Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2006 *Paper* 5

New Spiked-In Probe Sets for the Affymetrix HGU-133A Latin Square Experiment

Monnie McGee* Zhongxue Chen†

*Southern Methodist University, mmcgee@smu.edu

†University of Texas Southwestern Medical Center, zhongxue@smu.edu

New Spiked-In Probe Sets for the Affymetrix HGU-133A Latin Square Experiment

Monnie McGee and Zhongxue Chen

Abstract

The Affymetrix HGU-133A spike in data set has been used for determining the sensitivity and specificity of various methods for the analysis of microarray data. We show that there are 22 additional probe sets that detect spike in RNAs that should be considered as spike in probe sets. We assign each proposed spiked-in probe set to a concentration group within the Latin Square design, and examine the effects of the additional spiked-in probe sets on assessing the accuracy of analysis methods currently in use. We show that several popular preprocessing methods are more sensitive and specific when the new spike-ins are used to determine false positive and false negative rates.

New Spiked-In Probe Sets for the Affymetrix HGU-133A Latin Square Experiment

Abstract:

The Affymetrix HGU-133A spike in data set has been used for determining the sensitivity and specificity of various methods for the analysis of microarray data. We show that there are 22 additional probe sets that detect spike in RNAs that should be considered as spike in probe sets. We assign each proposed spiked-in probe set to a concentration group within the Latin Square design, and examine the effects of the additional spiked-in probe sets on assessing the accuracy of analysis methods currently in use. We show that several popular preprocessing methods are more sensitive and specific when the new spike-ins are used to determine false positive and false negative rates.

Introduction:

Since the appearance of the first Affymetrix GeneChip (1), there have been many preprocessing methods devised to deal with its unique design of perfect match (PM) and mismatch (MM) probe sets. The first of these was MAS 4.0 (2), later followed by MAS 5.0 (3). Other popular methods are dChip (4, 5), RMA (6), GCRMA (7), and PLIER (8).

In order to help assess the performance of these preprocessing methods, Affymetrix developed the HG-U95Av2 and HG-U133A spike-in data sets, available at http://www.affymetrix.com/support/technical/sample_data/datasets.affx , in which a subset of specific targets were added to the hybridization sample. Since the transcripts were spiked in at known concentrations for these data sets, it is possible to compare the sensitivity and specificity of the various analytical methods. The first data set, on the HGU-95Av2 platform, consists of 14 target transcripts that are spiked in at 14 different concentrations, ranging from 0 pM to 1024 pM. The experiment is arranged in a Latin Square design, such that each target transcript is spiked in at each concentration. With the inclusion of replicates, there are 59 arrays in this data set. It has been noted that two spiked-in probe sets, 407_at and 37777_at, contained poorly performing probes. In addition, two other probe sets behave similarly to the 14 probe sets that Affymetrix reported as recognizing spike-in transcripts (9).

Another series of spike-in experiments was performed on the HGU133A platform. This experiment differs from the HG-U95Av2 experiment in several important ways. First, the HG-U133A experiment consists of 42 specific transcripts that are spiked in at 14 concentrations ranging from 0 pM to 512 pM, again arranged in a Latin Square design. Therefore, there is a finer gradation of concentrations used than in the HG-U95Av2 experiment. Also, there are three transcripts spiked-in at each concentration and three replicate arrays for each experiment, thus a total of 42 arrays. Appendix A gives a table of the 42 probe sets that were defined as recognizing the spiked-in transcripts in the original experiment and their associated concentrations with each hybridization experiment. For convenience, we will call the triples of probe sets that recognize transcripts spiked-in at the same concentration “groups”. Group 1 consists of the probe sets in the first row; group 2 consists of the probe sets in the second row, and so on.

Another important way in which the HG-U133A experiment differs from the HG-U95Av2 experiment is that the former requires a completely different chip description file (CDF) than the commercially available platform on which it is based. The CDF for the HG-U133A spike-in experiment is named “HG-U133Atag”. This CDF contains information on the placement of seventeen more probe sets than are available in the CDF for the HG-U133A platform. This is important because eight of the extra probe sets recognized spiked-in transcripts, while the other nine do not. Table 1 shows the frequency of probe sets with given numbers of probe pairs (8, 10, 11, 13, 14, 15, 16, 20 and 69) in the HG-U133A and HG-U133Atag chip description files. For example, there are 482 probe sets in both CDFs that contain 16 probe pairs each. Similarly, there are 21748 probe sets with 11 probe pairs in the HG-U133A CDF, while the HG-U133Atag CDF has 21765 probe sets with 11 probe pairs each.

# Probe Pairs	8	10	11	13	14	15	16	20	69
HG-U133A	1	1	21748	4	4	2	482	40	1
HG-U133A tag	1	1	21765	4	4	2	482	40	1

Table 1: Number of probe sets containing 8, 9, 10, 11, 13, 14, 15, 16, 20, or 69 probe pairs each in HG-U133A and HG-U133A tag CDFs.

Table 2 gives the identification numbers of the additional seventeen probe sets and their classification as recognizing spiked-in transcripts or not.

Spike Ins	Non Spike Ins
AFFX-r2-TagA_at	AFFX-r2-TagO-3_at
AFFX-r2-TagB_at	AFFX-r2-TagO-5_at
AFFX-r2-TagC_at	AFFX-r2-TagIN-3_at
AFFX-r2-TagD_at	AFFX-r2-TagIN-5_at
AFFX-r2-TagE_at	AFFX-r2-TagQ-3_at
AFFX-r2-TagF_at	AFFX-r2-TagQ-5_at
AFFX-r2-TagG_at	AFFX-r2-TagJ-3_at
AFFX-r2-TagH_at	AFFX-r2-TagJ-5_at
	AFFX-r2-TagIN-M_at

Table 2: Probe IDs for the seventeen additional probe sets annotated in the HG-U133Atag CDF.

These probe sets were designed to recognize artificial sequences for quality control purposes, and should not cross-hybridize with either the spike-in transcripts or the combined RNA background used for the experiment.

Additional information included in files that are downloaded with the spiked-in data contains the currently known problems with cross-hybridization. More precisely, three additional probe sets contained probes exactly matching some of the spiked in transcript sequences. Table 3 is a reproduction of that table from the Affymetrix explanatory Excel spreadsheet for the HG-U133 Latin Square spike-in experiment.

Spike Probe Set	Additional Matching Probe Set	Number of Matching Probes
212827_at	209374_s_at	11/11
205398_at	205397_x_at	5/11
206060_s_at	208010_s_at	9/11

Table 3: Probe Sets known to behave similarly to Affymetrix Spike-in Probe Sets at the time that our analysis was conducted.

There is a further list of cross-hybridizing probe sets given in the affycomp package (9), a part of the Bioconductor software suite (10). A BLAST query of the sequences of the spiked-in clones against the all HG-U133A target sequences was performed. There are 145, 240, and 271 probe sets on the HG-U133A platform that match at least 200, 150, or 100 base pairs, respectively, of one or more spiked-in probe sets. The probe sets given in Table 3 are among the list of 145 probe sets with at least 200 bp (i.e. over 70%) matching sequences of the spiked-in probe sets.

In this paper we show that twenty-two additional probe sets can be considered as spiked-in probes according to criteria listed in the following section, apparently due to cross-hybridization to spiked-in transcripts. In other words, there are actually 64 spiked-in clones in the Affymetrix HG-U133 spike-in data set, instead of the original 42. We catalog each new spike-in probe set according to the pattern of target transcript concentrations within the Latin Square design. Further, we examine the

impact of the new spike-ins probe set annotation on the performance of popular analysis methods such as RMA, GCRMA, MAS 5.0, dChip, and PLIER.

Materials and Methods:

Two sets of tests were used to search the data for additional spike-ins. First, we analyzed the HGU133A spike-in data using MAS 5.0, RMA, and GCRMA, and examined the log fold changes in the expression values between pairs of experiments for which the concentrations of the spiked-in probe sets were separated by 1 permutation in the Latin Square design (*e.g*. Experiment 1 vs Experiment 2, Experiment 2 vs. Experiment 3, etc). The absolute value of the log base 2 fold change for each experiment was calculated and sorted in descending order. Any probe set which displayed absolute fold changes greater than 1 for at least one of the three methods mentioned previously (RMA, GCRMA, or MAS 5.0) were retained as possible additional spiked-in probe sets. For example, after all pairs of experiments were examined after preprocessing the data with RMA, there were 26 probe sets with a log base 2 fold change greater than 1.

Next, graphical displays of the probe-by-probe intensity levels for each candidate probe set across all arrays were examined. Intensity levels were normalized with quantile normalization before plotting. Probe sets that tended to display a range of intensity levels less than 1 (on a log base 2 scale) were eliminated from the consideration set. For the remaining candidates, we further examined the pattern of intensity levels across experiments to classify it into a particular spike-in group (defined in the introduction). If a suitable group was not found, the candidate was eliminated. We also considered biological similarity as a criterion. Biological similarity was determined by examining the reported functions of the various probe sets as given on the NetAffx Analysis Center website: http://www.affymetrix.com/analysis/index.affx.

Once the list of new spiked-in probe sets was complete, we reanalyzed the spike-in data using RMA, RMA with no background correction (RMA-noBG), GCRMA, MAS, dChip, and PLIER in order to determine the performance of these methods with the new spiked-in probe sets included. Although all 42 files were preprocessed together for each method, we compared pairs of experiments that were separated by the same number of permutations of the Latin Square (where d = number of permutations), and obtained average true and false positive rates for each preprocessing method within each grouping. In this Latin Square design, d can be thought of as the log2 fold difference in spike-in transcript levels for a majority of the transcripts. For example, Experiments 1 and 2, 2 and 3, 3 and 4, etc. are separated by one shift in the Latin Square design; therefore, d = 1 for these experiments. For twelve of these fourteen pairs of experiments, there is a 2 fold difference in spike-in transcript levels. Similarly, experiments 3 and 5, 4 and 6, and 5 and 7 are separated by two permutations in the Latin Square design; therefore, d = 2. Eleven of the experiments have fold changes of 2 on the log base 2 scale. We compared experiments with d = 1 through d = 7, since d=8 is equivalent to d=6, d=9 equivalent to d=5 and so on. ROC curves and the corresponding areas under the curves (AUCs) are given in the next section.

Results:

Obtaining Candidate Spiked-In Probe Sets: Table 4 illustrates how RMA was used to obtain the expression values for each probe and each experiment. In this example, we examine the log base 2 fold changes between two experiments (e.g. Experiment 1 and Experiment 2). Absolute values of the log base 2 fold changes were calculated and arranged in descending order. The first twenty such probe sets, their expression values per experiment and their absolute log base 2 fold changes are shown. The designated spike-in genes are shown in boldface type.

There are several interesting things to note about this table. First, the probe sets showing the largest fold change differences are not labeled as spiked-in probe sets. In fact, only four of the top twenty are so labeled. Next, the expression values in the first three columns and the next three columns

are remarkably similar. In other words, there is good consistency across the experiments where the expression values are concerned. This is a characteristic of RMA, which tends to give reasonably reproducible results across replicates of the same experiment (9). However, similar tables where the data were analyzed using MAS 5.0 and GCRMA (see Appendix B) also show the same pattern.

ProbeID	Exp1R1	Exp1R2	Exp1R3	Exp2R1	Exp2R2	Exp2R3	Abs FC	Spike In
AFFX-r2-Bs-lys-5_at	5192	4986	5432	10	9	10	9.071	N
AFFX-r2-Bs-phe-5_at	6045	5707	6000	12	11	12	9.004	N
AFFX-r2-Bs-phe-M_at	5560	5501	5331	11	11	11	8.994	N
AFFX-LysX-5_at	5641	5330	5559	11	12	11	8.931	N
AFFX-ThrX-M_at	5926	5348	5873	14	14	14	8.708	N
AFFX-r2-Bs-lys-M_at	4582	4434	4707	12	11	10	8.695	N
AFFX-PheX-M_at	5088	5000	5159	13	12	14	8.604	N
AFFX-PheX-5_at	5725	5510	5704	15	15	14	8.588	N
AFFX-r2-Bs-thr-5_s_at	7928	8113	7632	22	24	19	8.509	N
AFFX-r2-Bs-thr-M_s_at	7630	8065	7865	23	18	24	8.5	N
AFFX-r2-Bs-thr-3_s_at	7724	7556	7959	22	27	22	8.346	N
AFFX-LysX-3_at	**4892**	**4478**	**4858**	**15**	**15**	**14**	**8.307**	**Y**
AFFX-ThrX-3_at	**8430**	**8343**	**8344**	**29**	**26**	**26**	**8.282**	**Y**
AFFX-r2-Bs-lys-3_at	3547	2923	3928	12	11	12	8.213	N
AFFX-LysX-M_at	5061	5227	4935	19	18	19	8.113	N
AFFX-r2-Bs-phe-3_at	4573	4124	4482	18	18	18	7.919	N
AFFX-ThrX-5_at	5324	4730	4744	26	24	24	7.656	N
AFFX-PheX-3_at	**4844**	**4910**	**4951**	**40**	**39**	**35**	**7.001**	**Y**
205267_at	**116**	**108**	**100**	**210**	**231**	**238**	**1.069**	**Y**

Table 4: Expression values and absolute value of log base 2 fold change (Abs FC) from an analysis of Experiments 1 and 2 in the HGU133A spike-in data set using RMA. Note that the 11 largest fold changes do not belong to probe sets that had not been described as recognizing spiked-in transcripts.

As previously mentioned, data preprocessed using RMA yielded 26 such probe sets. For GCRMA and MAS 5.0, there were 27 and 28, respectively. A log base 2 fold change cutoff of 4 (instead of 1) for MAS 5.0 was used, because the cutoff of 1 produced hundreds of possible candidates. A list of the final 30 probe sets meeting the criterion is given in Table 5. Note that there are 30 because some probe sets passed the criterion based on only one method, while others passed for all methods.

204891_s_at	209374_s_at	AFFX-PheX-5_at
AFFX-PheX-M_at	AFFX-ThrX-5_at	AFFX-ThrX-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-DapX-5_at	AFFX-DapX-M_at	AFFX-LysX-5_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-M_at	203173_s_at
204890_s_at	AFFX-LysX-M_at	AFFX-r2-Bs-lys-5_at
213060_s_at	208010_s_at	204613_at
219607_s_at	220502_s_at	201744_s_at

Table 5: List of new spiked-in probe sets after initial analysis with RMA, MAS 5.0, GCRMA, and dChip.

To further examine the behavior of these candidate probe sets, we plotted their log 2 PM and MM intensities across all 14 experiments (Figure 1). If these probe sets truly recognize spike-in

transcripts, then there should be large differences in intensities for some experiments. We defined a "large difference" as larger than a fold change of 2 across all 14 experiments. For brevity, we give only four examples here: three probe sets that clearly behave as spike-ins and another that does not.

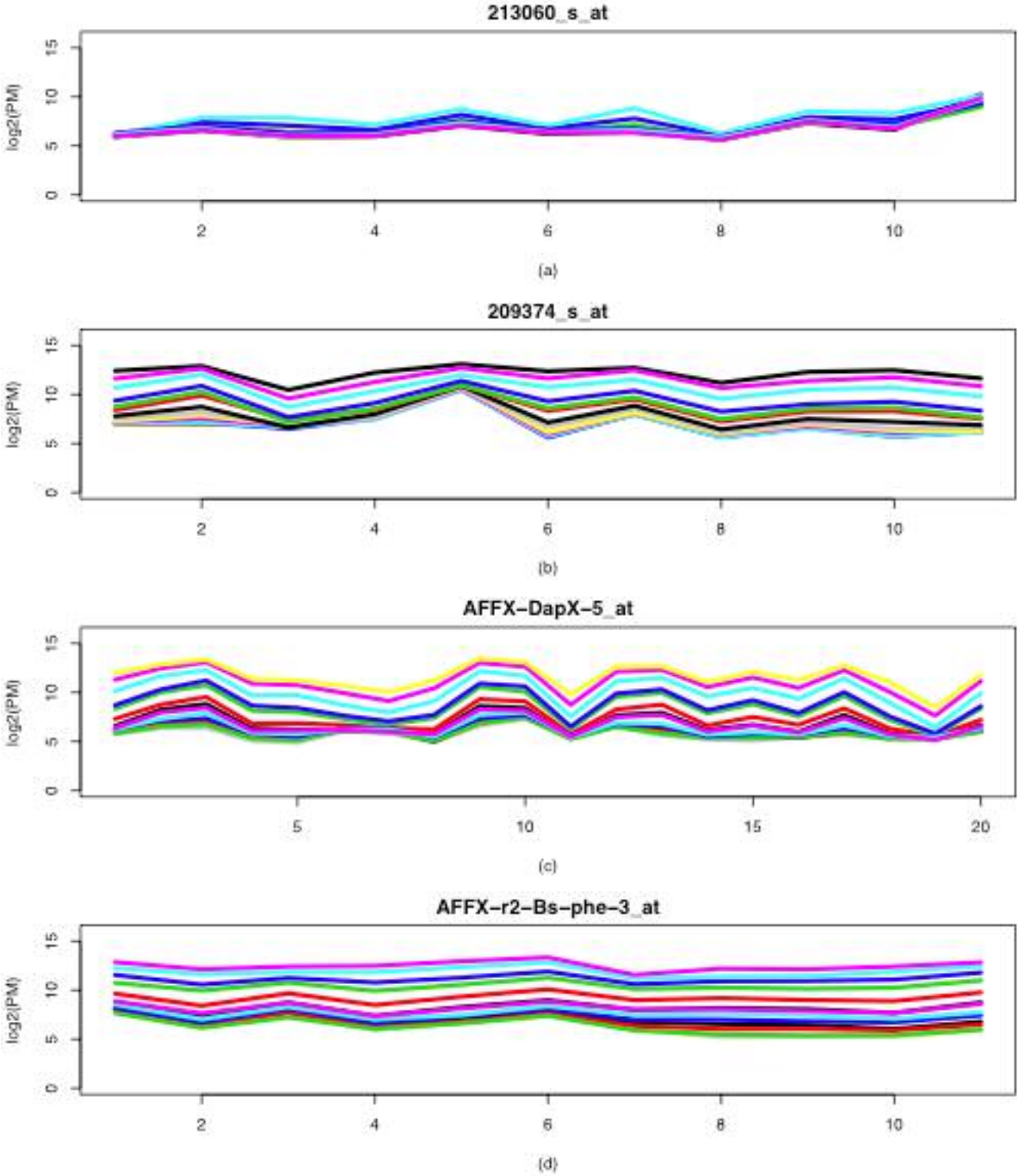


Figure 1: Probe set intensity plot for one non-spiked-in transcript and three spiked-in transcripts. Each line represents the average of three replicates for one experiment from the HG-U133 spike-in data. The x-axis gives the individual probes per probe set, and the y-axis marks the log base 2 intensity for each probe.

The three new probes sets that appear to recognize spike-in transcripts were randomly selected from different concentration groups. Each line in the plot represents the average of each probe intensity for the three replicates from a different experiment (from 1 to 14). The numbers on the x-axis are the probe numbers. There are 11 probe pairs for sets 213060_s_at, 209374_s_at and AFFX-r2-Bs-phe-3_at. AFFX-DapX-5_at has 20 probe pairs. Only the PM intensities are reproduced here, as the MM intensities behave similarly.

The probe set "213060_s_at", (Figure 1(a)), appeared as differentially expressed only for the MAS 5.0 method. Once its intensity levels were plotted, we see that it has almost the same log base 2 intensity across all arrays for each probe. The concentrations for some experiments are so close that they overlap, thus giving the illusion that only three or four experiments were plotted. In contrast, the probe sets "209374_s_at", "AFFX-DapX-5_at", and "AFFX-r2-Bs-phe-3_at" (plots b, c, and d) have very different intensity values for each probe across experiments, ranging from 5 to 12 on a log base 2 scale. Therefore, they follow similar patterns to the original spike-in genes, even though they were not originally designated as such. The average range (maximum value – minimum value) of the probe sets determined to be non-spiked-in was 1.3 with a standard deviation of 0.6, while the average range for the spiked-in probe sets was 5.9 with a standard deviation of 1.0 (on a log base 2 scale). Based on the graphical analysis and the average and standard deviation of the ranges, probe sets 201744_s_at, 220502_s_at, 219607_s_at, 204613_at, 213060_s_at, 204890_s_at, 203173_s_at, and 204891_s_at were eliminated, leaving twenty-two new spiked-in probe sets.

Determining Group Membership of Candidate Probe Sets: Appendix C gives a summary of the group membership for each of the new spike-in probe sets. We found that 15 of the twenty-two new spike-ins have patterns similar to the fourteenth group. Five of the remaining probe sets have patterns similar to the penultimate group, and the probes 209374_s_at and 208010_s_at are placed in the fifth and eighth groups, respectively. These placements are reasonable from a biological, as well as a statistical, perspective. For example, the proposed spike-in 209374_s_at is known to have the exact same probe sequence as the original spiked-in transcript 212827_at. Furthermore, probe set 208010_s_at shares 9 of its 11 probes with the known spike in 206060_at (Table 3). Therefore, the intensity levels of these pairs of transcripts should be highly correlated.

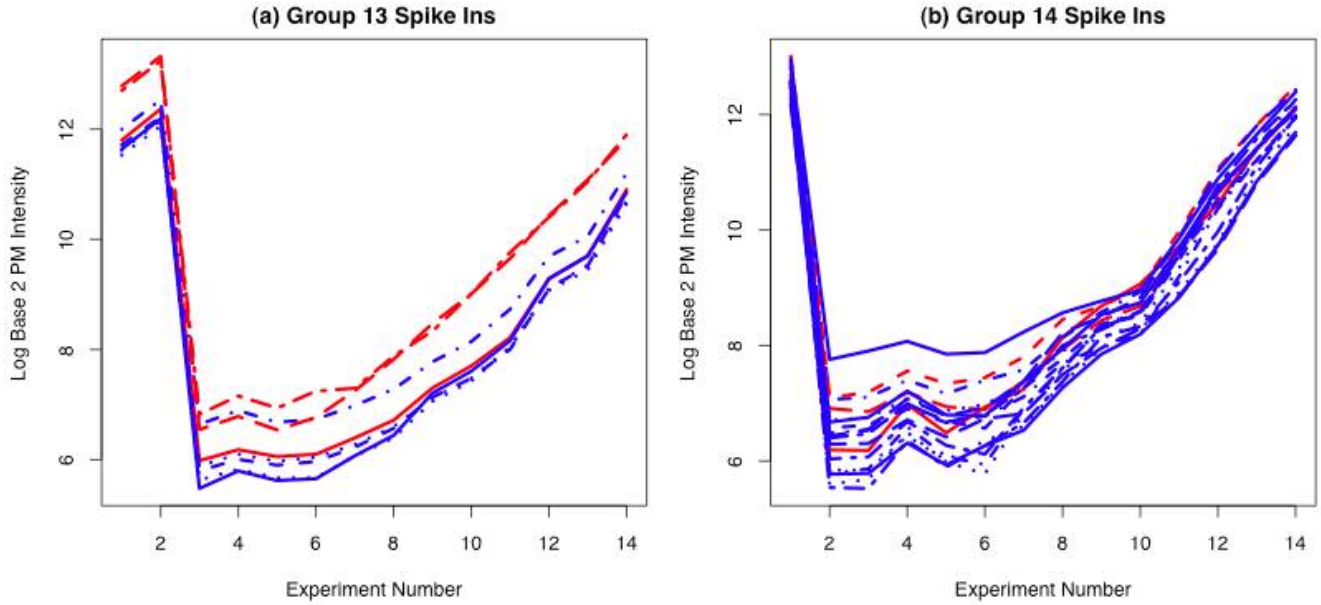


Figure 2: New (blue) and old (red) spiked-in clones from concentration group 13 and 14. It is clear that the new spike-ins follow roughly the same pattern as the old ones.

Figure 2 shows the average PM intensities of original and proposed spike-ins for concentration groups 13 (Figure 2a) and 14 (Figure 2b). It is clear that the new spiked-in genes follow the patterns within these groups. From this point forward, when we use the term "proposed spike-ins", we mean the set of the 42 original spiked-in probe sets plus the 22 new spike-ins discussed in this paper.

Sequence and Functional Similarity:
We have already explained the nearly identical sequence structure between 209374_s_at, 208010_s_at and originally spiked-in members of their respective groups. Further investigation of the function of the original and proposed spiked-in genes in concentration groups 13 and 14 indicate that the proposed spike-in probe sets are likely recognizing other sections of the same spike-in transcripts as the original spike-in probe sets.

For example, a BLAST search on the probe set AFFX-DapX-3_at returns a definition of Bacillus subtilis clone YAC15-6B ypiABF gene. However, BLAST searches on the probe sets AFFX-DapX-5_at and AFFX-DapX-M_at reveal the same annotation. The designations -3, -5, and –M indicate the section of the gene (3', 5' or middle, respectively) from which the sequences were extracted to create the probe set. It makes sense that these probe sets would behave in a similar fashion since the spike-in transcripts appear to be derived from full-length gene clones. The same is true of the probe sets AFFX-LysX-3/5/M_at, AFFX-ThrX-3/5/M_s_at and AFFX-PheX-3/5/M_at.

BLAST queries on the AFFX-r2-Bs-dap series of probe sets revealed that their structure and function are similar to the AFFX-DapX series. Since their sequence structure and functions are so close to originally spiked-in genes, these probe sets behave as spike-ins, also. Analogously, the AFFX-r2-Bs-lys, AFFX-r2-Bs-thr, and AFFX-r2-Bs-phe series behave as AFFX-LysX, AFFX-ThrX, and AFFX-PheX series, respectively.

Effect on Current Analysis Methods:

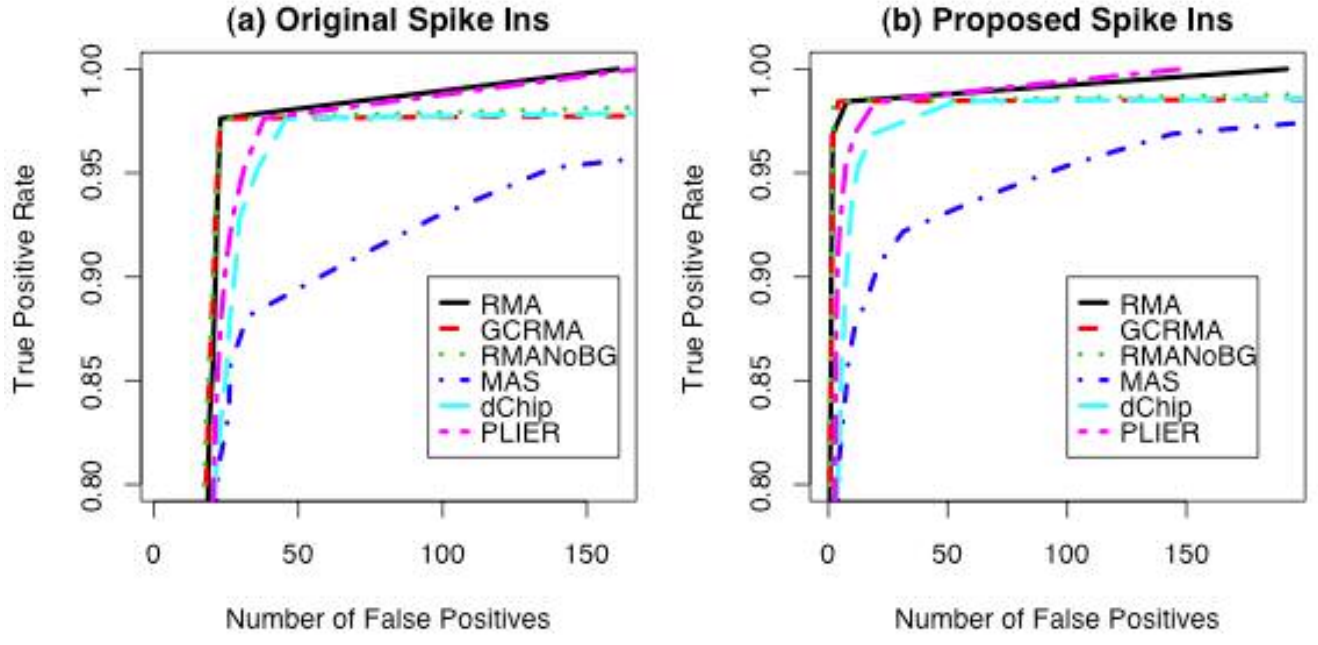


Figure 3: Average ROC curve for pairs of experiments for d = 5 for the original 42 spike-ins (a) and the proposed 64 spike-ins (b). The scale of the y-axis is from 0.80 to 1.0 in order to show differences. All methods perform better with the proposed spike-ins. There is not much practical difference among the methods when the fold changes are this large, except to say that MAS 5.0 seems to perform the worst of all the methods.

When new spike-in genes are discovered, it is important to examine their effect on the assessment of currently accepted analysis methods since they would likely be viewed as false positive results in any differential expression assessment when annotated as non-spike-ins. To accomplish this, we looked at average ROC curves of experiments with d=1 to 7 (Figure 3 and Appendix D) for analysis based on RMA, RMA-noBG, GCRMA, MAS, dChip, and PLIER.. For all ROC curves and AUC measurements that follow, we obtain results for both the 42 original and the 64 new spiked-in probes for each value of d.

Figure 3(a) shows the ROC curves for six preprocessing methods using the original 42 spiked-in probe sets when d = 5. In other words, most of the spiked-in transcripts are spiked-in at concentrations that differ by 5 on a log base 2 scale. Figure 3(b) shows the same six methods for all 64 spike-ins. For such a large fold change difference, any method ought to be quite accurate, and this is the case. Note that the scale of the y-axis is from 0.80 to 1.0 in order to make the differences easier to see. However, the difference in the performance of the methods using the original spike-ins versus the proposed spike-ins is striking. All methods perform markedly better when the proposed spike-ins are used. With the original spike-ins, the number of false positives at a true positive rate of 0.80 is approximately 20 for all methods. The number of false positives is close to zero for all methods when the proposed spike-ins are used. The story is very similar for all values of d (see Appendix D). Differences in the performance of all methods become more pronounced between the two sets of spike-ins as d gets larger.

Table 6 displays the average AUC for spike-in experiments which d from 1 to 7 using the original spiked-in probe sets. Table 7 displays the same information, except that the proposed spiked-in probe sets are used. We calculated the AUCs using 200 false positives for all six methods and both sets of spike-ins. The data are displayed in this manner so that one can see how the magnitude of the differences in intensity affects the results. AUCs for all methods using the proposed spike-ins tend to be larger, no matter the value of d. As would be expected, it is much easier for all methods to detect large differences than to detect the smaller ones.

d	RMA	RMA-NoBG	GCRMA	MAS 5.0	dChip	PLIER
1	0.746	0.733	0.715	0.085	0.620	0.363
2	0.785	0.744	0.786	0.277	0.752	0.741
3	0.882	0.874	0.835	0.613	0.824	0.865
4	0.914	0.921	0.881	0.798	0.880	0.903
5	0.929	0.922	0.925	0.873	0.909	0.926
6	0.935	0.939	0.948	0.903	0.932	0.937
7	0.934	0.942	0.949	0.910	0.933	0.937

Table 6: AUC for RMA, RMA-noBG, GCRMA, MAS 5.0, dChip, and PLIER for detection of spiked-in genes in the HG-U133A spikein experiment, where the original 42 spiked-in probe sets are used to calculate true and false positives. To calculate the AUCs, the number of false positives was set to 200.

d	RMA	RMA-NoBG	GCRMA	MAS 5.0	dChip	PLIER
1	0.738	0.734	0.643	0.060	0.600	0.365
2	0.831	0.812	0.720	0.307	0.709	0.752
3	0.904	0.908	0.747	0.561	0.811	0.883
4	0.964	0.964	0.855	0.837	0.913	0.951
5	0.990	0.983	0.982	0.939	0.971	0.985
6	0.998	0.996	0.998	0.968	0.989	0.994
7	0.999	0.999	0.999	0.978	0.992	0.996

Table 7: AUC for RMA, RMA-noBG, GCRMA, MAS 5.0, dChip, and PLIER for detection of spiked-in genes in the HG-U133A spikein experiment, where 64 spike ins (original 42 plus 22 proposed) are used. To calculate the AUCs, the number of false positives was set to 200.

In Tables 6 & 7, Figure 3, and Appendix D, we see that the use of the proposed spiked-in probe sets improve the overall performance of all methods in comparison with the use of the original spiked-in probe sets. However, the use of the new spiked-in probe sets does not substantially change the relative performance these methods in comparison with each other, except that GCRMA tends to perform comparably to RMA and RMA-noBG under the original spike-ins, but is inferior to both under the proposed spike-ins. In general, RMA and RMA-noBG perform better than the others with regard to specificity and sensitivity.

Discussion:

In this study, we proposed that 22 new probe sets that recognize spike-in transcripts be added to the current list of 42 spiked-in probe sets in HG-U133A experiment. The existence of new spike-ins has precedent, as the actual number of spike-in genes in the HG-U95Av2 experiment is debated. Affymetrix acknowledged that two of the probe sets contained poorly performing probes, which may affect their resulting expression values (9). Other researchers have discovered additional spiked-in probe sets (11). Our analysis of the HG-U95Av2 data did not reveal any further spiked-in probes than the ones already discovered.

None of the spiked-in probe sets in the HG-U95Av2 experiment were Affymetrix control transcripts. In the HG-U133A experiment, however, twelve of the original 42 spiked-in probe sets are controls (8 are artificial control sequences, four are prokaryotic controls). The probe sets initially described as recognizing the spiked-in bacterial controls were targeted at the 3' end of the genes. However, in the generation of the labeled RNA target mixture for hybridization, efficient polymerization would extend the labeled targets toward the 5' end of these relatively short bacterial transcripts. Thus, it is not surprising that probe sets that recognize the 5' and middle regions of these transcripts would also hybridize to the same spike-in RNAs as would probe sets targeted at the 3' end of the gene. This explains the inclusion of the probe sets AFFX-DapX-5_at, AFFX-DapX-M_at, AFFX-LysX-5_at, AFFX-LysX-M_at, AFFX-PheX-5_at, AFFX-PheX-M_at, AFFX-ThrX-5_at, and AFFX-ThrX-M_at, as new spike-ins. One can also see this in the placement of the probe sets into groups. The five probe sets in the penultimate group are all associated with the "dap" gene, which implies that they would have the same pattern as AFFX-dapX-3_at. Similarly, the fifteen new spike-in probe sets with the designation "lys", "phe" and "thr" all have similar patterns as the original spiked-in genes based on these genes. The same is true of the AFFX-r2-Bs series.

A graph of ROC curves shows that the relative standing of the currently accepted methods of analysis has not changed substantially with the addition of these new spike-in probe sets. MAS 5.0 seems to perform a bit worse than before, while PLIER seems to perform better. Some have argued that RMA-noBG performs best under most conditions (9). This may be true for the HG-U95Av2 data set, but in the analysis of the HG-U133A spike-in data set reported here, standard RMA seems to perform best. However, even though the relative standing of the methods is not affected by the inclusion of new spike-ins, the overall performance of all methods is affected. All methods perform better when the proposed spike-ins are included.

Even though the inclusion of these probe sets as recognizing spiked-in transcripts does not change the current status of the popular algorithms, use of the new spike-ins may have implications for methods not tested here, and for methods that have yet to be developed. Further, the placement of 15 of the new spiked-in genes in the last concentration group implies that the Latin Square is no longer balanced. The implications of this for current and future methods have yet to be worked out. This is also indicates that some probe sets act in clusters, and thus do not represent independent assessments when designing statistical tests for differentially expressed genes.

References

1. Lipshutz,R.J., Fodor,S.P.A., Gingeras,T.R. and Lockhart,D.J. (1999) High density synthetic oligonucleotide arrays *Nat. Genet.* **21**, 20-24.
2. Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., V,G.M., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Norton,H. and Brown,E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays *Nat. Biotechnol.* **14**, 1675-1680.
3. Affymetrix (2002) Statistical Algorithms Description Document
4. Li,C. and Wong,H.W. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection *Proceedings of the National Academy of Sciences* **98**, 31-36.
5. Li,C. and Wong,H.W. (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application *Genome Biology* **2**, research0032.1-0032.11.
6. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, Normalization, and summaries of high density oligonucleotide array probe level data *Biostatistics* **4**, 249-264.
7. Wu,Z., Irizarry,R.A., Gentleman,R., Martinez-Murillo,F. and Spencer,F. (2004) A model-based background adjustment for oligonucleotide expression arrays *Journal of the American Statistical Association* **99**, 909-917.
8. Affymetrix,I. (2005) Technical Note: Guide to probe logarithmic intensity error (PLIER) estimation
9. Cope,L.M., Irizarry,R.A., Jaffee,H., Wu,Z. and Speed,T.P. (2003) A benchmark for Affymetrix GeneChip Expression Measures *Bioinformatics* **20**, 323-331.
10. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B.M., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J., Hornik,K., Hothorn,T., Huber,W., S,I., Irizarry,R.A., Cheng,F.L., Maechler,M., AJ,R., Sawitizki,G., C,S., Smyth,G., Tierney,L., Yang,J.W.H. and Zhang,J. (2004) Bioconductor: Open software development for computational biology and bioinformatics *Genome Biology* **5**, R80.
11. Wolfinger,R. and Chu,T. (2003) Who are those strangers in the Latin Square? *Methods of Microarray Data Analysis III*

Appendix A: Spiked-in Probe Sets for the HG-U133 A Spike-In Experiment and the concentrations (in pM) of their associated transcripts.

Probe ID	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
203508_at 204563_at 204513_s_at	0	0.125	0.25	0.5	1	2	4	8	16	32	64	128	256	512
204205_at 204959_at 207655_s_at	0.125	0.25	0.5	1	2	4	8	16	32	64	128	256	512	0
204836_at 205291_at 209795_at	0.25	0.5	1	2	4	8	16	32	64	128	256	512	0	0.125
207777_s_at 204912_at 205569_at	0.5	1	2	4	8	16	32	64	128	256	512	0	0.125	0.25
207160_at 205692_s_at 212827_at	1	2	4	8	16	32	64	128	256	512	0	0.125	0.25	0.5
209606_at 205267_at 204417_at	2	4	8	16	32	64	128	256	512	0	0.125	0.25	0.5	1
205398_s_at 209734_at 209354_at	4	8	16	32	64	128	256	512	0	0.125	0.25	0.5	1	2
206060_s_at 205790_at 200665_s_at	8	16	32	64	128	256	512	0	0.125	0.25	0.5	1	2	4
207641_at 207540_s_at 204430_s_at	16	32	64	128	256	512	0	0.125	0.25	0.5	1	2	4	8
203471_s_at 204951_at 207968_s_at	32	64	128	256	512	0	0.125	0.25	0.5	1	2	4	8	16
AFFX-r2-TagA_at AFFX-r2-TagB_at AFFX-r2-TagC_at	64	128	256	512	0	0.125	0.25	0.5	1	2	4	8	16	32
AFFX-r2-TagD_at AFFX-r2-TagE_at AFFX-r2-TagF_at	128	256	512	0	0.125	0.25	0.5	1	2	4	8	16	32	64
AFFX-r2-TagG_at AFFX-r2-TagH_at AFFX-DapX-3_at	256	512	0	0.125	0.25	0.5	1	2	4	8	162	32	64	128
AFFX-LysX-3_at AFFX-PheX-3_at AFFX-ThrX-3_at	512	0	0.125	0.25	0.5	1	2	4	8	16	32	64	128	256

Appendix B

Top twenty absolute fold changes between experiments 1 and 2 using MAS 5.0 and GCRMA analyses. Absolute fold change is calculated by taking the absolute value of the log base 2 quotient of the sum of intensities from each experiment. The Spike In column gives the original spike-in designation given at the time the data were released.

Top 20 Differentially Expressed Genes Found Using MAS 5.0								
Probe Set ID	Exp1R1	Exp1R2	Exp1R3	Exp2R1	Exp2R2	Exp2R3	Abs FC	Spike In
AFFX-PheX-M_at	4868.01	4596.08	5086.07	1.17	1.09	0.93	12.16	N
AFFX-PheX-5_at	4906.05	4501.03	5167.72	1.13	1.31	1.42	11.88	N
AFFX-ThrX-3_at	5514.58	5284.16	5768.56	2.44	2.66	2.22	11.14	Y
AFFX-r2-Bs-thr-3_s_at	5715.36	5385.35	6078.51	2.10	4.24	2.11	10.99	N
AFFX-LysX-5_at	4453.25	3976.38	4600.66	1.03	1.22	4.73	10.87	N
AFFX-r2-Bs-phe-5_at	5442.98	4813.04	5592.10	4.13	0.68	3.94	10.82	N
AFFX-ThrX-5_at	4306.27	3962.28	4606.67	2.58	2.17	2.72	10.75	N
AFFX-r2-Bs-phe-M_at	5168.37	4568.73	5129.72	3.37	1.17	7.03	10.33	N
AFFX-r2-Bs-lys-5_at	4781.86	4335.74	4952.05	7.23	0.44	5.04	10.11	N
AFFX-LysX-3_at	5193.96	4862.86	5264.42	2.59	6.67	6.85	9.89	Y
AFFX-ThrX-M_at	4806.26	4304.89	4977.08	2.15	10.68	5.65	9.57	N
AFFX-r2-Bs-lys-M_at	4509.13	3882.26	4820.12	8.88	6.92	3.55	9.42	N
AFFX-LysX-M_at	5241.89	5157.68	5440.15	10.94	2.41	14.74	9.14	N
AFFX-r2-Bs-thr-M_s_at	6241.71	6306.88	6540.25	16.44	11.79	6.79	9.09	N
AFFX-PheX-3_at	4872.27	5069.32	4996.06	13.26	13.99	16.55	8.41	Y
AFFX-r2-Bs-phe-3_at	5418.63	4957.40	5482.87	19.71	18.52	20.33	8.08	N
AFFX-r2-Bs-lys-3_at	3944.22	3323.01	4127.48	21.18	13.65	13.06	7.89	N
AFFX-r2-Bs-thr-5_s_at	5112.14	5243.97	5371.47	30.09	24.16	28.54	7.57	N
209638_x_at	8.96	12.52	8.61	0.85	1.07	1.02	3.36	N

Top 20 Differentially Expressed Genes Found Using GCRMA								
Probe Set ID	Exp1R1	Exp1R2	Exp1R3	Exp2R1	Exp2R2.	Exp2R3	Abs FC	Spike In
AFFX-r2-Bs-thr-3_s_at	12673.829	13342.300	13448.834	4.043	3.961	3.992	11.684	N
AFFX-r2-Bs-thr-M_s_at	11242.576	12270.045	11818.744	4.139	4.061	4.094	11.489	N
AFFX-r2-Bs-phe-M_at	21249.008	20158.612	21708.436	8.040	7.709	7.768	11.390	N
AFFX-ThrX-3_at	11987.608	12050.150	12059.392	4.753	4.364	4.599	11.362	Y
AFFX-r2-Bs-thr-5_s_at	11227.866	11044.406	11651.056	4.577	4.543	4.522	11.280	N
AFFX-PheX-M_at	11588.552	11997.882	11573.435	4.857	4.728	4.762	11.259	N
AFFX-PheX-5_at	15260.341	14124.818	15084.659	6.253	6.102	6.071	11.237	N
AFFX-LysX-M_at	12629.040	13769.093	12322.634	6.004	5.353	5.438	11.171	N
AFFX-r2-Bs-lys-5_at	20340.893	20542.154	21368.915	9.628	8.759	9.108	11.145	N
AFFX-r2-Bs-phe-5_at	22049.740	19869.950	20533.954	9.656	9.006	9.092	11.136	N
AFFX-ThrX-M_at	14595.730	15240.599	15342.225	6.902	6.789	6.871	11.101	N
AFFX-PheX-3_at	12329.760	13513.873	12821.668	6.291	6.162	6.160	11.021	N
AFFX-LysX-5_at	15309.709	14827.204	14737.202	7.626	7.250	7.417	10.975	N
AFFX-LysX-3_at	13436.706	13385.509	12866.751	6.733	6.570	6.478	10.970	Y
AFFX-ThrX-5_at	10847.603	10566.455	11117.822	5.576	5.535	5.566	10.930	N
AFFX-r2-Bs-phe-3_at	14293.626	13037.730	14119.617	7.454	6.875	6.940	10.928	N
AFFX-r2-Bs-lys-M_at	15385.833	14150.520	15378.243	8.242	7.577	7.292	10.924	N
AFFX-r2-Bs-lys-3_at	13560.790	11299.483	14271.304	9.354	8.220	8.422	10.556	N
205267_at	119.176	103.815	107.496	265.603	301.105	306.834	1.402	Y

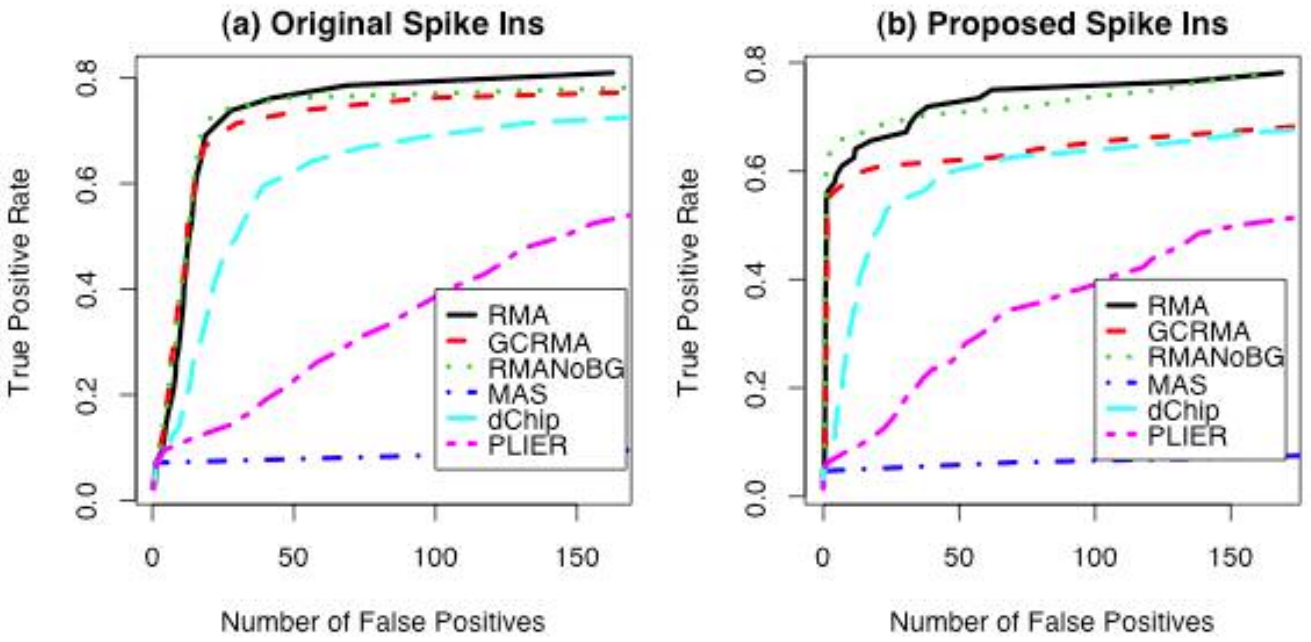
Appendix C

Table of concentration groups for new spike-in probe sets.

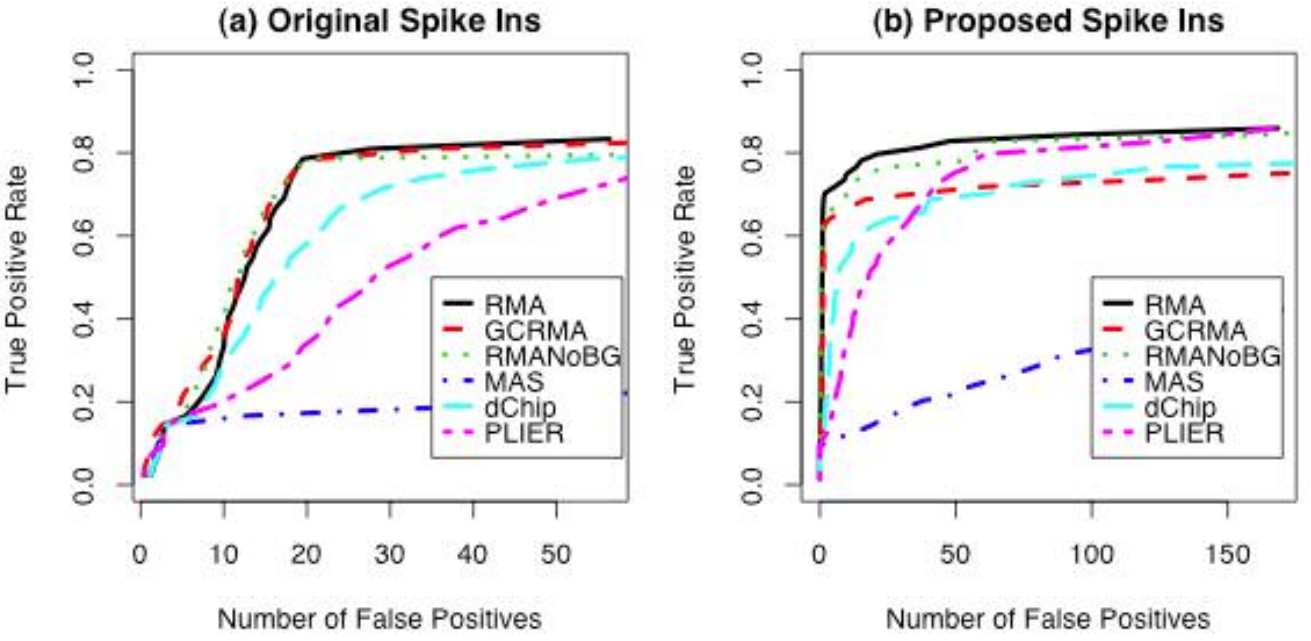
Concentration Group	Current Members	Proposed Members	Comments
5	207160_at 205692_s_at 212827_at	209374_s_at	Shares 11 of 11 probes with 212827_at
8	206060_s_at 205790_at 200665_s_at	208010_s_at	Shares 9 of 11 probes with 206060_s_at
13	AFFX-r2-TagG_at AFFX-r2-TagH_at AFFX-DapX-3_at	AFFX-DapX-5_at AFFX-DapX-M_at AFFX-r2-Bs-dap-3_at AFFX-r2-Bs-dap-5_at AFFX-r2-Bs-dap-M_at	All proposed spike-ins are associated Bacillus subtilis clone YAC15-6B ypiABF genes.
14	AFFX-LysX-3_at AFFX-PheX-3_at AFFX-ThrX-3_at	AFFX-PheX-5_at AFFX-PheX-M_at AFFX-ThrX-5_at AFFX-ThrX-M_at AFFX-LysX-5_at AFFX-LysX-M_at AFFX-r2-Bs-lys-5_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-Bs-lys-3_at AFFX-r2-Bs-lys-M_at	The probes with similar designations (lys, phe, or thr) have similar functions. For example, AFFX-LysX-*_at (where * = 3, 5, or M) and AFFX-r2-Bs-lys-*_at are both associated with Bacillus subtilis lys gene for diaminopimelate decarboxylase.

Appendix D

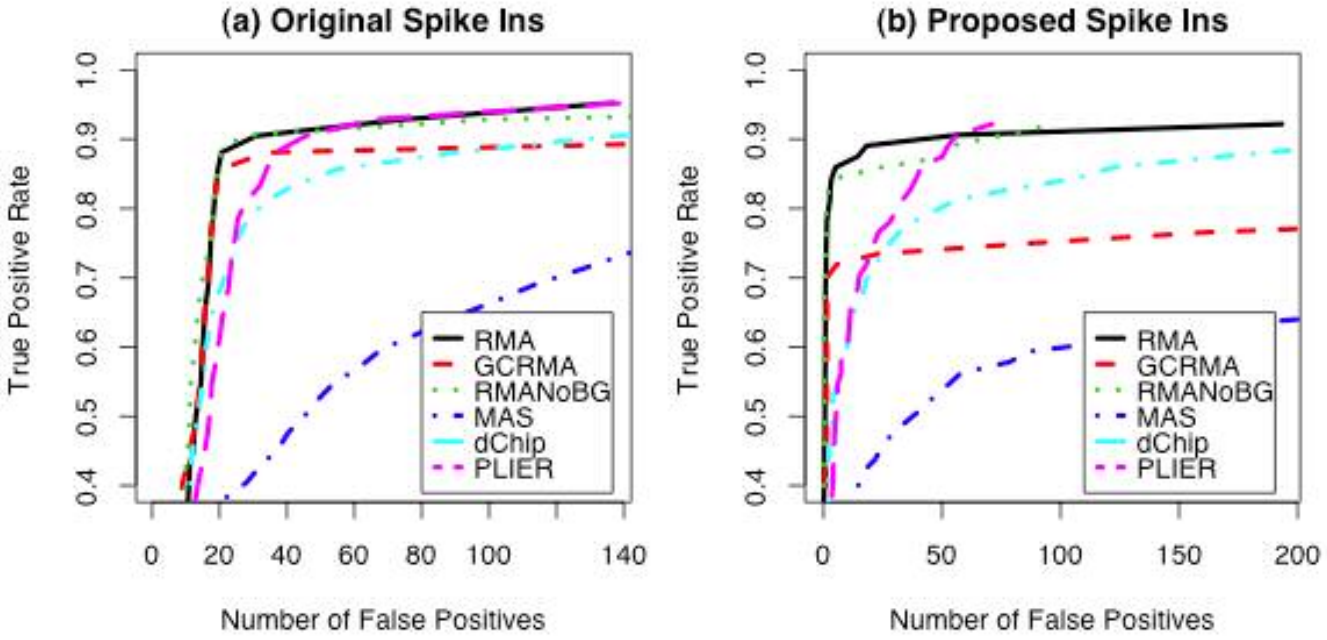
ROC curves for experiments according to d, the number of permutations of the Latin Square design, which separate the experiments.



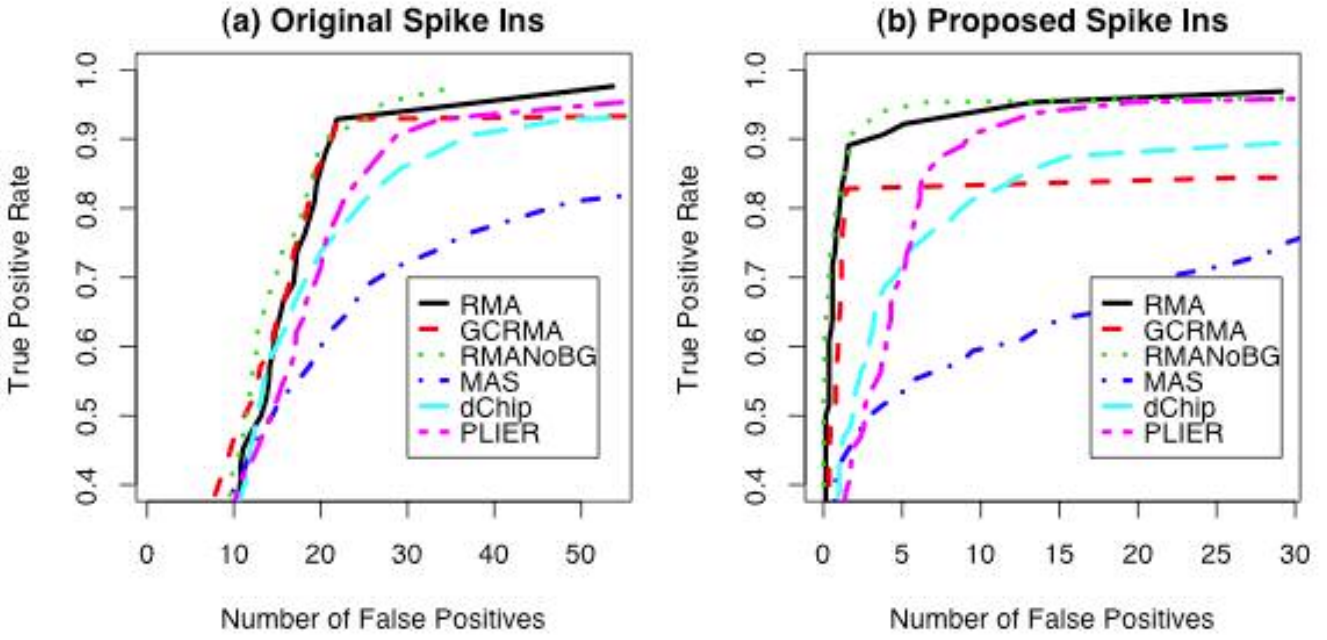
Supplemental Figure 1: Average ROC curve for pairs of experiments for d = 1 for the original 42 spike-ins (a) and the proposed 64 spike-ins (b). For these experiments, the log fold changes between spiked-in transcripts are quite small, which implies that detection of differentially expressed genes will require more sensitive methods. Performance is somewhat improved when all 64 spike-ins are used, except for MAS and PLIER, which remain unspecific. GCRMA performs slightly worse than RMA and RMA-noBG when the proposed spike-ins are used.



Supplemental Figure 2: Average ROC curve for pairs of experiments for d = 2 for the original 42 spike-ins (a) and the proposed 64 spike-ins (b). All methods perform better when using the proposed spike-ins.



Supplemental Figure 3: Average ROC curve for pairs of experiments for d = 3 for the original 42 spike-ins (a) and the proposed 64 spike-ins (b). The scale for the y-axis goes from 0.4 to 1.0 so that differences among the methods are visible. The scale of the y-axis also explains why the number of false positives for the original spike-ins (a) begins at approximately 15. This indicates that all of the methods are performing better with the proposed spike-ins (b). The best methods (RMA and RMA-noBG) are able to find 90% of true positives with between 40 and 50 false positives when the proposed spike-ins are used. Ninety percent of true positive are found with PLIER, also, but with the number of false positives close to 60.



Supplemental Figure 4: Average ROC curve for pairs of experiments for d = 4 for the original 42 spike-ins (a) and the proposed 64 spike-ins (b). The scale for the y-axis goes from 0.4 to 1.0 so that differences among the methods are visible. The story for these plots is much the same as in previous ones: all methods perform better with the prposed 64 spike-ins (b). The number of false positives at a true positive rate of 0.4 using the original spike-ins is close to 10, while, for the proposed spike-ins, the number of false positives for the same true positive rate is nearly 0 for all methods. GCRMA performs worse than RMA and RMA –noBG in plot (b), but performs equally well in plot (a).