medicine, [5]Jackson Laboratories, [6]The Rockefeller University, [7]University of Cambridge, [8]Memorial Sloan-Kettering Cancer Institute, [9]University of Virginia School of Medicine, [10]Harvard Partners Genome Center

Abstract

Experimentalists utilizing DNA microarray technology must overcome many challenges to obtain results in which the variability of the data is due solely to biological activity. It is important to be aware of and account for potential sources of variability in the experimental design and results. The goal of this study is to identify non-biological factors that contribute to variation in microarray results. A retrospective study, using data collected by ABRF MARG members in their core labs, was conducted using both Affymetrix GeneChip and spotted microarray technologies. The GeneChip study analyzed the metrics and meta data from a total of over 800 murine U74A and human U95A expression arrays. The spotted microarray study analyzed the effects of slide fabrication, hybridization and scanner settings from over 100 slides in which the same reference RNA was used. The effect of tissue type, array lot, hybridization procedure and scanner settings on the variability of results was investigated for both types of microarray technologies. The results of these studies provide insight on potential sources of experimental error with microarray technologies and suggest experimental strategies to correct them.

Introduction

The Human Genome Project has served to catalyze the development of new tools used to manage, analyze, and interpret the enormous amount of sequence information that has been generated. DNA microarray technology has emerged as one of these more powerful tools to analyze the transcriptome. Generally, DNA microarrays utilize base pairing rules to specifically match hundreds and thousands of sample gene targets to an ordered arrangement of characterized or partially characterized DNA probes. The number of target molecules bound to each probe can then be detected which provides quantitative information on, as well as identity of, each gene target in the sample. Currently, two DNA microarray platforms are widely employed. Namely, the slide-based technologies developed by the laboratories of Patrick Brown and Ronald Davis at Stanford University and the GeneChip technology developed by Affymetrix, Inc.

The utilization of high throughput gene expression technologies has presented challenges in the area of experimental design and data analysis. Investigators employing DNA microarray technology must overcome many challenges to obtain results in which the variability of the data is due solely to the biological perturbation being studied. It is important to be aware of the potential sources of variability in microarray data and account for them in the experimental design and results. The goal of this study is to identify factors not associated with biological activity that may contribute to variation in microarray results.. The preliminary study presented herein represents an effort to determine the data variances related to hardware, system and processing issues. A retrospective study using data that were previously generated in the microarray core facilities of the members of the MARG is described.

Methods

Experimental Approach. **This preliminary study utilized previously generated microarray data by members of the MARG to assess the variability in cDNA and Affymetrix GeneChip microarray experiments, identify factors contributing to errors, and develop strategies and algorithms that can be used to minimize these errors. These meta and metric data, are used in the quality control and standardization of microarray analysis outputs. The analysis described herein utilizes these outputs as measures of variability.**

Data Collection for custom array studies. **Ten test chips were analyzed with same extracts to determine process variability and the effect of scanner PMT voltage on ratios. Five Human 9K chips and 8 Mouse 9K chips were analyzed for hybridization and biological reproducibility.**

Data Collection for Affymetrix studies**. Average difference values, which are an indication of transcript abundance were collected for β-actin, GAPDH, and spike controls (BioB, BioC, and CreX) for the Human U95A and Murine U74A expression GeneChip arrays. In addition, the values used to assess the "noise" of the system (RawQ, Background, and Scaling) were collected. Global scaling was applied with the target intensity set at 2500. Data from at least 5 arrays with the same lot number (per each sample type) were collected. A macro was written in Excel to collect and collate the data from 835 arrays.**

Statistical analysis of Affymetrix GeneChip Data**. Kruskal-Wallis tests were performed to compare the distributions of the 8 continuous response variables (ADall.BioB, ADall.BioC, ADall.CreX, Scaling.Factor, Background.Level, AD35.GAPDH, AD35.Actin, Raw.Q.Noise) by groups (Lab, Tissue, Lot number). The p-value for each test represents the probability of observing data at least this extreme under the null hypothesis that the distributions are identical for all groups; thus small p-values are evidence that at least one group has a different distribution than the others. Each test is followed by a table of the medians (that value such that half the data are smaller and half are larger) by group; under the null hypothesis all medians would be identical, so differences in medians can be used to some extent to identify the primary contributions to the significance of each test. Finally, density estimates of the log10 continuous variables are plotted by [binary] group for CellLine vs. Tissue; where these curves are high, there is a high concentration of data. Comparison of these curves can lend insight into the distribution of each variable. Although the Tissue and CellLine groups represent diverse biology they were collapsed for analysis due to statistical restraints.**

Results of the Affymetrix GeneChip Study

Results And Discussion Of The Affymetrix Study

In each of the 16 graphs shown on the left, the average difference (AD) values comparing samples obtained from cell lines (solid lines) and tissue (dashed lines) were compared. In each comparison, the effect of sample type was significant ($p<0.001$). However, for the housekeeping genes (Figs. 1 thru 4), the distribution of the 3/5 ratios appear to be similar for β-actin (Figs. 1 and 2) and more varied for GAPDH (Figs. 3 and 4). This may reflect the influence of varying oxygen levels, known to affect GAPDH expression, in the incubators used to grow the cell lines.

Interestingly, the AD values for the exogenous spike controls (Figs. 5 thru 10) were significant between cell lines and tissues. In theory, they should not be affected by sample type. This suggests that the spike controls may not be adequate to normalize data for comparative analysis. The distribution of the AD values for BioB and C (Figs. 5 thru 8), which are included to represent low and medium copy number values, appear to be similar. The AD value distribution for CreX, however, was bimodal and probably due to the difference in PMT settings. AD values for CreX (Figs. 9 and 10) , which represents a high copy number target, achieved saturation when scanned with the high PMT setting. This notion is supported by the results shown in Table 1 that illustrates that PMT setting (along with lab) accounted for the largest source of error in the analysis.

The data distributions for Background and Scaling Factor (Figs. 11, 12, 15 and 16) appeared to be bimodal, possibly reflecting the differences in the PMT settings used by the different laboratories. The results shown in Table 1 also suggest that PMT and Lab were the greatest contribution of error in thee analyses of Background and Scaling Factor. However, the distribution profiles look similar between tissue and cell line in many cases. The Raw Q noise values appeared similar between cell lines and tissue.

Table 1 illustrates that the greatest source of error for the analyses shown in the density plots (Figs. 1-16) was due to lab-to-lab variation. This suggests that the .CHP array data generated by different laboratories may not be used without further normalization in comparative analyses.

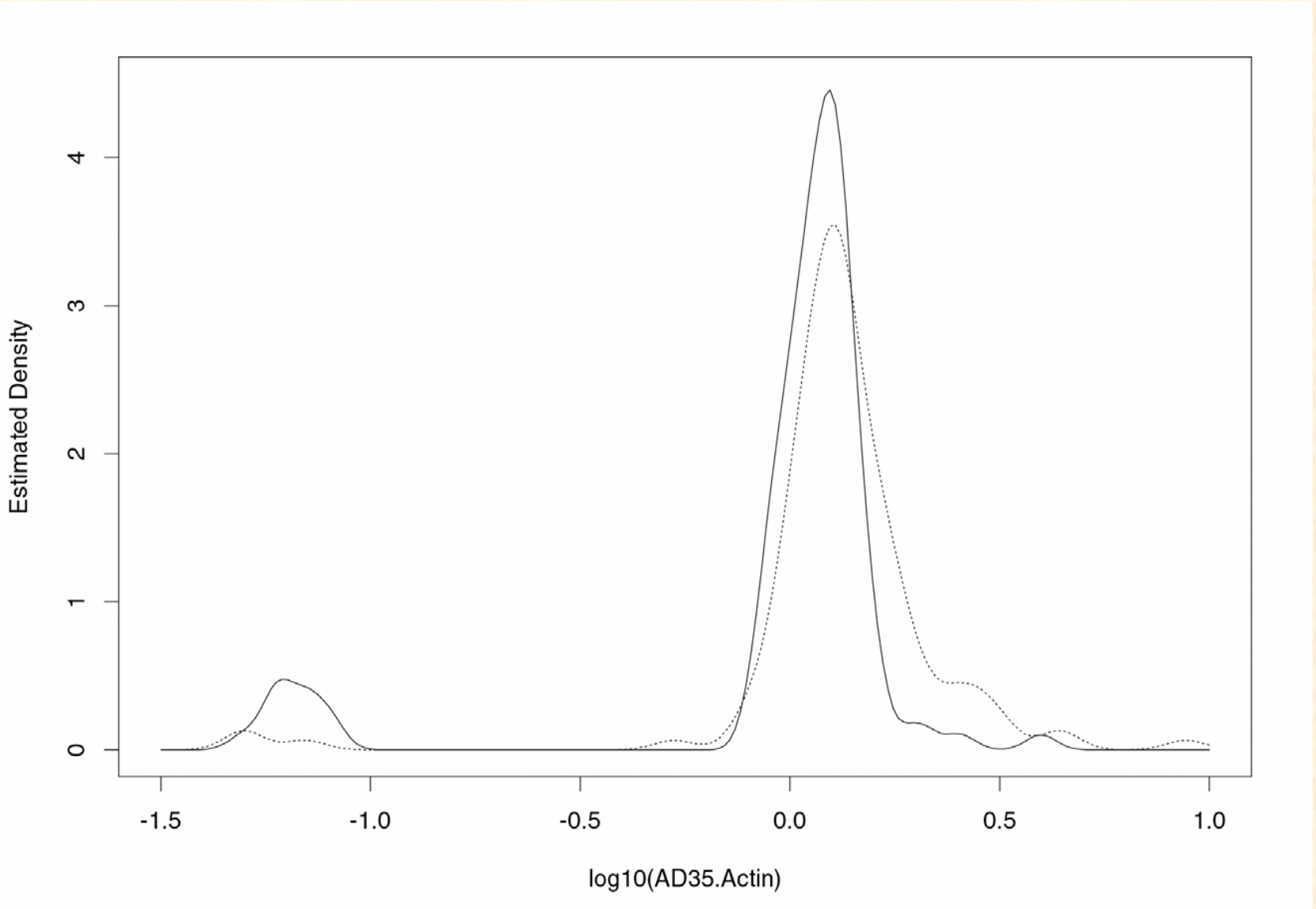

Figure 1. Density plots of the distribution of average difference values for the 3'/5' ratios of β-actin on Affymetrix Human U95A expression arrays. Cell lines = solid line and tissue = dashed line.

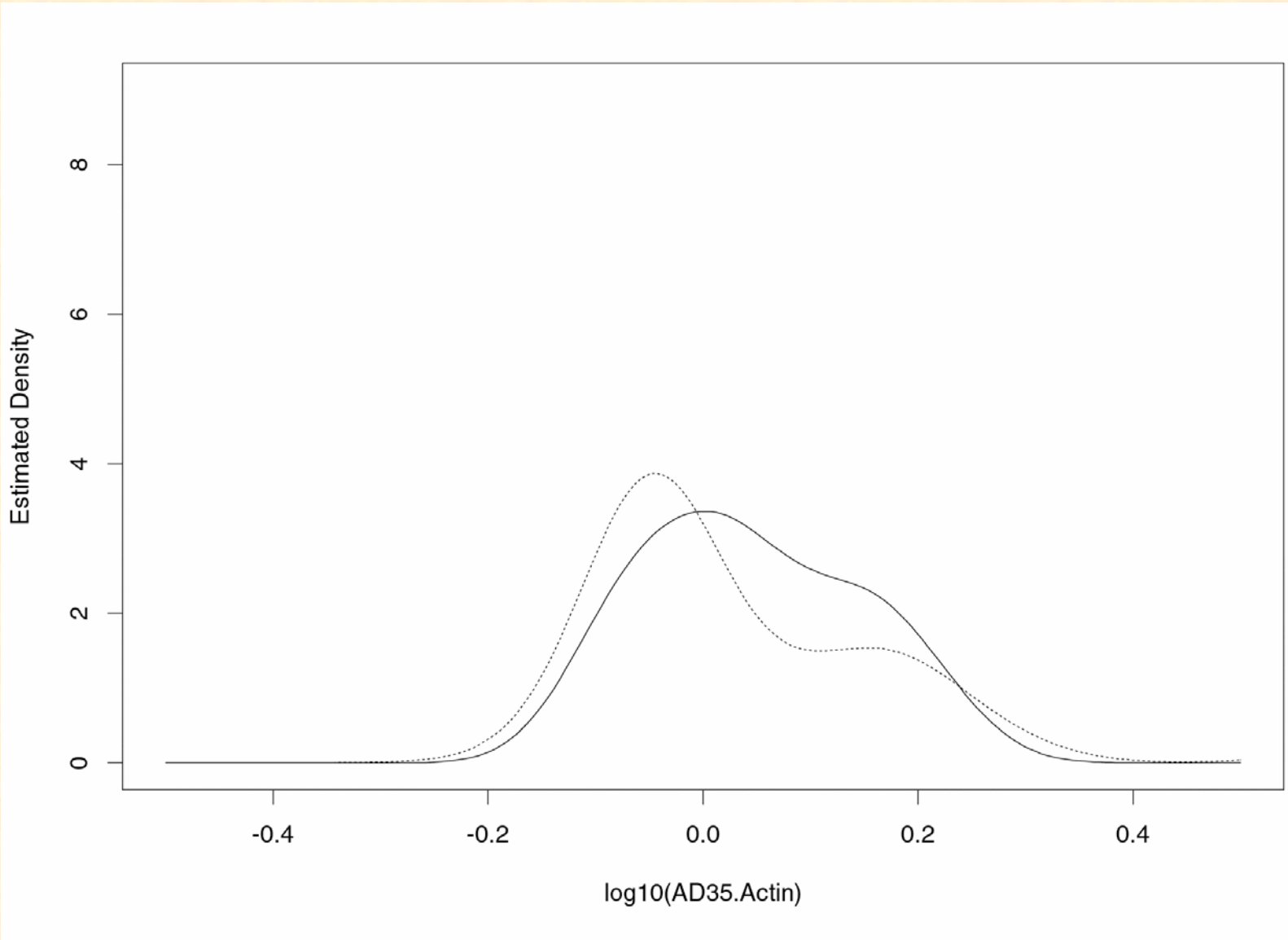

Figure 2. Density plots of the distribution of average difference values for the 3'/5' ratios of β-actin on Affymetrix Murine U74A expression arrays. Cell lines = solid line and tissue = dashed line.

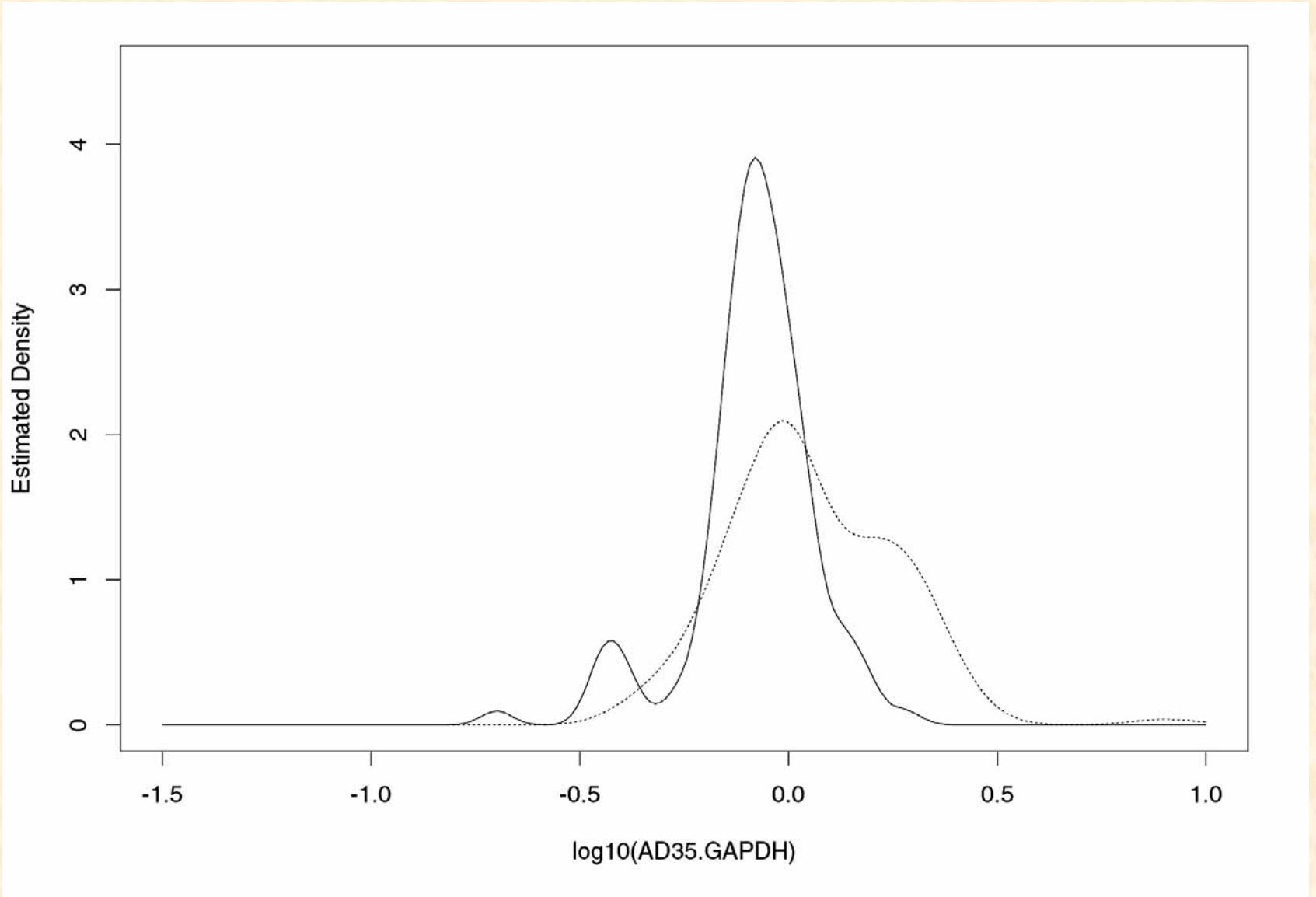

Figure 3. Density plots of the distribution of average difference values for the 3'/5' ratios of GAPDH on Affymetrix Human U95A expression arrays. Cell lines = solid line and tissue = dashed line.

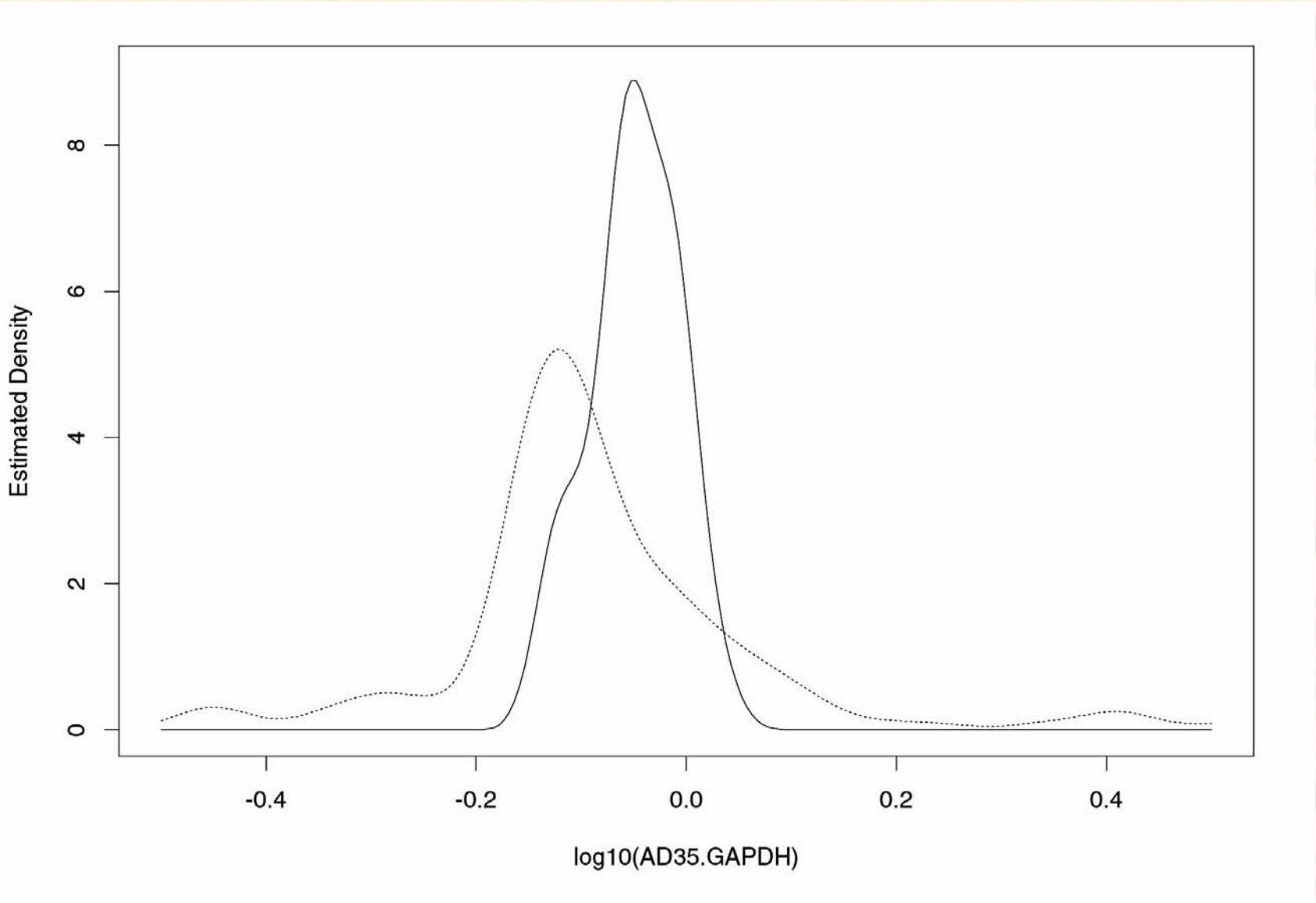

Figure 4. Density plots of the distribution of average difference values for the 3'/5' ratios of GAPDH on Affymetrix Murine U74A expression arrays. Cell lines = solid line and tissue = dashed line.

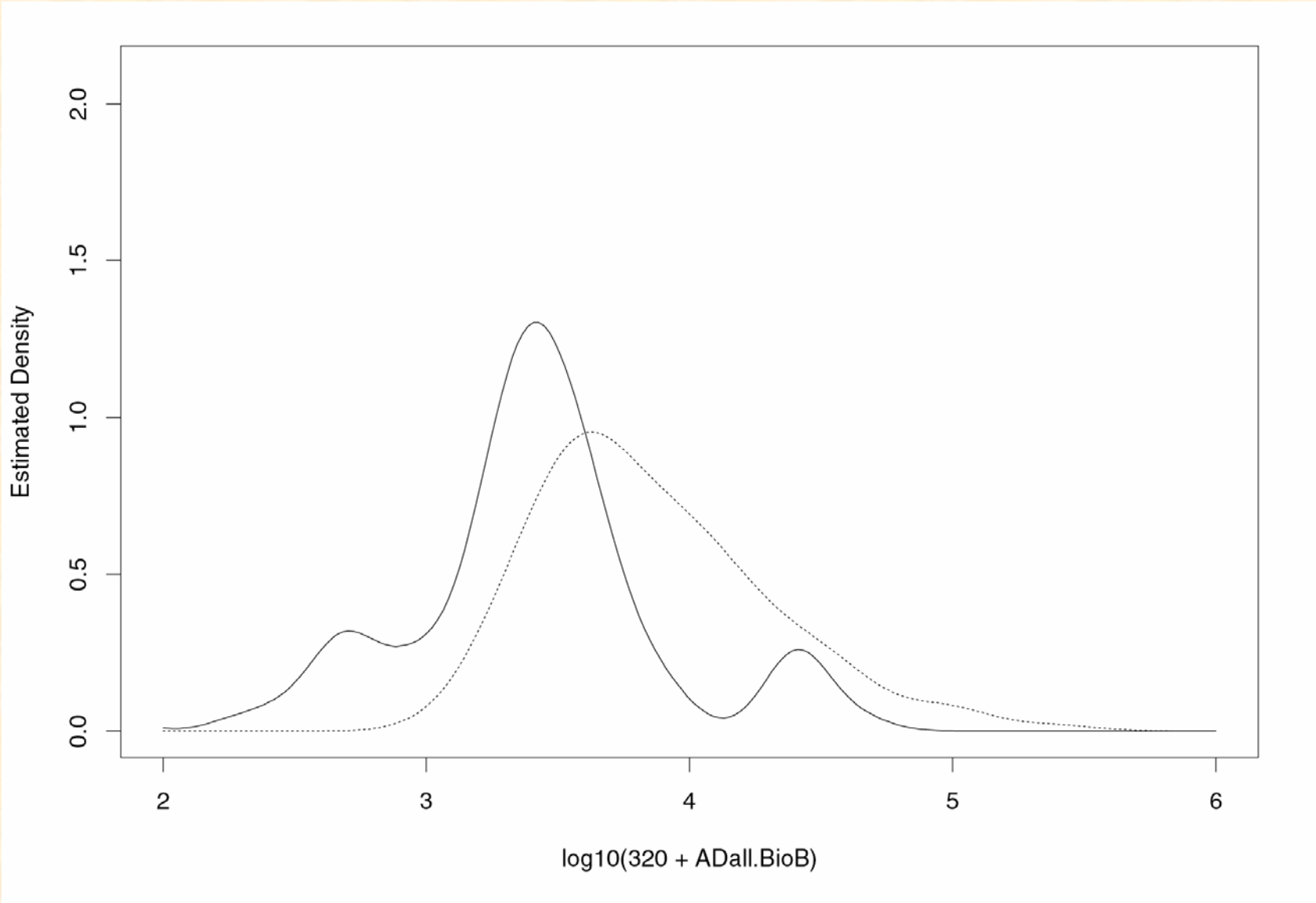

Figure 5. Density plots of the distribution of average difference values for the spike control BioB on Affymetrix Human U95A expression arrays. Cell lines = solid line and tissue = dashed line.

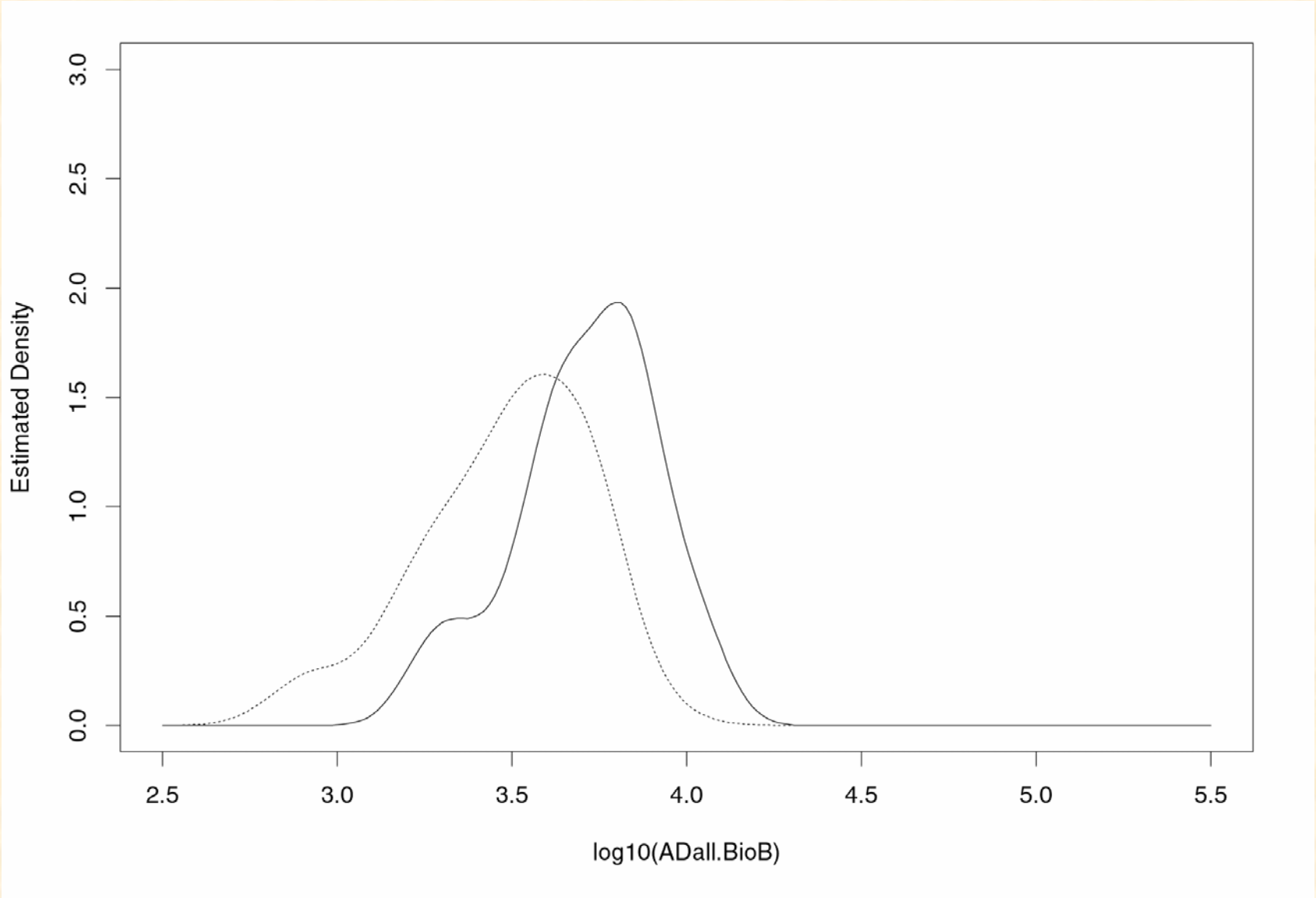

Figure 6. Density plots of the distribution of average difference values for the spike control BioB on Affymetrix Murine U74A expression arrays. Cell lines = solid line and tissue = dashed line.

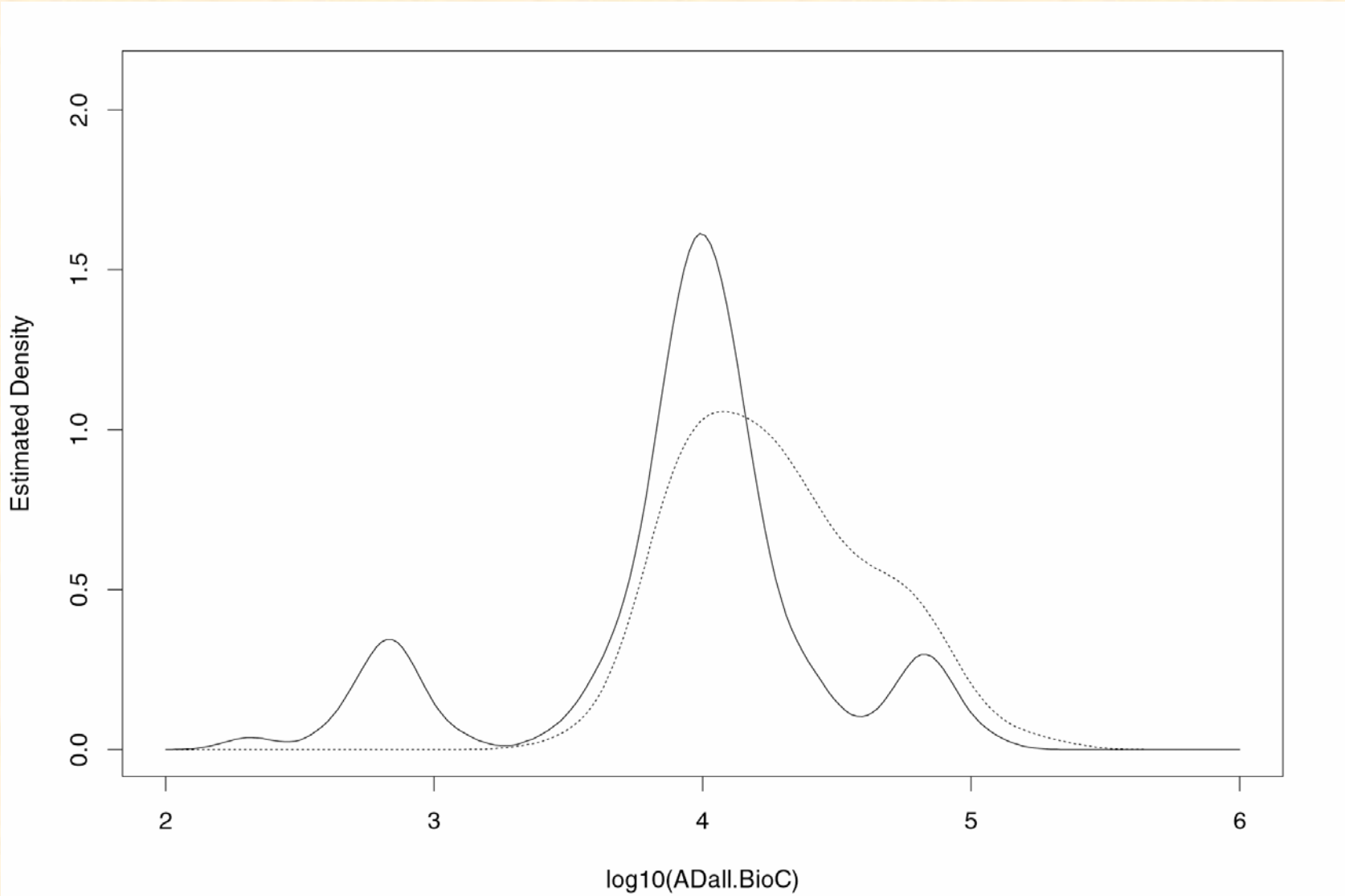

Figure 7. Density plots of the distribution of average difference values for the spike control BioC on Affymetrix Human U95A expression arrays. Cell lines = solid line and tissue = dashed line.

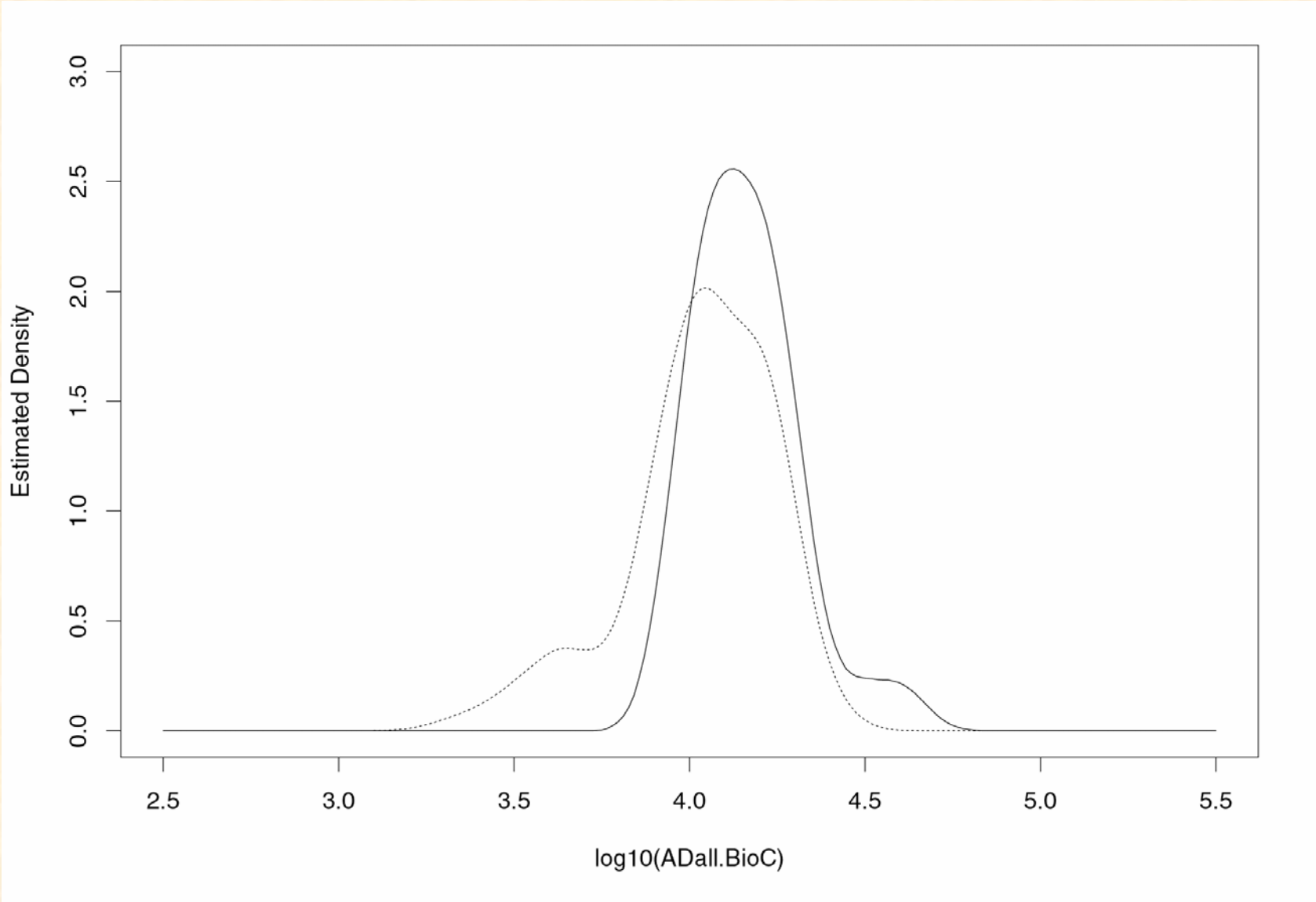

Figure 8. Density plots of the distribution of average difference values for the spike control BioC on Affymetrix Murine U74A expression arrays. Cell lines = solid line and tissue = dashed line.

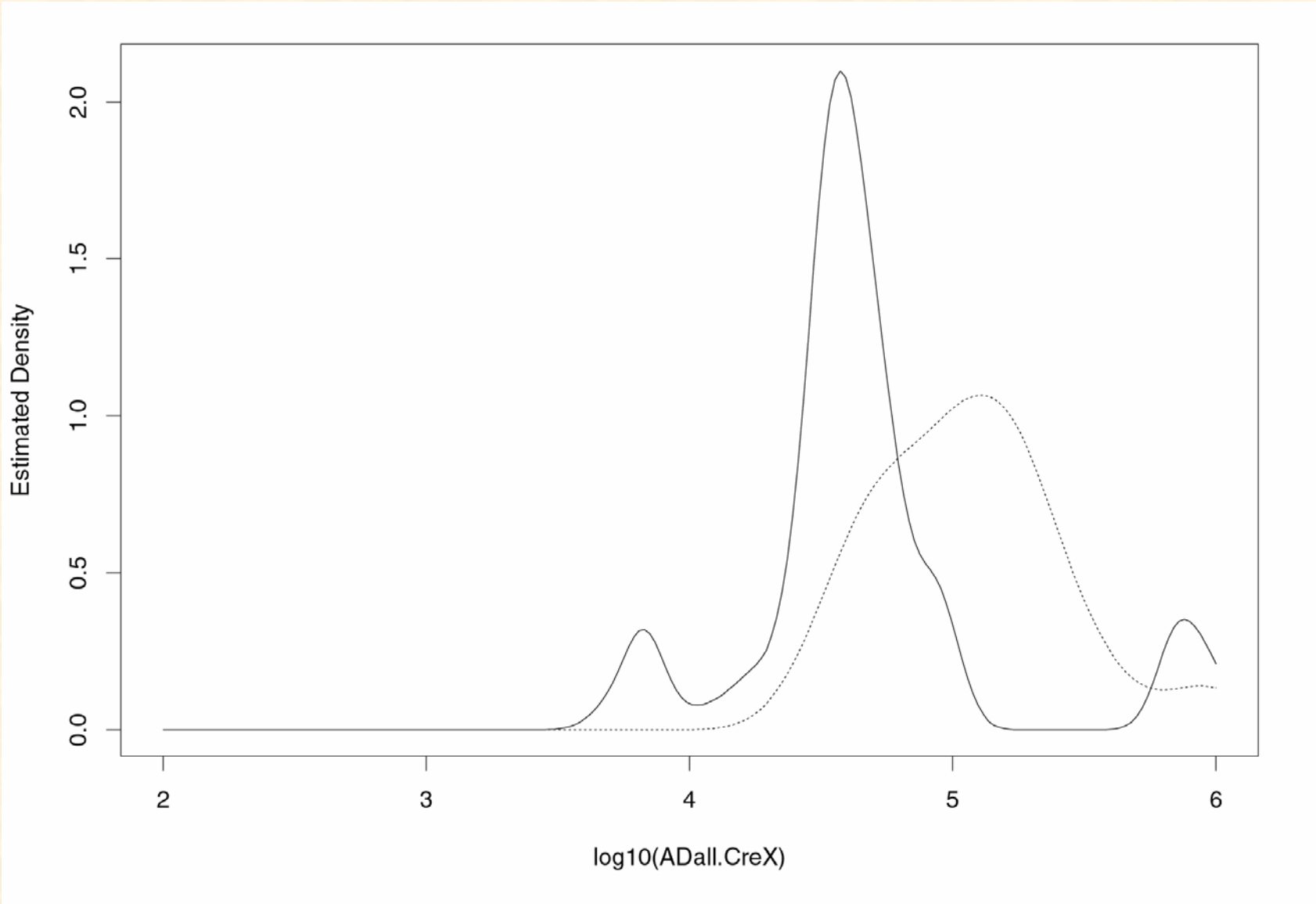

Figure 9. Density plots of the distribution of average difference values for the spike control CreX on Affymetrix Human U95A expression arrays. Cell lines = solid line and tissue = dashed line.

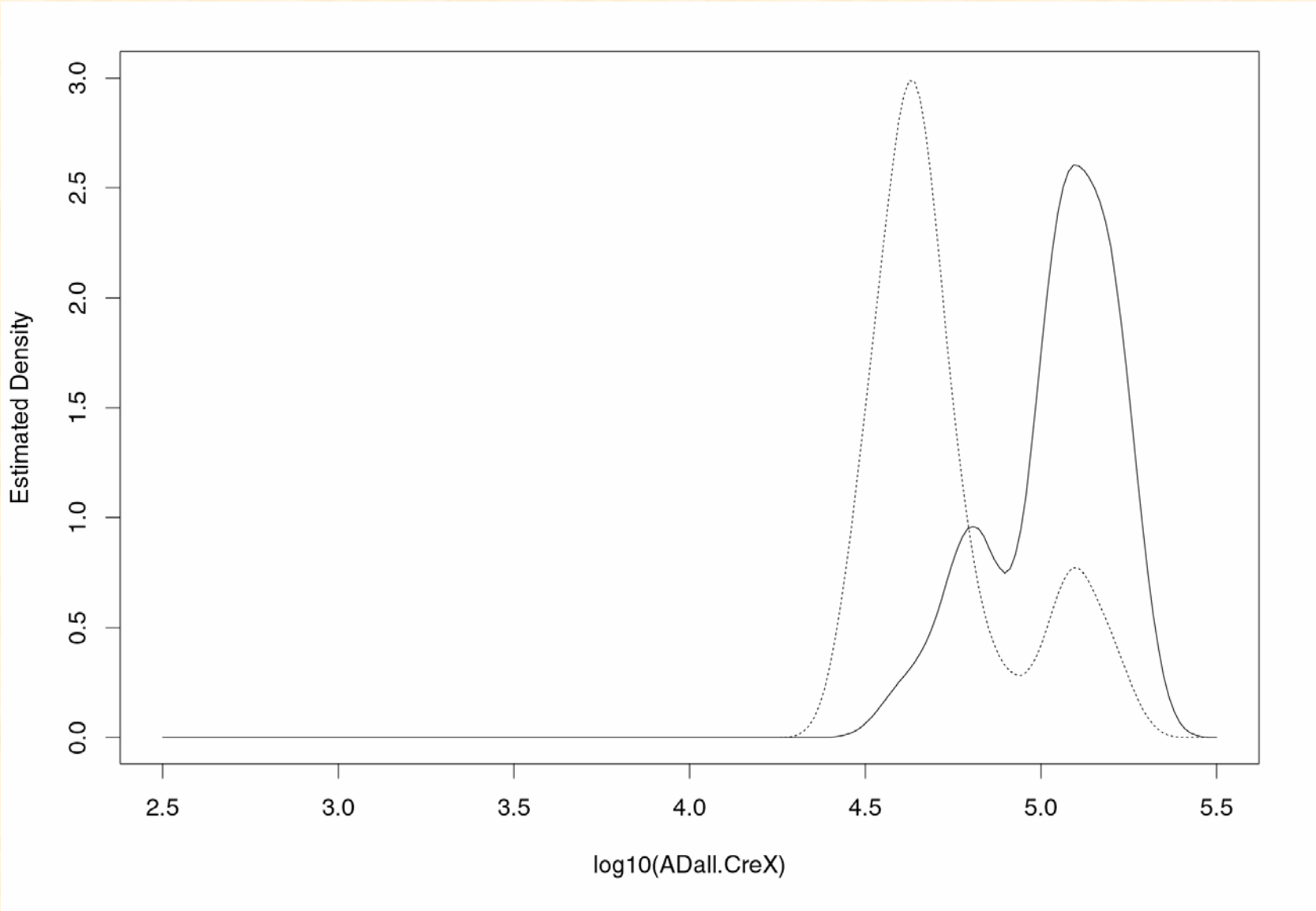

Figure 10. Density plots of the distribution of average difference values for the spike control CreX on Affymetrix Murine U74A expression arrays. Cell lines = solid line and tissue = dashed line.

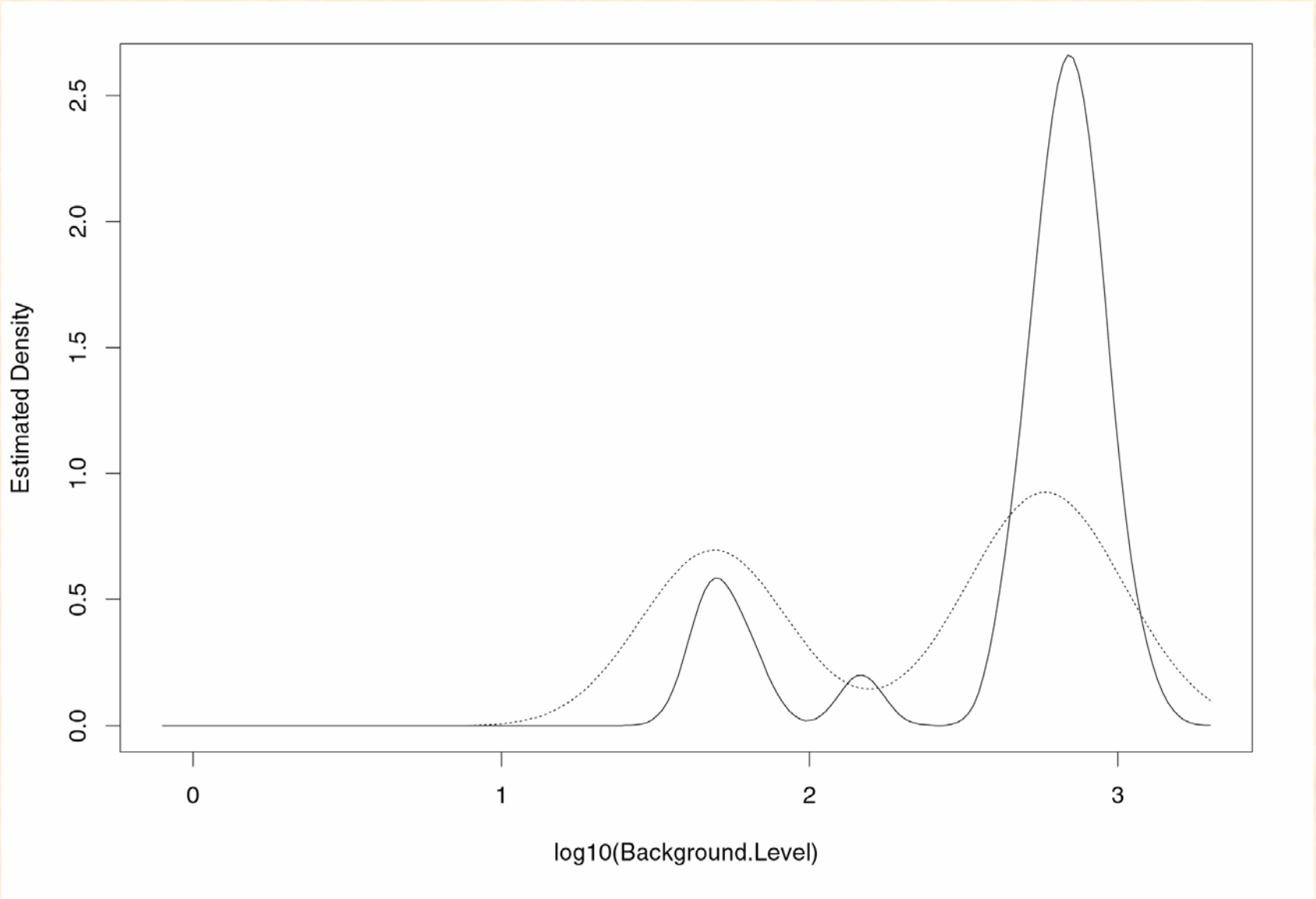

Figure 11. Density plots of the Background Level values on Affymetrix Human U95A expression arrays. Cell lines = solid line and tissue = dashed line.

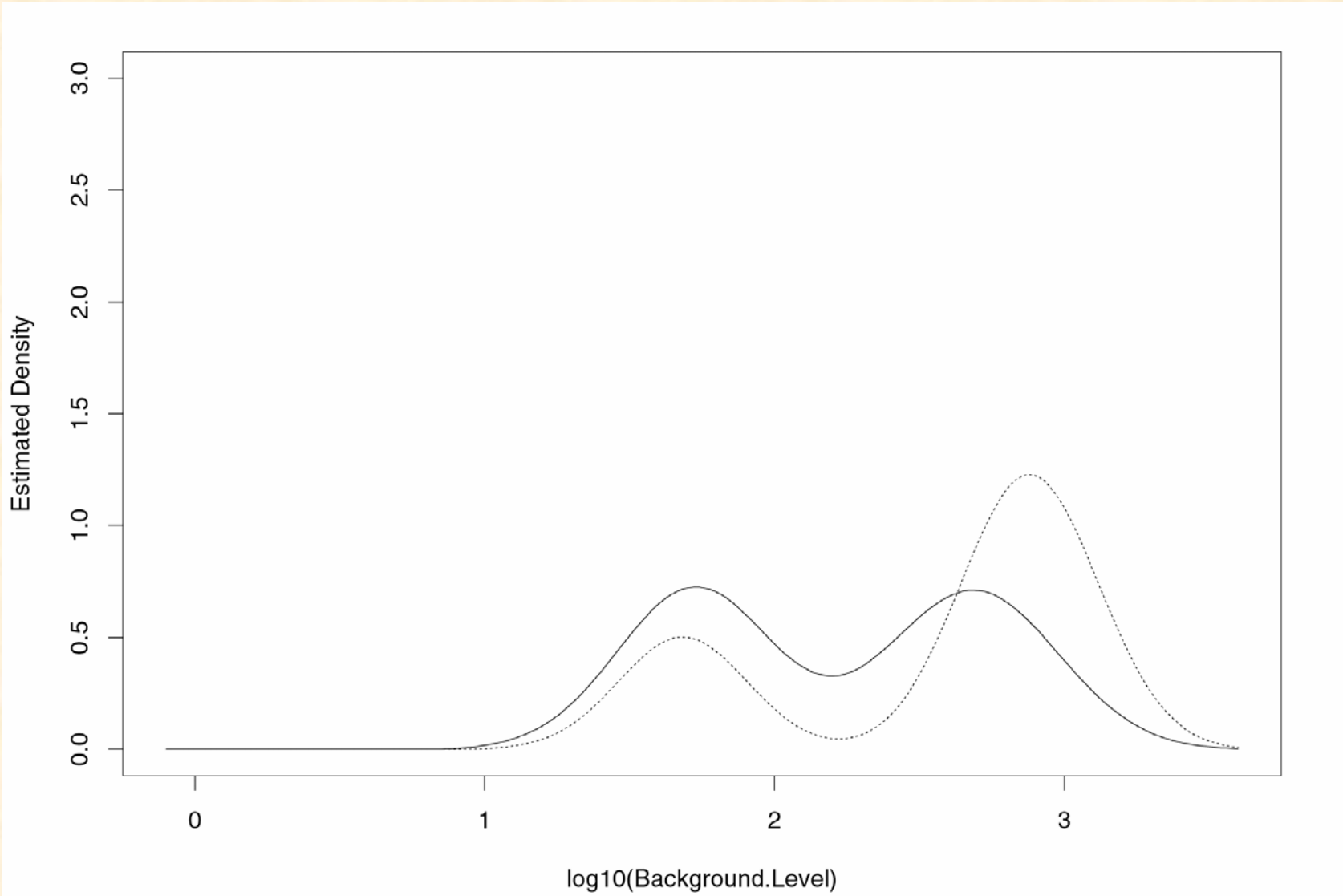

Figure 12. Density plots of the Background Level values on Affymetrix Murine U74A expression arrays. Cell lines = solid line and tissue = dashed line.

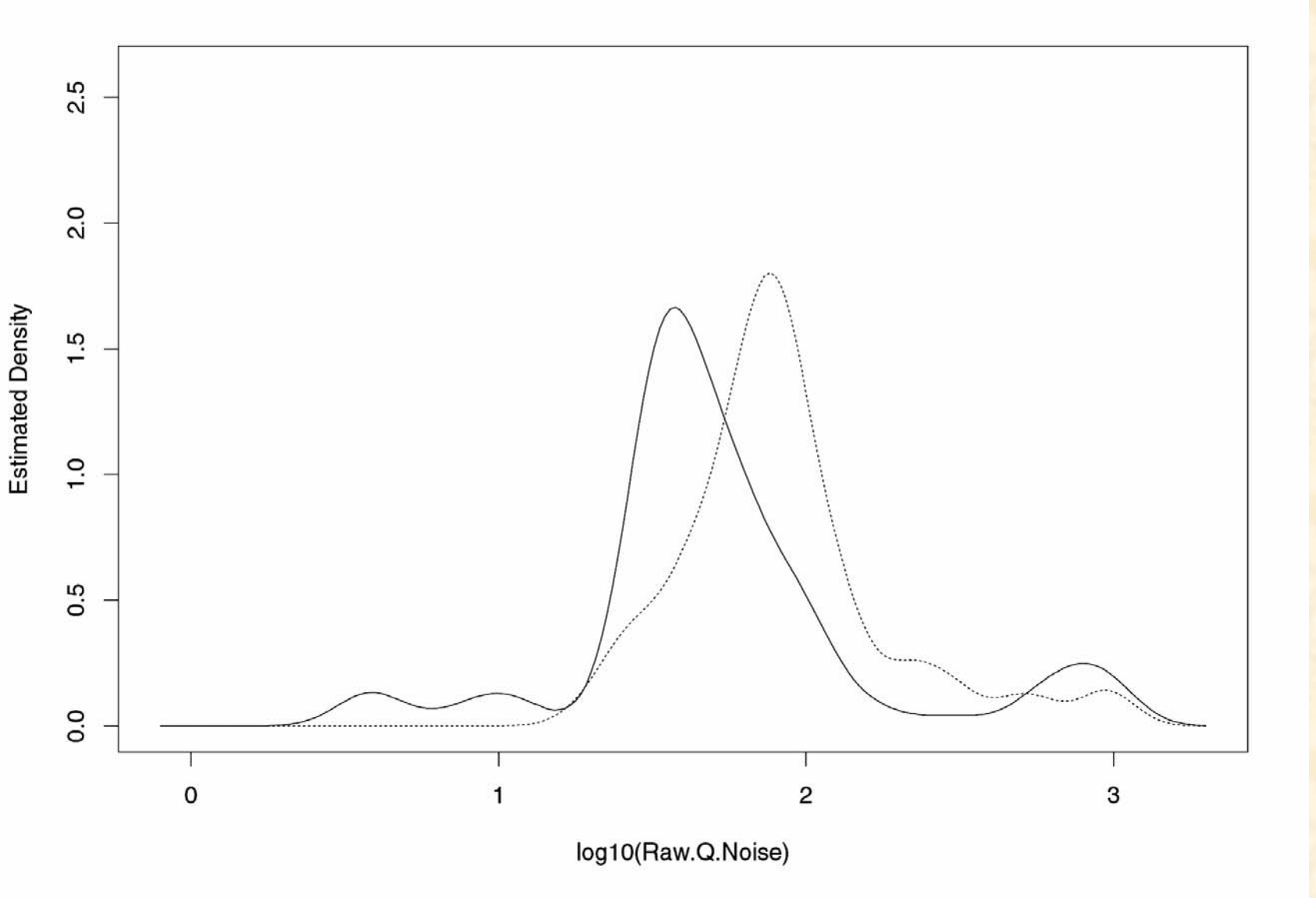

Figure 13. Density plots of the Raw Q noise values on Affymetrix Human U95A expression arrays. Cell lines = solid line and tissue = dashed line.

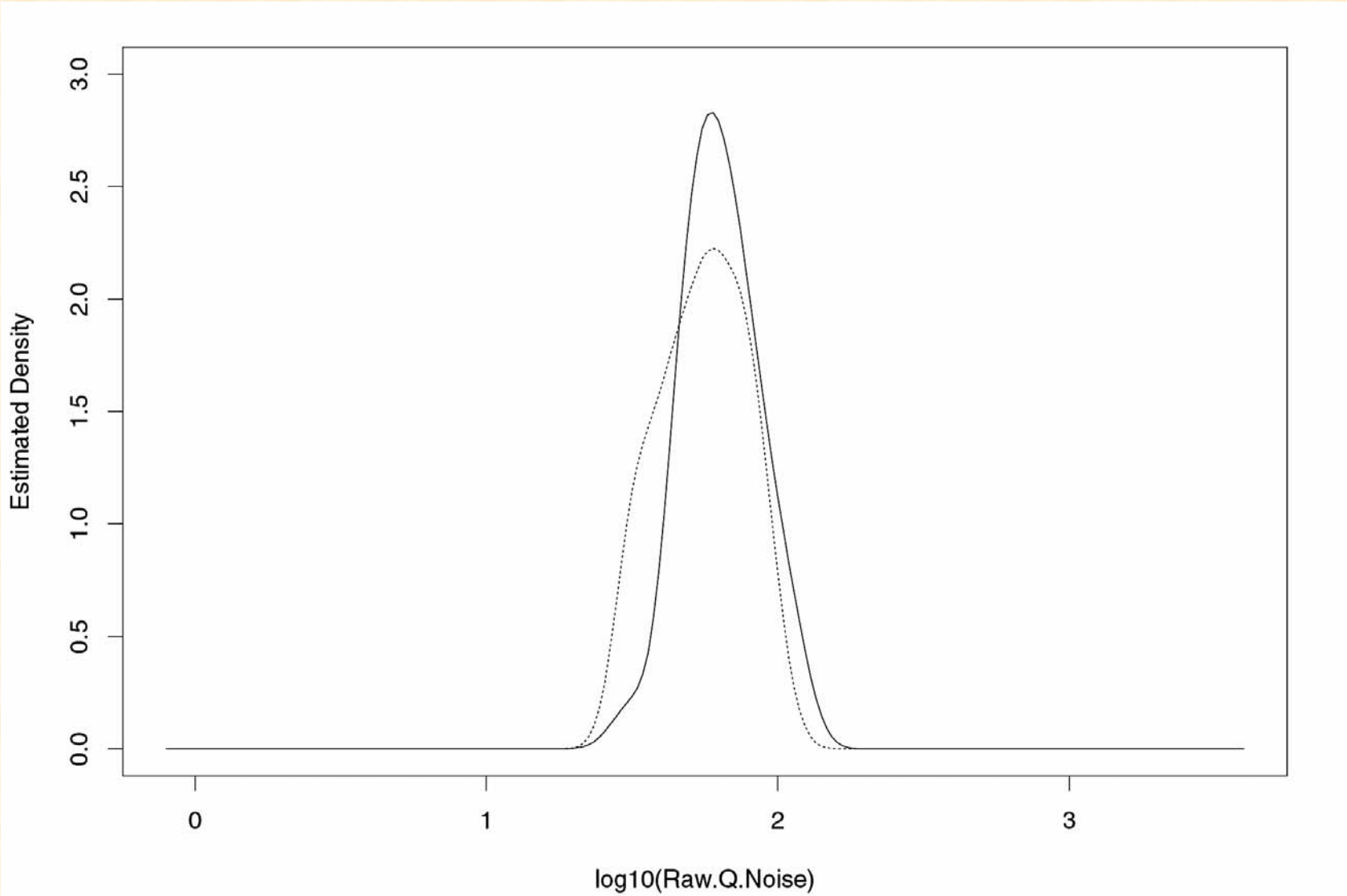

Figure 14. Density plots of the Raw Q noise values on Affymetrix Murine U74A expression arrays. Cell lines = solid line and tissue = dashed line.

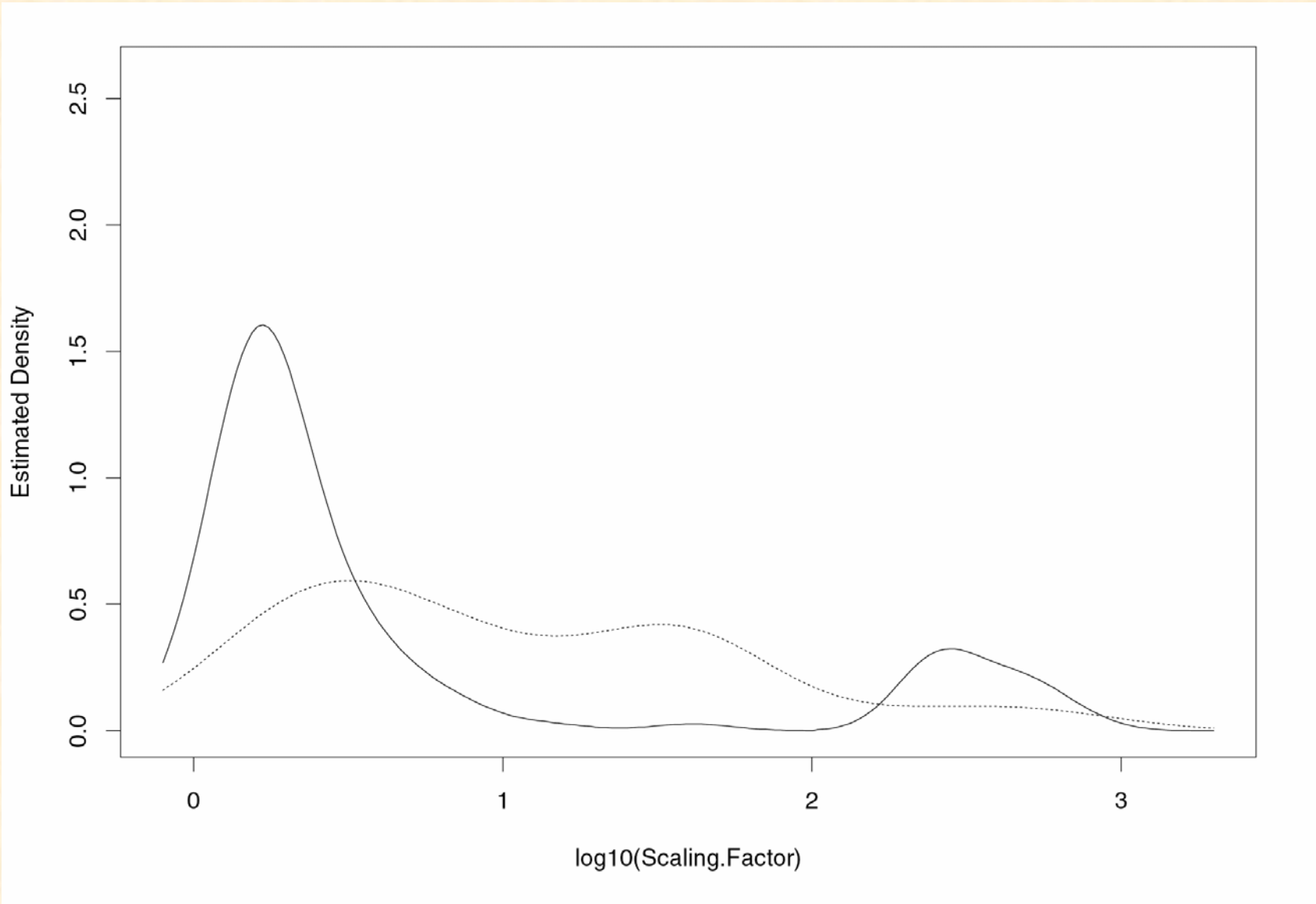

Figure 15. Density plots of the Scaling Factor values on Affymetrix Human U95A expression arrays. Cell lines = solid line and tissue = dashed line.

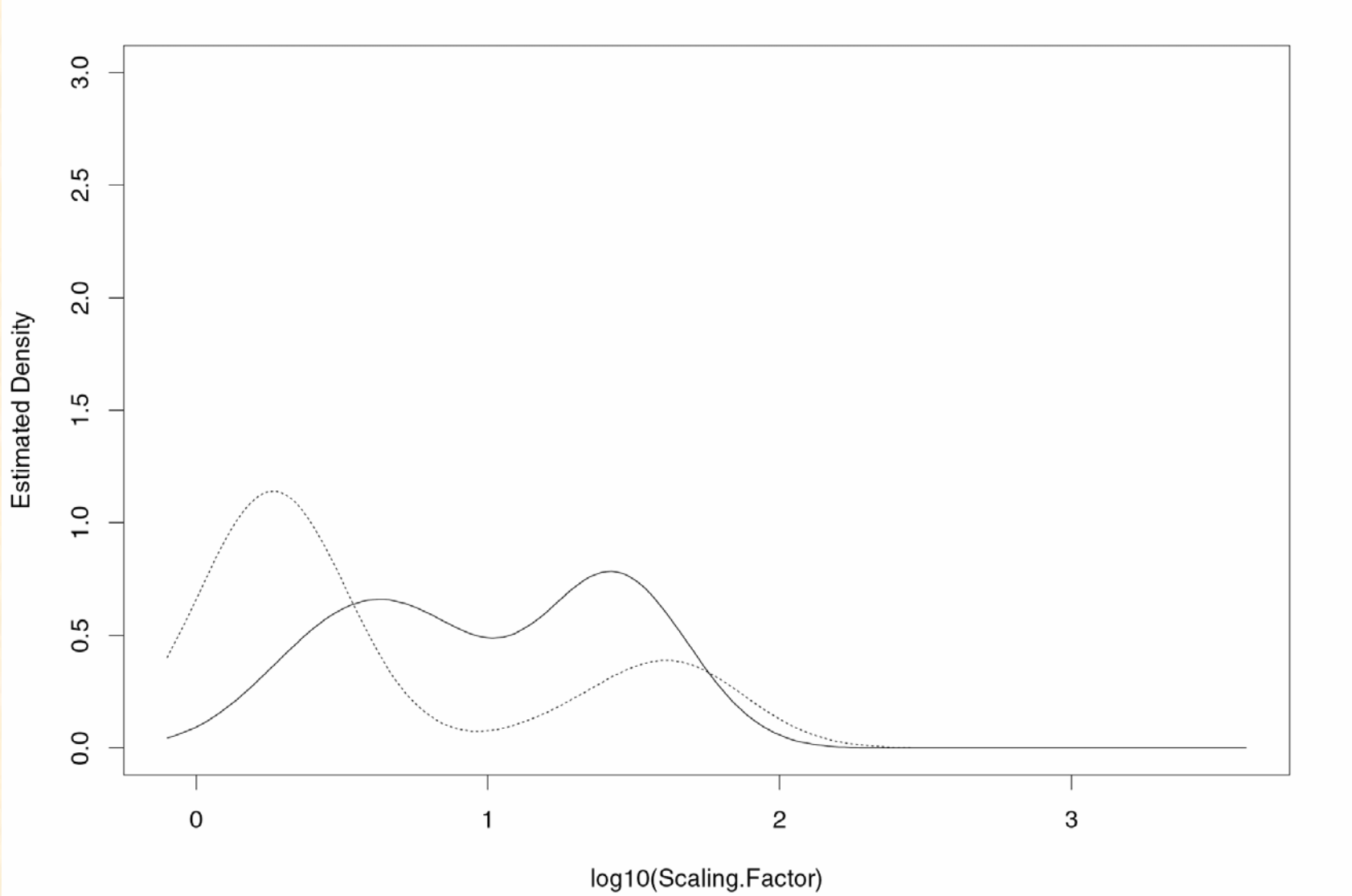

Figure 16. Density plots of the Scaling Factor values on Affymetrix Murine U74A expression arrays. Cell lines = solid line and tissue = dashed line.

Table 1. Ranking of the sources of error contributing to the variability for Affymetrix expression array metric and meta data

Scaling Factor	PMT[a]	>	Lab	>	type
Raw Q noise	Lab	>	PMT	=	type
Background level	PMT	>	Lab	>	type
Actin AD 3:5	Lab	>	PMT	>	type
GAPDH AD 3:5	Lab	=	PMT	>	type
BioB AD all	Lab	>	PMT	=	type
BioC AD all	Lab	>	Type	>	PMT
CreX AD all	Lab	=	PMT	>	type

[a]PMT = photomultiplier tube, Lab = site in which the arrays were run, type = tissue or cell line

Results of the Slide-Based Custom DNA Microarray Study

Table 2. Some sources of process errors

CAUSE(S)	CONSEQUENCES
Different self-extinguishing characteristics of the two fluorescent dyes and different light-to-current characteristics of the two photomultipliers	Necessity to use different accelerating voltages to improve signal-to-noise ratio -> Different amplifications in the two channels -> Multiplicative effect on expression ratio
Different surface tension, non-homology of pen tips, and/or non-uniform printing pressure, uneven glass slide thickness	Non-uniform spotting -> Systematic error manifested as abnormal, complex pen-domain periodicity in all chips of the spotting series
Random, non-uniform hybridization rate and/or incomplete removal of unhybridized probe and buffer	Detected fluorescence non-proportional to the cDNA abundance -> low reproducibility in repetitive experiments
Incomplete and/or non-uniform neutralization of the slide coat electric charge, non-uniform slide coating, or non-uniform post-printing slide processing	Slide coat non-uniform adsorption of labeled molecules -> Non-Gaussian distribution around the unit of the background pixel intensities -> reduced detection for higher adsorption (additive effect)
Different (unknown) probabilities to label the same cDNA with the two dyes, different (unknown) probabilities of a spot to capture distinctly labeled cDNAs, and different (unknown) fluorescent efficiencies of distinctly labeled cDNAs	Unknown, non-proportional relation between cDNA amounts ratio and net signals ratio

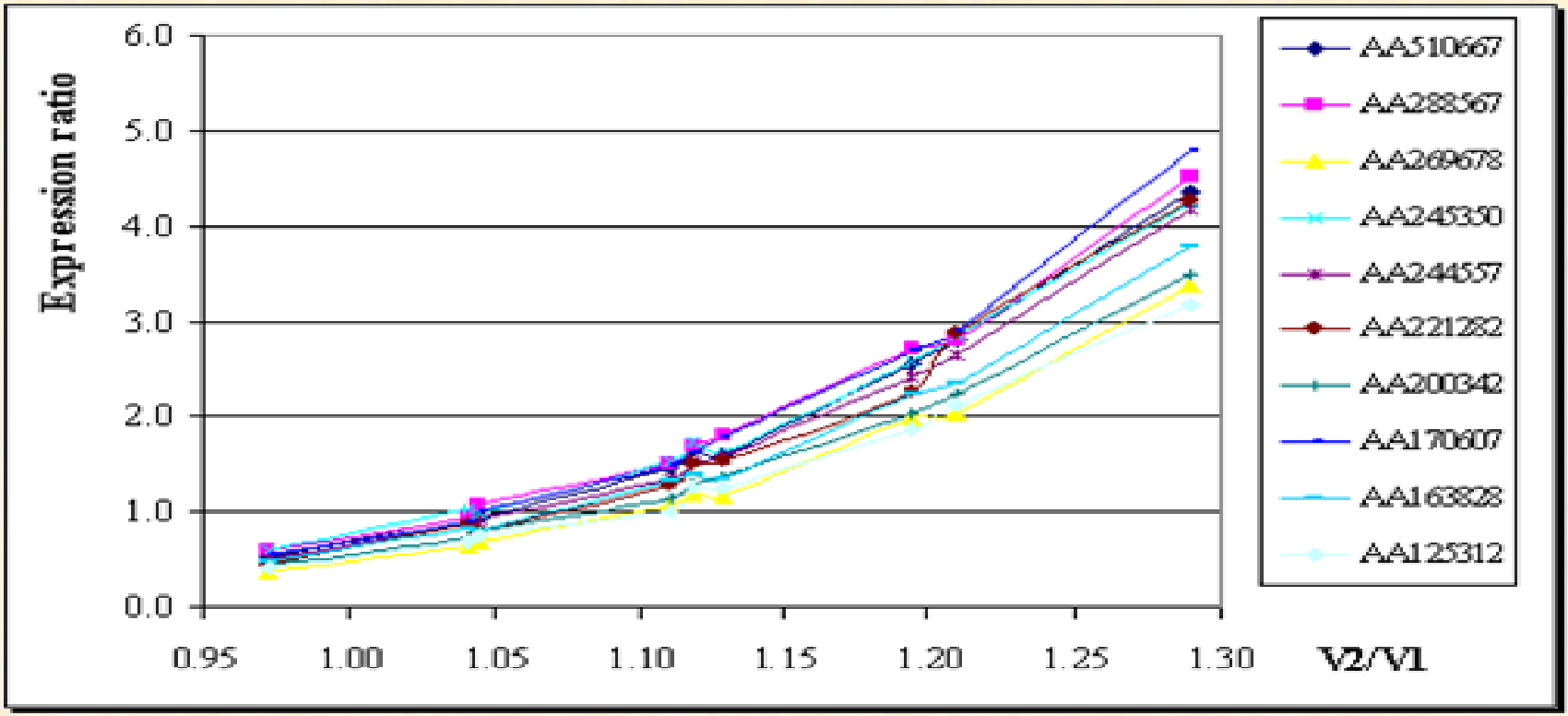

Figure 17. Assessing scanner variability: dependence of the observed gene expression ratio on the accelerating voltage ratio. Values determined for the first ten clearly expressed genes indicated by their accession numbers. The 9k microarray chip, hybridized with reverse transcribed RNA extracts from wild type and Cx43 knockout mice brain, has been scanned with the following accelerating voltage pairs (V2/V1): 700V/720V = 0.972, 750V/720V = 1.042, 700V/670V = 1.045, 800V/720V = 1.111, 750V/670V = 1.119, 700V/620V = 1.129, 800V/670V = 1.194, 750V/620V = 1.210, 800V/620V = 1.290. The expression ratios have been modified between the two extreme voltage ratios by factors ranging from 7.24 to 9.21.

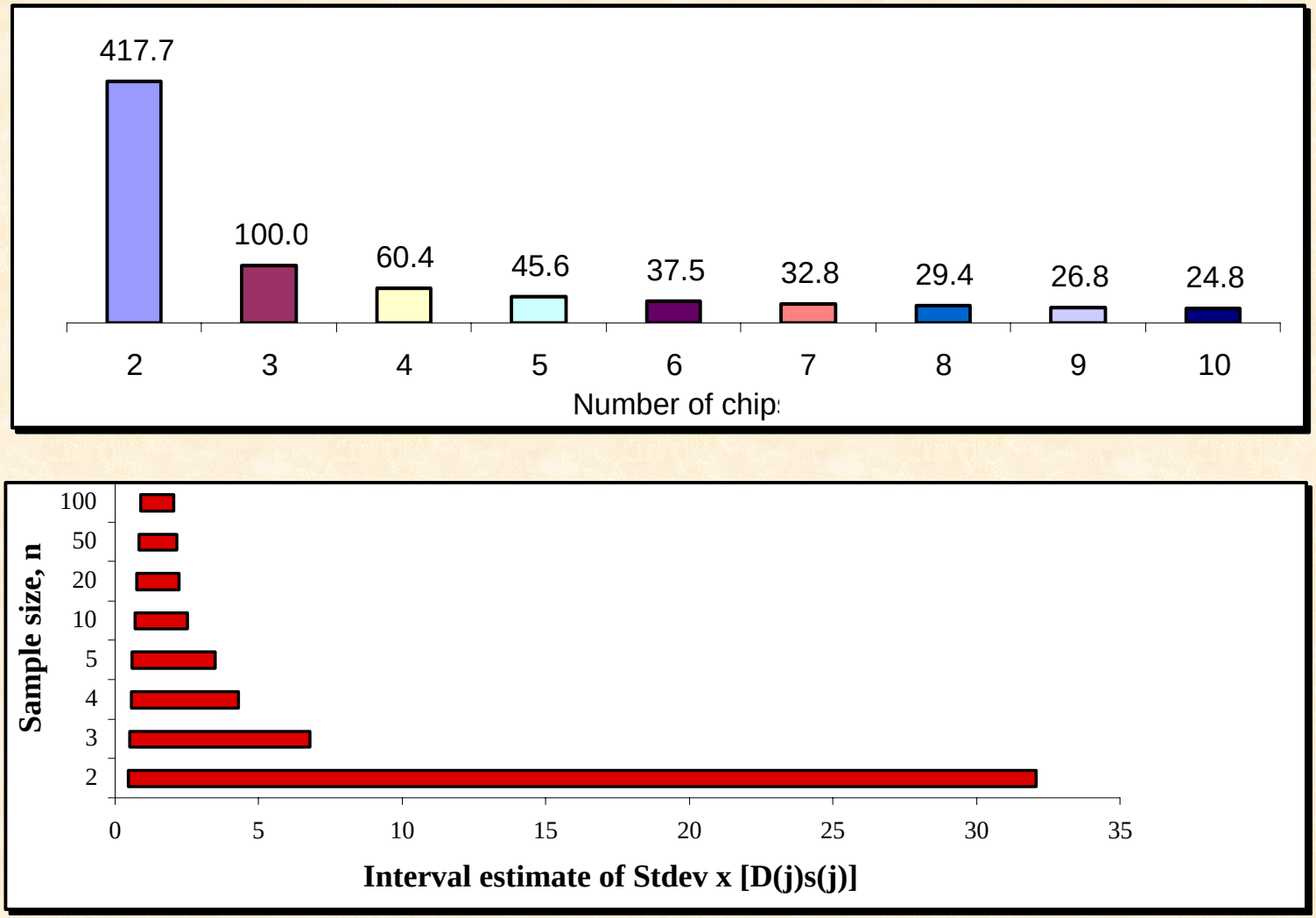

Figure 18. Need for chip replication. The estimated error of the expression ratio for various number of replications, using a relative reference of 3 chips = 100 (Top). The interval estimate of the standard deviation of the expression ratio as a function of number of replications (Bottom).

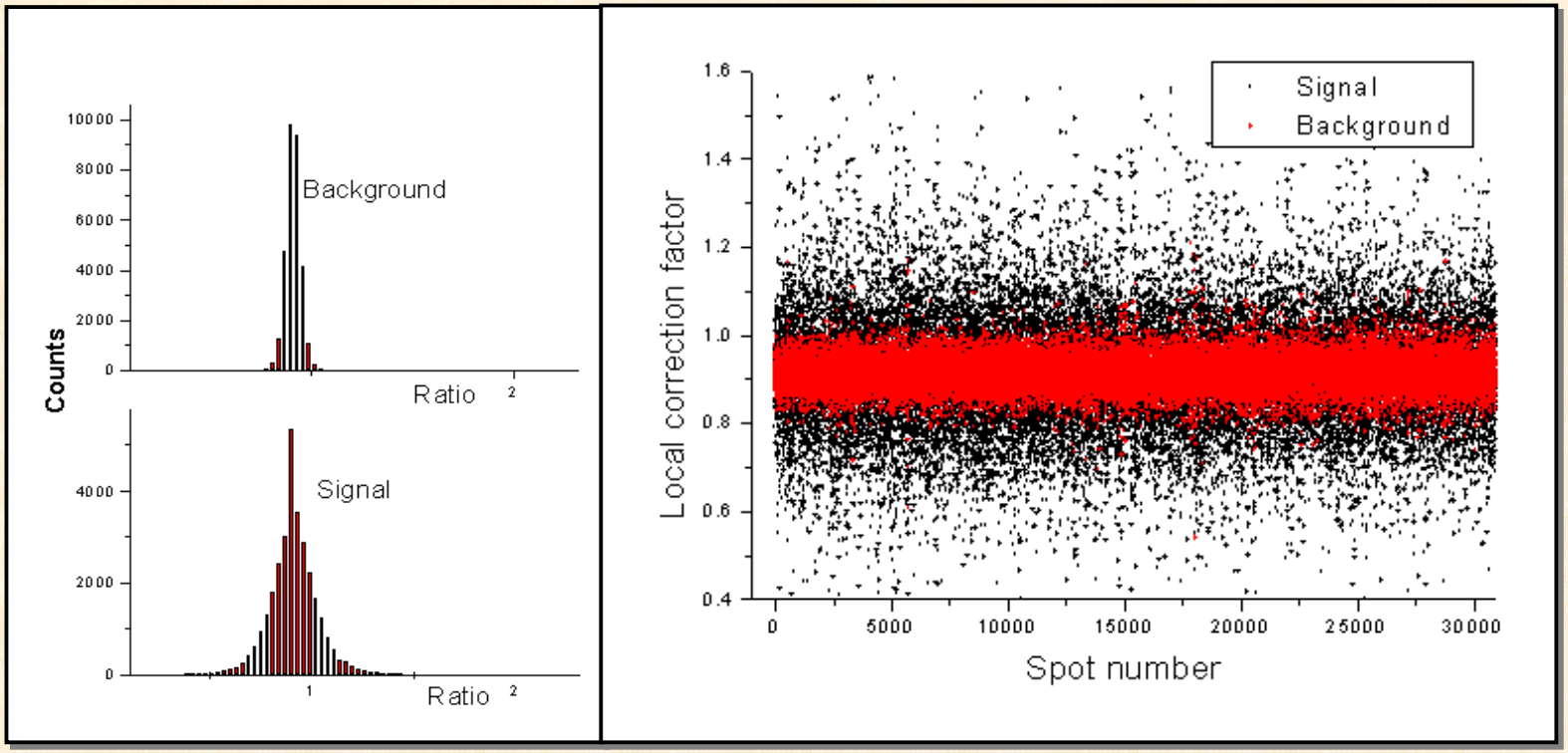

Figure 19. Signal and background local correction factors in the yellow test performed on a dually labeled spleen extract hybridized with a 31,200 spots microarray. The correction factors are restoring the Gaussian distributions of the ratios of medians of spot or background pixel intensities around the global correction factor that balances the two channels. The range of signal correction factors is higher than that of background (left panel) because of larger number of sources of variability affecting the signal.

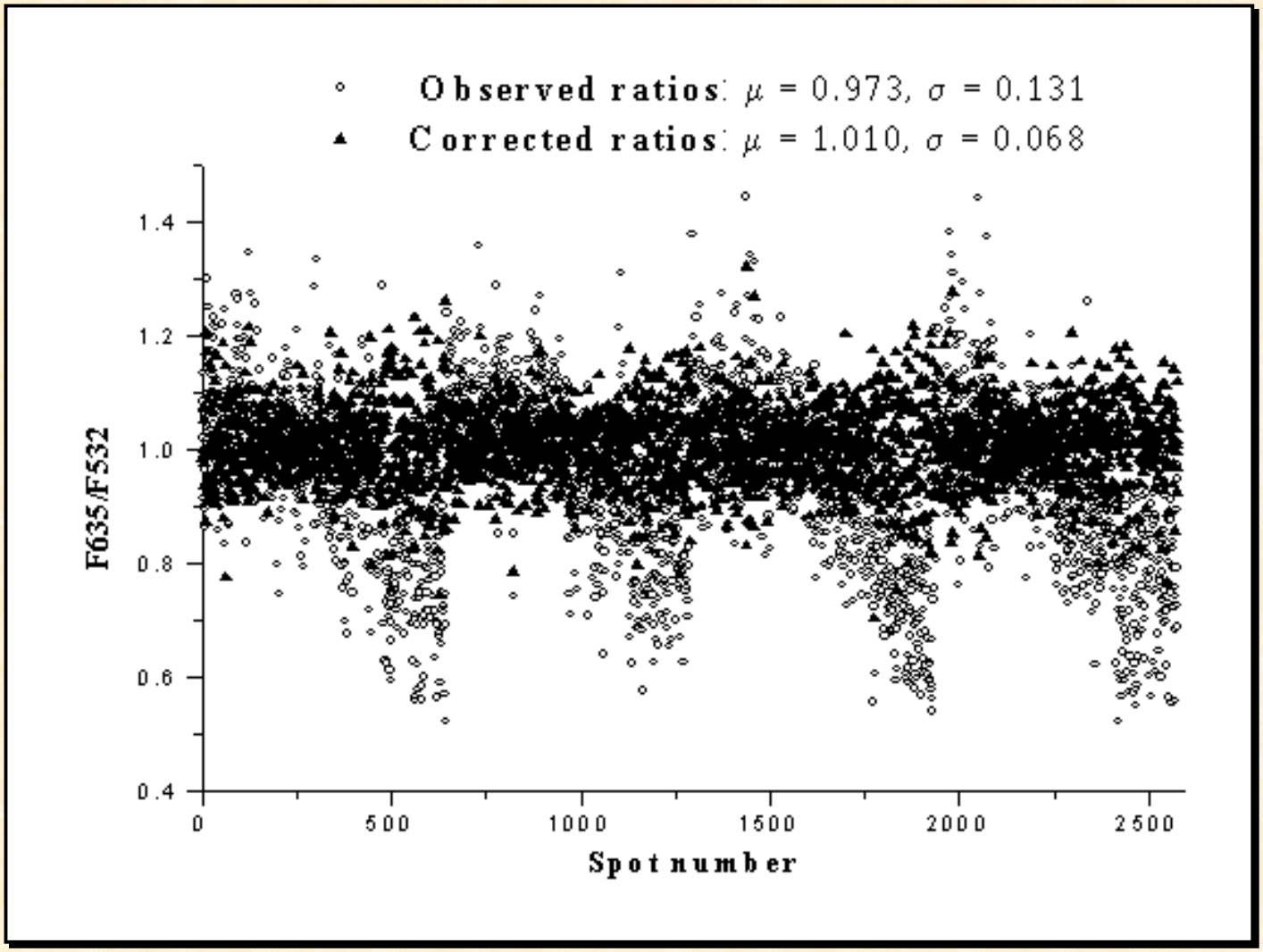

Figure 20. Correction of processing error. 2.5k mouse test chip hybridized with differently labeled halves of the same extract from N2A mouse neuroblastoma. The correction restores the Gaussian distribution of the ratio of medians of spot pixel intensities around the global correction factor *h*.

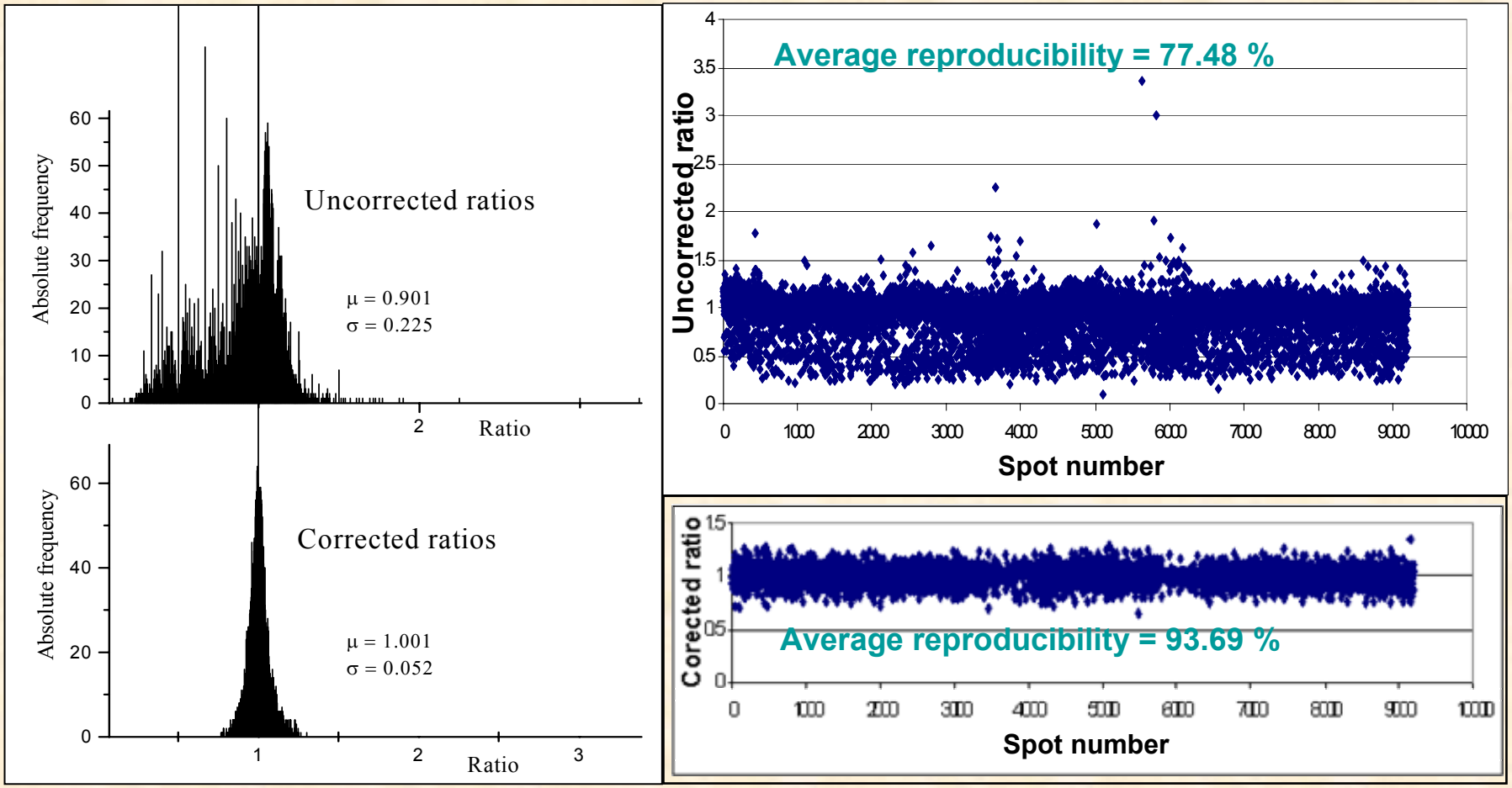

Figure 21. Correction efficiency. Verified by the "yellow test" in the experiment alpha

General Conclusions

(1) Lab-to-lab variation accounted for the greatest source of error in our Affymetrix study. This suggests that .CHP data generated by different institutions may not be easily compared without further normalization in comparative analyses.

(2) The observed variance in the AD values for the exogenous control spikes suggests that the controls may not be an adequate tool to normalize data for comparison analysis. It has been theorized that AD values should be independent of sample and array type.

(3) "Yellow" test is important in characterizing the errors due to the Process (probe labeling, sequence-dependent dye biases, slide treatment, hybridization, scanning etc.).

(4) Multiple experiments can best be compared by either using a pooled reference in one channel or by alternate labeling (dye flipping).

(5) Biological reproducibility should be demonstrated by repeating each experiment a minimum of 3 or 4 times with different extracts of the same type.

(6) Systemic, reproducible errors can be minimized by applying various algorithms and improve the average reproducibility from ~77% to ~93%. The caveat however is that one should not try to rescue a poor hybridization result with mathematical manipulations.