Research article

Reproducibility of oligonucleotide arrays using small samples

Jeanette N McClintick[1,2], Ronald E Jerome[1,3], Charles R Nicholson[1,3], David W Crabb[3,4,5] and Howard J Edenberg*[1,2,3]

Address: [1]Center for Medical Genomics, Indiana University School of Medicine, Indianapolis, Indiana, USA, [2]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana, USA, [3]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana, USA, [4]Department of Medicine, Indiana University School of Medicine, Indianapolis, Indiana, USA and [5]VA Medical Center, Indianapolis, Indiana, USA

Email: Jeanette N McClintick - jnmcclin@iupui.edu; Ronald E Jerome - rjerome@iupui.edu; Charles R Nicholson - charnich@indiana.edu; David W Crabb - dcrabb@iupui.edu; Howard J Edenberg* - edenberg@iupui.edu

* Corresponding author

Abstract

Background: Low RNA yields from small tissue samples can limit the use of oligonucleotide microarrays (Affymetrix GeneChips®). Methods using less cRNA for hybridization or amplifying the cRNA have been reported to reduce the number of transcripts detected, but the effect on realistic experiments designed to detect biological differences has not been analyzed. We systematically explore the effects of using different starting amounts of RNA on the ability to detect differential gene expression.

Results: The standard Affymetrix protocol can be used starting with only 2 micrograms of total RNA, with results equivalent to the recommended 10 micrograms. Biological variability is much greater than the technical variability introduced by this change. A simple amplification protocol described here can be used for samples as small as 0.1 micrograms of total RNA. This amplification protocol allows detection of a substantial fraction of the significant differences found using the standard protocol, despite an increase in variability and the 5' truncation of the transcripts, which prevents detection of a subset of genes.

Conclusions: Biological differences in a typical experiment are much greater than differences resulting from technical manipulations in labeling and hybridization. The standard protocol works well with 2 micrograms of RNA, and with minor modifications could allow the use of samples as small as 1 micrograms. For smaller amounts of starting material, down to 0.1 micrograms RNA, differential gene expression can still be detected using the single cycle amplification protocol. Comparisons of groups of four arrays detect many more significant differences than comparisons of three arrays.

Background

The ability to measure the expression of thousands of genes at once using microarrays has opened new areas of research, including global examination of the effects of perturbations on cells or animals and the classification of tumors by their pattern of gene expression. Microarrays using cDNAs [1,2] and oligonucleotides [3–5] have both proven valuable.

Commercially available oligonucleotide microarrays provide a standardized tool that allows assay of thousands of mRNAs at one time. Affymetrix GeneChips® contain pairs

of 25-nucleotide sequences (probe pairs) synthesized on silica wafers; one of each pair exactly matches the sequence of interest and the other contains a single mismatching nucleotide in the center [6,7]. A single sequence is queried by a group of 8 to 16 probe pairs that constitute a probe set. RNA from the sample is converted to double-stranded cDNA and then labeled by *in vitro* transcription with biotinylated nucleotides. The biotinylated cRNA is hybridized to the GeneChip®, unhybridized material is washed off, and the signal is detected using fluorescein-labeled Streptavidin [8].

The standard Affymetrix protocol [8] uses as starting material 10 μg of total RNA, from which biotinylated cRNA is synthesized. This can limit the use of this system for small samples from biopsies, laser capture microdissection or tissues from model organisms such as mice. Mahadevappa and Warrington [9] examined the effect of using less biotinylated cRNA in the hybridization. Hybridization reactions that contained the recommended 10 μg of cRNA (from human endometrium adenocarcinoma cells) detected 35% of the 1779 transcripts on the GeneChip®. Reducing the amount of cRNA in the hybridization to 5 μg reduced the fraction detected to 30%, and further reducing the cRNA to 2.5 μg allowed detection of only 27% of the sequences [9]. Ohyama et al. [10] tested a modified protocol for synthesizing biotinylated cRNA from very small amounts of starting material. Total RNA from laser capture microdissected human oral cancer tissues was converted into cDNA and transcribed *in vitro*; the resulting cRNA was converted into cDNA and transcribed *in vitro* a second time to generate more cRNA; this cRNA was again converted into cDNA and biotinylated cRNA was synthesized by a third *in vitro* transcription. This procedure produced 10 μg of biotinylated cRNA from 0.1 μg of starting total RNA. Hybridization with 10 μg of biotinylated cRNA generated by this amplification protocol allowed detection of 30% of the 7000 transcripts on the HuGeneFL GeneChip® [10]. In contrast, hybridization with 10 μg of biotinylated cRNA generated by the standard protocol from total RNA extracted from similar tissues resulted in detection of 35% of the transcripts being detected [10].

Rather than focusing upon the number of transcripts detected, the real test of a microarray protocol is the extent to which it allows differences in expression levels to be reliably detected. The biological variability inherent in most experimental models, including both genetic and environmental differences between animals or even replicate cell cultures, limits the detection of such differences. Additional variability in the treatment and handling of the models and in the RNA extraction typically occur outside the microarray laboratory, and can be reduced (but not eliminated) by careful experimental design. There is the

potential for introducing additional technical variability during the synthesis of biotinylated cRNA. In evaluating a new protocol or comparing existing protocols, measures such as the yield of cRNA or the fraction of probe sets detected can be useful, but the key measure is the extent to which differences in gene expression can be detected in a realistic experiment.

We have systematically explored the use of smaller amounts of starting material (total RNA) in a model experiment that retains the individual-to-individual biological variation of a real experiment. This allowed us to compare technical variability to the biological variability in a typical experiment. We started with total RNA from individual rats exposed to two different nutritional regimens and used serial dilutions of the RNA to simulate experimental systems that provide smaller quantities of total RNA. We used the Rat Genome RGU34A GeneChip® for all of the experiments. Our first goal was to determine a reasonable lower bound for total RNA that could successfully be used in the standard protocol. Second, we wanted to test a modified version of the Ohyama amplification procedure [10] that we thought would be faster, simpler and less likely to skew results due to truncation of the labeled cRNA. We examined the variability and the ability to detect significant changes between animals fed the 2 different dietary regimens when different amounts of starting material and different protocols were used.

Results
Yields and quality of cRNA

The standard protocol uses 10 μg of total RNA to produce biotinylated cRNA [8]. In our experiment, the average yield of biotinylated cRNA was 97 μg (± 41 μg; mean ± standard deviation) when we started with 10 μg of RNA and used one-half of the double stranded cDNA product in the *in vitro* transcription reaction. Starting with 2 μg of total RNA, we obtained an average of 50 μg (± 20 μg; standard deviation) of biotinylated cRNA after using all of the double stranded cDNA product in the *in vitro* transcription reaction. The 1 μg pooled samples produced an average of 5.6 μg of biotinylated cRNA (yields ranged from 2.4 μg to 7.7 μg). No difference in yield was observed between the samples prepared with the ENZO and Epicentre T7 polymerases. This lack of difference led us to select the Epicentre Ampliscribe™ T7 High Yield Kit for the extra *in vitro* transcription step to decrease expense.

The standard Affymetrix protocol uses 15 μg of biotinylated cRNA to make a 300 μl hybridization cocktail, of which 200 μl is injected into the chip for hybridization. The yield of cRNA from both 10 μg and 2 μg of starting RNA (above) were more than sufficient for this. Because yields of cRNA starting from 1 μg total RNA were too low, we decided to use an additional amplification step for RNA

Table 1: Quality control measures as a function of starting material.

	Amount of starting RNA			
	10 μg	2 μg	0.5 μg	0.1 μg
% Present	38.6%	37.8%	30.2%	24.5%
Scaling Factor	14.1	14.3	33.4	89.0
3'/5' GAPDH	1.7	1.6	6.1	5.5
3'/5' β-actin	1.5	1.5	9.2	8.2

Data are averages of all arrays in each RNA sample size group. 3'/5' signal ratios are given for GAPDH and β-actin.

samples less than 1 μg. Our modified protocol uses the initial *in vitro* transcribed cRNA as starting material for a second round of cDNA synthesis and *in vitro* transcription (Methods). The cRNA yields from the 0.5 μg samples with our protocol averaged 32 μg (± 13 μg), more than enough for the standard hybridizations. The cRNA yields from the 0.1 μg samples using the same protocol averaged 10 μg with a range from 5 to 21 μg, so some samples had too little to prepare a hybridization mixture at the same concentration. Therefore, the 0.1 μg samples were hybridized to GeneChips® using 5 μg to 7.5 μg of cRNA to assess results using these limited amounts.

Aliquots of the biotinylated cRNA samples were analyzed by agarose gel electrophoresis to check the quality and length. The cRNA for both the 10 μg and 2 μg samples ranged from 200 to over 2,000 bases (before fragmentation). The cRNA from the 0.5 μg sample prepared by the amplification protocol ranged from 200 to 850 bases, a considerable decrease in maximum length. The 0.1 μg samples were too faint to judge their size range.

Changes in sensitivity as measured by detection of probe sets

For a particular tissue or cell type, the percent present and the scaling factors should be similar among all arrays in the same experiment in the absence of variability introduced by preparation, labeling, and handling of the individual samples. Because we created groups of samples diluted from the same individual RNA preparations, any group differences reflect differences in labeling and handling of the samples. We used the percent of probe sets called present and the scaling factor (see Methods) for an initial comparison among the groups (Table 1). The 2 μg samples were essentially equivalent to the 10 μg samples by these measures (Table 1). For the amplified samples (0.5 μg and 0.1 μg starting material), the percent present was decreased and the scaling factor increased compared to the non-amplified samples. The percent present dropped to 30.2% in the 0.5 μg group (78% of that detected in the 10 μg group) and even further, to 24.5%, for the

0.1 μg samples. Two other quality control measures, noise and background, were similar across all 30 GeneChips® (data not shown).

If the decrease in starting amount of RNA or the differences in protocol had no effect on the outcome, the signals from all of the reduced RNA sample size groups would be distributed similarly to the 10 μg group. SAS was used to analyze the signals from each RNA sample size group. Probe sets were separated by detection call (absent, present and marginal) and analysis was performed separately for each group (only 2% of the probe sets were called marginal; these were omitted from Table 2). Signals for the 2 μg samples are distributed similarly to the 10 μg samples (Table 2). For the amplified samples the range of signals is increased and the distribution is shifted toward higher signal values. The decrease in percent of probe sets called present on the GeneChips® from the amplified samples has the effect of lowering the average signal on the chip, requiring a higher scaling factor (Table 1). This results in an inflation of the signal values for all of the probe sets on these arrays.

To examine the effects of starting with smaller amounts of RNA on the variability in detection of probe sets, we examined the number of probe sets that changed from present to absent or from absent to present when comparing the 10 μg sample to the smaller RNA samples from the same animal (Table 3). The average number of probe sets called present on the 10 μg chip and absent on the 2 μg chip (P_{10} to A) and called absent on the 10 μg chip and present on the 2 μg chip (A_{10} to P) changes were similar. The bulk of these changes were for probe sets with lower levels of expression (Table 3). This distribution is consistent with the greater variability seen in probe sets with low signals (see below); there is no significant loss of low-level transcripts in the 2 μg samples.

Because the amplified samples (from 0.5 μg and 0.1 μg starting material) have a decrease in the percent of probe sets called present (Table 1), the number of probe sets

Table 2: Expression levels for probe sets called absent and present by MAS5.

	Absent				Present				Maximum Signal
	25%	50%	75%	90%	25%	50%	75%	90%	
10 µg	25	69	169	328	446	1047	3220	10497	117836
2 µg	27	71	170	337	443	1052	3272	10769	123626
0.5 µg	34	94	223	438	628	1406	4315	14321	234230
0.1 µg	48	134	301	521	888	1877	5582	18228	439116

For each starting amount of RNA, the signals corresponding to the 25th percentile (25%), 50th percentile (50%), 75th percentile (75%) and 90th percentile (90%) are shown, along with the maximum signal.

Table 3: Variability in the detection of probe sets.

	Changes in Detection[1]					
	<300[2]		300–600[2]		>600[2]	
	P_{10} to A[3]	A_{10} to P[4]	P_{10} to A[3]	A_{10} to P[4]	P_{10} to A[3]	A_{10} to P[4]
2.0 µg	104 (58%)	131 (74%)	51 (28%)	35 (20%)	25 (14%)	1 2 (7%)
0.5 µg	241 (28%)	128 (74%)	245 (29%)	32 (18%)	365 (43%)	1 4 (8%)
0.1 µg	346 (28%)	52 (68%)	355 (29%)	15 (19%)	523 (43%)	10 (13%)

[1]Changes in Dectection are the average number of Present to Absent and Absent to Present changes per chip when comparing the 10 µg chip to the reduced RNA sample chip from the same animal. [2]Number of changes found among those probe sets that had a signal less than 300, 300 to 600 or greater than 600 (as noted) in the 10 µg chip. [3]Present in 10 µg chip, absent in reduced RNA sample chip from the same animal. [4]Absent in 10 µg chip, present in reduced RNA sample chip from the same animal.

called present in the 10 µg samples and absent in the amplified sample from the same animal must be greater than the number called absent in the 10 µg sample and present in the amplified samples (Table 3). Loss of signal is expected in low-level transcripts for the amplified samples because of the decrease in starting material, but probe sets present in the 10 µg and absent in the amplified samples are not confined to those probe sets with low signals. Forty-three percent of the probes not detected in the amplified samples have a signal greater than 600 in the 10 µg sample, and 5% have a signal of at least 3200. In comparison, for the 2 µg samples, only 14% of the probe sets present in the 10 µg sample and absent in the 2 µg sample had signals greater than 600, and none had signals over 3200.

The probe sets changing from absent in the 10 µg samples to present in the amplified samples (Table 3) were mostly those with low signals in the 10 µg samples, reflecting the greater noise found in probe sets with low signals (Figure 1).

Biological and technical variability

The Affymetrix MAS5 comparison analysis tool was used to compare expression levels between pairs of Gene-Chips®. This analysis directly compares two arrays at each individual probe pair rather than merely comparing the signal computed from all of the probe pairs for a probe set [11]. Comparisons among the 10 µg samples from different animals within each of the treatment groups showed that the biological (between animal) plus random technical variability is considerable (Table 4, 10 *vs.* 10). There was an average of over 700 apparent differences in expression level between pairs of GeneChips® within a single treatment group, 12% of which were 2-fold or greater. The number of increases and decreases were comparable, suggesting random rather than systematic changes. Increases and decreases were randomly distributed across probe sets with different levels of expression. This can be visualized as a scatter plot comparing two different 10 µg samples both from the vitamin-deficient group (Figure 1a and 1b); results are similar for a pair of control samples. There is noticeable scatter from the expected diagonal, and the scatter is greatly exacerbated at low signal levels. This shows variability between animals within a single
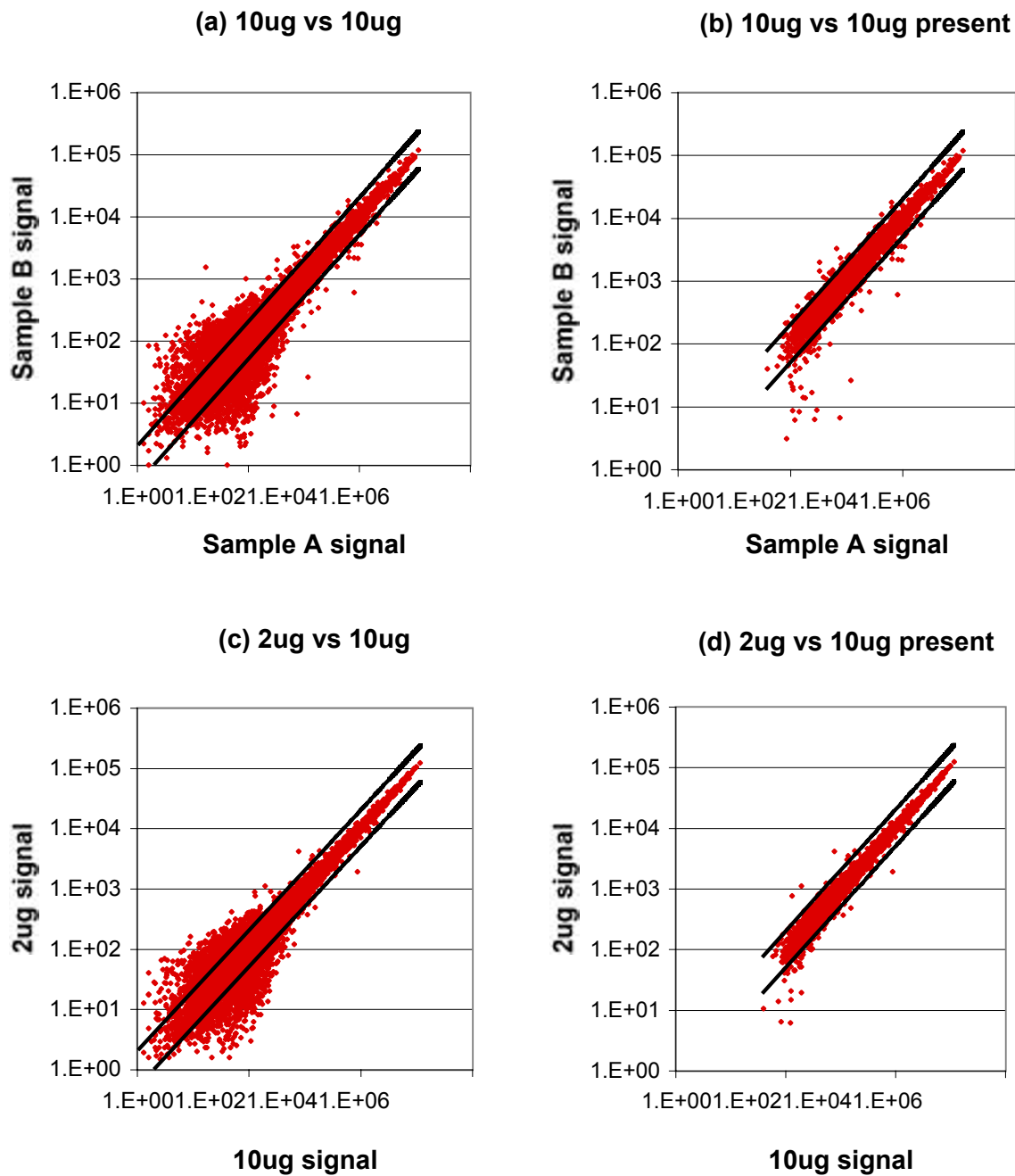
(a) 10ug vs 10ug

(b) 10ug vs 10ug present

(c) 2ug vs 10ug

(d) 2ug vs 10ug present

Figure 1
Log scale scatter plots of samples labeled by the standard protocol. Lines indicate a 2-fold difference. (a) Two 10 μg treated samples, all probe sets shown. (b) Same as figure 1a, but limited to the probe sets called present in Sample A (x-axis). (c) 2 μg vs. 10 μg from the same animal (Sample A in figures 1a and 1b). (d) Same as figure 1c, but limited to probe sets called present in the 10 μg sample.

Table 4: Changes in comparisons between animals within treatment.

	Between animal comparisons					
	Increases		Decreases		ALL	
Samples	Total	≥ 2-fold	Total	≥ 2-fold	Total	≥ 2-fold
10 *vs.* 10	335	31	374	51	709	82
0.5 *vs.* 0.5	426	143	506	144	932	287

The average number of changes detected by MAS5 per comparison between 2 GeneChips® (among probe sets called present in the baseline sample). Comparisons were made between animals in the same treatment group for both the 10 μg and 0.5 μg sample size groups to measure technical plus biological variability. "Total" is the number of changes regardless of magnitude; "≥ 2-fold" is the number of probe sets with changes of 2-fold or more. "Marginal Increases" and "Marginal Decreases" are included.

Table 5: Changes in comparisons between RNA sample size groups.

	Between sample size groups					
	Increases		Decreases		ALL	
Samples	Total	≥ 2-fold	Total	≥ 2-fold	Total	≥ 2-fold
2 *vs.* 10	136	3	134	7	270	10
0.5 *vs.* 10	528	436	1043	654	1571	1090
0.1 *vs.* 10	255	252	1119	647	1374	899
0.1 *vs.* 0.5	103	42	593	69	696	111

The average number of changes detected by MAS5 per comparison between 2 GeneChips® (among probe sets called present in the baseline sample). Comparisons were made between samples from the same animal in two different RNA sample size groups to measure technical variation. "Total" is the number of changes regardless of magnitude; "≥ 2-fold" is the number of probe sets with changes of 2-fold or more. "Marginal Increases" and "Marginal Decreases" are included.

treatment group (plus the technical variability in handling two samples, even with the same protocol). Figure 1b is the same pair of GeneChips® but is restricted to probe sets that were called present in the first sample (Sample A, x-axis). Analysis of the signals for all of the 10 μg arrays shows that 59% of the probe sets have signals below 300, and that 89% of these are called absent. Removing the probe sets called absent from further analyses removes most of the variability seen in this low signal range. The data in Table 4 were limited to probe sets that were called present in the baseline sample for each comparison to avoid the noise of low signal absent and marginal calls (cf. Figure 1).

GeneChips® from the lower RNA sample size groups were compared to the 10 μg chip from the same animal; these comparisons represent technical variation only, because the RNAs were dilutions from the same RNA preparations. To compare variability introduced by the same labeling protocol with different amounts of starting material, in Figure 1c and 1d we compared a 2 μg sample to a 10 μg

sample from the *same* animal (sample A, x-axis in a and b). As can be seen, the variation due to differences in sample size plus the technical variability is less than the between-animal variation shown in Figure 1a and 1b. The 2 μg arrays have an average of 134 decreases and 136 increases when compared to the 10 μg samples from the same animals (Table 5, 2 *vs.* 10); this reflects technical variability of samples prepared by the same protocol from different amounts of starting material. Only 4% of these changes had a magnitude of 2-fold or greater, for an average of 10 changes per chip. This variability was much smaller than the biological variation seen between pairs of arrays from different animals (10 μg arrays, Table 4). There were a balanced number of increases and decreases. The changes in the 2 μg samples appear random: different probe sets change in different comparisons. Only 8 probe sets changed consistently (in at least 5 of the 6 comparisons).

To examine variability introduced by the amplification protocol, we compared a 0.5 μg sample to a 10 μg sample

from the same animal (Figures 2a and 2b). This comparison includes both systematic and random variability introduced by the amplification protocol used for the 0.5 µg sample. There was much more variability between the 10 µg and 0.5 µg samples than between 10 µg and 2 µg samples or between 10 µg samples (compare Figures 1 and 2). Only probe sets that were called present in the 10 µg sample have been plotted in Figure 2b; the variability is still high. This remaining variability extends over a wider range of signals than for the comparisons in Figure 1. More of the probe sets decreased in signal than increased in signal. Many more of these changes are at least 2-fold. The comparison of 0.1 µg to 10 µg data (not shown) is very similar to the 0.5 µg to 10 µg comparison.

The number of changes observed from the comparison analysis of the amplified samples to the 10 µg samples is significantly higher than for the 2 µg samples, and there are many more decreases than increases (Table 5, 0.5 *vs.* 10 and 0.1 *vs.* 10). The number of decreases exceeds 1000 per chip for both amplified groups. For the amplified samples, nearly 2/3 of the changes are 2-fold or greater (Table 5), another indication of the increased variability also seen in Figures 2a and 2b. The changes for the amplified samples were spread evenly across the range of signals, except there were fewer increases seen in the low-level transcripts of the 0.1 µg samples than for the 0.5 µg samples, reflecting the loss of more low-level transcripts in the 0.1 µg samples. Decreases in the amplified samples were more consistent, with 727 and 794 probe sets that decreased in at least 7 of the 8 samples for the 0.5 µg and 0.1 µg groups, respectively. Of the 727 probe sets that consistently changed in the 0.5 µg samples, 712 decreased in at least 6 of the 0.1 µg samples. This indicates that a group of probe sets is being *systematically* affected in both of the amplified groups (see below). A percentage of these decreases actually result in loss of detection of probe sets, 33% for the 0.5µg and 43% for the 0.1 µg samples.

To measure the level of variability within an amplified group, the 0.5 µg arrays within a treatment group were compared (Table 4, 0.5 *vs.* 0.5). The number of changes within the 0.5 µg arrays was greater than within the 10 µg arrays, 932 per chip compared to 709 for the 10 µg. Not only were there more changes, a larger percentage of the changes were 2-fold or greater, 30% *vs.* 12% for the 10 µg samples. This indicates that additional noise was introduced by the amplification.

The 0.1 µg and 0.5 µg samples from the same animal were similar to each other (Figures 2c and 2d, Table 5). Only 16% of the differences between the 0.1 µg and 0.5 µg groups were ≥ fold, compared to 65% that were ≥ 2 fold when comparing the amplified samples to 10 µg samples. The variation between the amplified groups (0.1 µg and

0.5 µg) is greater than the variation between the non-amplified groups (2 µg and 10 µg), and decreases outnumber increases because of the greater loss of signal in the 0.1 µg samples.

Amplification truncates the 5' ends of the RNA

A likely cause for the consistent decreases of particular probe sets in the amplified samples, not related to low signal level, is the loss of the 5' end of the transcript. Synthesis of cDNA from the cRNA initially prepared is expected to lead to some truncation of the 5' ends of the original mRNA, due to the requirement for priming during synthesis of the second strand. Indeed, the cRNA prepared by amplification from the 0.5µg samples was noticeably shorter than that prepared by the standard protocol, as detected by agarose gel electrophoresis (above). Another way to detect such potential shortening of the probe sets is to compare the signals from the Affymetrix control probe sets. There are 3 probe sets each for GAPDH and β-actin, designated 3', Middle and 5' based on their relative distance from the 3' end of the transcript. The average 3'/5' ratio for both 10 µg and 2 µg samples was 1.7 or below (Table 1), representing good samples http://www.affymetrix.com. The 3'/5' ratios of the amplified samples all exceeded 3 and were as high as 14, with the average near 6 for GAPDH and 8.5 for β-actin (Table 1). These ratios indicate a differential loss of the 5' ends of the transcripts for the amplified samples (see below).

Examination of the Affymetrix comparison analyses for the GAPDH and β-actin probe sets gives an even better picture of the 5' loss. None of the comparisons of the 2 µg samples to the 10 µg samples from the same animals showed a significant change in signal for these probe sets. All 8 of the 0.5 µg and 0.1 µg samples had significant decreases as compared to the 10 µg sample from the same animal, the magnitude of which increases as the distance of the probe sequence from the 3' end of the transcript increases (Table 6). The amplified samples both show similar progressive loss of the more 5' sequences. For example, both amplified samples have an average of 33% of the GAPDH 3' signal (mean $\log_2$ ratio -1.6), but only 11% as much GAPDH 5' signal ($\log_2$ ratio-3.1).

To determine if truncation of the 5' ends of the RNA may be a significant problem for many of the sequences on the GeneChip®, the distance of the target sequence (from which probe sets were designed) from the 3' end of the transcripts was determined for as many probe sets as possible by BLASTing the target sequence against the **nr** database (Figure 3). We then compared the percent of probe sets that decreased in signal at different 3' distances. The differences between the 2 µg and 10 µg samples were evenly distributed across the 3' distances (Figure 4), reinforcing the idea that these are random differences. In con-
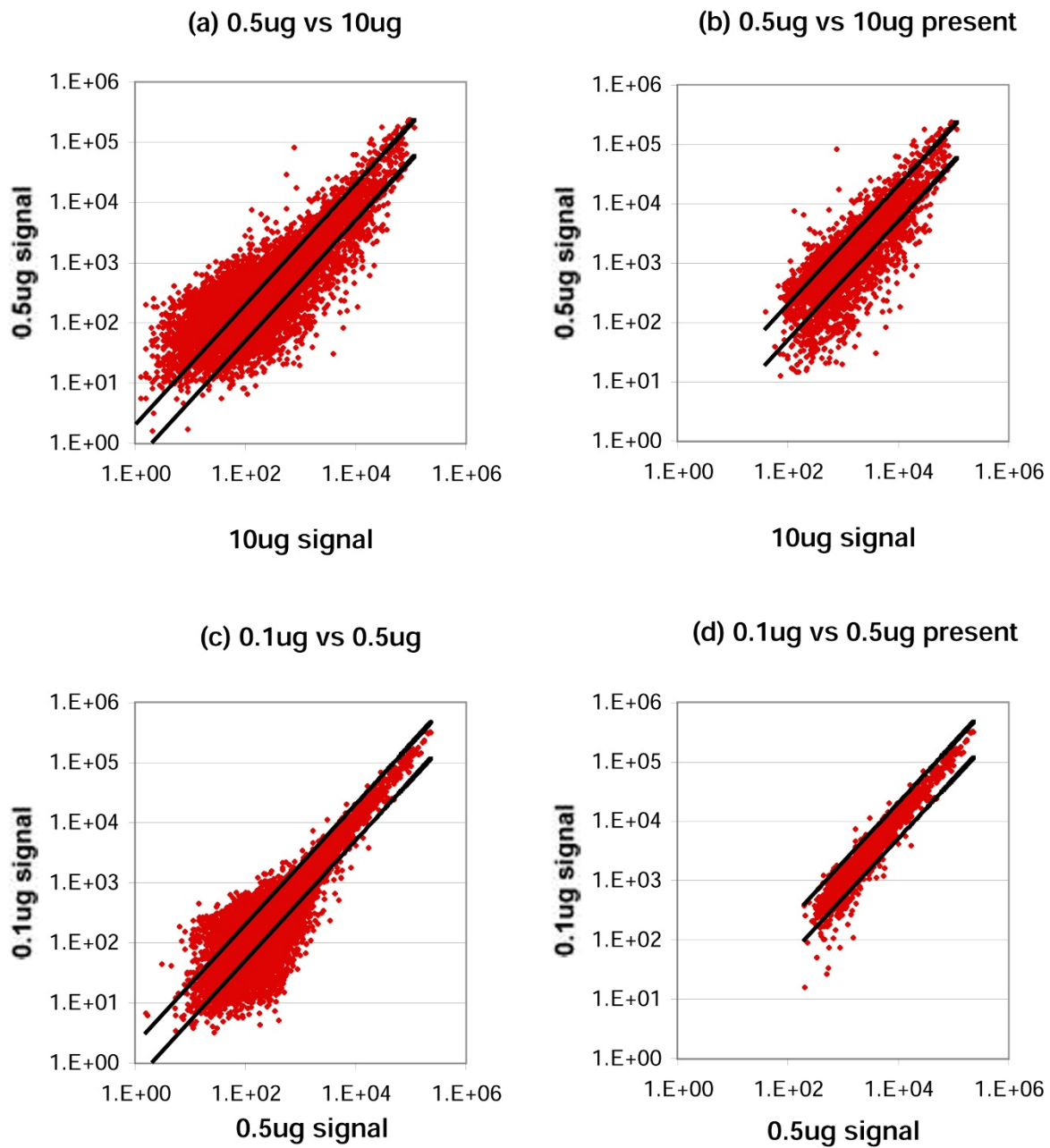
(a) 0.5ug vs 10ug

(b) 0.5ug vs 10ug present

(c) 0.1ug vs 0.5ug

(d) 0.1ug vs 0.5ug present

Figure 2
Log scale scatter plot of amplified samples. Lines indicate a 2-fold difference. (a) 0.5 μg sample vs. 10 μg sample from the same animal as in figure 1c. (b) Same as a, but limited to probe sets called present in the 10 μg sample. (c) 0.1 μg vs. 5 μg from the same animal as in figures 2a and 2b. (d) Same as figure 2c, but limited to probe sets called present in the 0.5 μg sample.
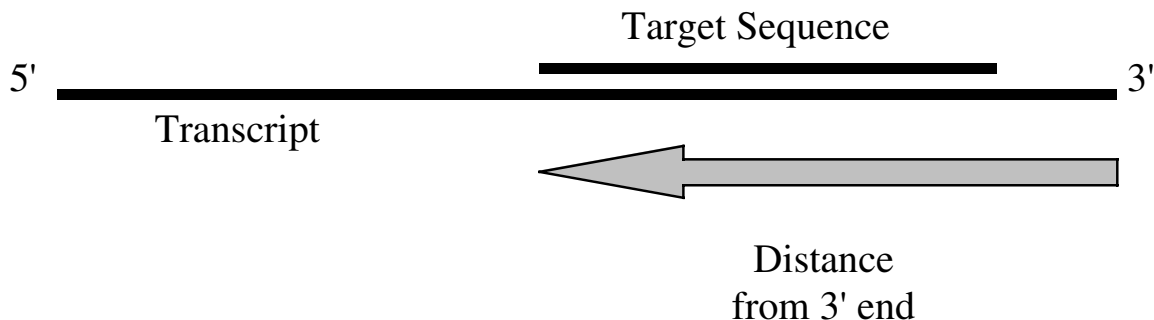
Figure 3
Distance of the target sequence from the 3' end of the transcript. The target sequence from which the probe sets are designed is aligned to the transcript. The distance in bases from the 3' end is calculated by subtracting the position of the 5' end of the alignment from the length of the transcript.

Table 6: Amplified samples show differential loss of 5' sequences.

	Mean log$_2$ratio		Target range
Probe set	0.1 μg	0.5 μg	(nt from 3')
GAPDH 5'	-3.1	-3.1	883–1153
GAPDH M	-2.6	-2.7	541–835
GAPDH 3'	-1.6	-1.6	210–492
β-actin 5'	-3.2	-3.7	798–1116
β-actin M	-2.5	-2.6	432–726
β-actin 3'	-1.0	-1.3	0–390

Average log$_2$ ratios of expression in the amplified samples to that in the control sample (that started with 10 μg RNA) from the same animal. Target range is the position of the target sequence from the 3' end of the transcript.

trast, the amplified samples (0.1 μg and 0. 5 μg) show a marked increase in the percent of probe sets that were decreased as the target sequence is moved farther from the 3' end (Figure 4).

Both sets of amplified samples were affected in the same manner by this 5' truncation. The decreases seen when comparing the 0.1 μg to the 0.5 μg samples were not associated with distance from the 3' end and were not consistent for particular probe sets (only 26 probe sets decreased in at least 7 of the 8 samples). This indicates that the truncation is a result of the single cycle of amplification, rather than the amount of starting material.

Ability to detect differences in expression between treatment groups
The main goal in an experiment comparing two treatment groups is to find genes whose expression differs significantly. To assess whether the lower starting amounts of total RNA can be used successfully, a comparison of results from standard t-tests was performed. Based upon the data in Figure 1, we filter out those probe sets that are not detected in at least one of the treatment groups to be compared before performing statistical comparisons. To be conservative, we only eliminated probe sets that are not called "present" on at least half of the GeneChips® in either of the treatment groups (rather than demand that a probe set be present in all of the GeneChips® in a set; others can choose different fractions), and call this a
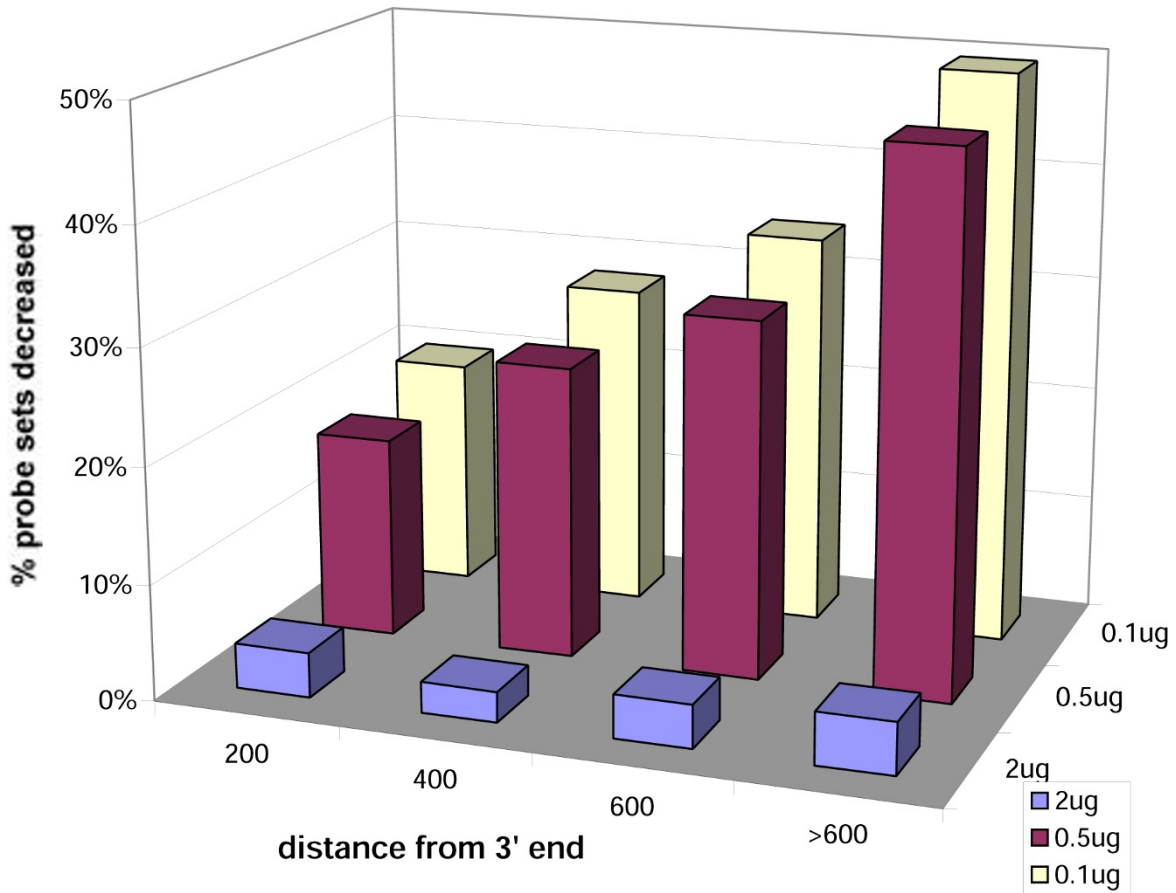
Figure 4
Reduction in signal as a function of the distance from the 3' end. Affymetrix comparison analysis was used to compare the reduced RNA samples to the 10 μg sample from the same animal. This plot shows the percentage of probe sets called "decreased" as a function of distance from the 3' end of the transcript. Analysis was limited to probe sets called present in the 10 μg sample.

"detection filter." This does not eliminate probe sets that are either turned on or off, because these would be present in one of the two treatment groups.

Table 7 gives the number of probe sets that met our criteria for significance: they passed the "detection filter" and were significant at p ≤ 0.01 for the t-test or at the appropriate level for the Wilcoxon rank sum test (non-parametric) [13]; both tests give generally parallel results. Even

though both the 0.5 μg and 0.1 μg comparisons are 4 × 4 comparisons, there is a 50% drop in the number of significant probe sets as compared to the 10 μg samples. Only part of this drop can be attributed to the decrease in the percent of probe sets that met the "detection filter" for these samples (Table 7); the extra noise introduced by the amplification, as seen in the 0.5 μg group (cf. Figure 2, Tables 4 and 5), contributes substantially, since the t-test is

Table 7: Number of significant probe sets for different groups of samples.

		Significant Difference	
Group	Detection[1]	t-test[2]	Wilcoxon[3]
10 μg(4)*	3765	150	340
0.5 μg (4)	3090	75	175
0.1 μg(4)	2461	80	193
10 μg(3)	3662	86	NA
2 μg (3)	3666	83	NA

The number significant is the number of probe sets significantly different between vitamin sufficient and -deficient animals among those that are present in at least half of the arrays in at least one of the groups, [1]number of probe sets that passed the detection filter, [2]p-value ≤ 0.01 (t-test); [3] p-value ≤ 0.0143 for Wilcoxon test; *Number in parentheses is the number of GeneChips® in each treatment group for the comparison;

sensitive to increases in standard deviations of the two groups being tested.

Because of the loss of two of the 2 μg arrays, the table also contains data for the subset of six that match the six remaining 2 μg arrays. Note the sharp decline in the number of probe sets meeting the p ≤ 0.01 significance criteria (t-test) when the number of samples in the 10 μg class is reduced from a 4 × 4 comparison to a 3 × 3 comparison (Table 7); this attests to the additional power gained by using the additional array. The best (lowest) p-value that can be achieved in the Wilcoxon test with four samples for each treatment group is 0.0143; for three samples per group it is 0.05 [13]. The 2 μg samples also produced a similar number of significant probe sets for the Wilcoxon with a 0.05 p-value as the 10 μg samples with the same number of arrays, 585 for the 2 μg samples and 620 for the corresponding 10 μg samples. These numbers can be compared to the 869 found in the complete set of 10 μg samples (sample size of 4) with a p-value of 0.0571 for the Wilcoxon.

As expected, probe sets with low expression level (low signal) were less reproducible in comparisons between the different sample groups, as were probe sets with very low fold-changes. Reproducibility for the 2 μg samples was best for probe sets with a fold change ≥ 1.7 (log_2 ratio ≥ 0.75). For the amplified samples, good reproducibility was achieved for probe sets with fold changes ≥ 2 (log_2 ratio ≥ 1). Calculated fold changes for the concordant probe sets were reasonably stable across the different groups.

Estimate of technical false positive rate

The 2 μg and 10 μg groups were from the same original RNA extractions and were labeled by the same protocol. Because they were so comparable in all of our measures, these 2 groups were used to estimate the number of false positives due to technical variability to be expected using our standard t-tests. The three 2 μg samples from the nor-

mal diet animals were compared to the three 10 μg samples from the same animals using the t-test, and the three 2 μg samples from the diet deficient animals were similarly compared to the 10 μg samples from the same animals. Since both of these comparisons are between samples from the same set of similarly treated animals, one should expect no changes. Therefore, any probe sets that were found to be significantly different between the 2 μg and 10 μg samples from the same animals would be false positives. For each comparison, normal and deficient, 14 probe sets were found to be significantly different between the 2 μg and 10 μg samples, which is a false positive rate of 0.4% of the probe sets that met the "detection filter." Of these 28, only 3 of the normal group and 4 of the deficient group had a fold-change greater than 1.5 and only one in each group had a fold-change that exceeded 2. Fold-changes larger than 1.5 were seen only in probe sets with average signals < 900. Those probe sets with signals over 900 had smaller fold changes, most less than 1.3.

In comparison, for those probe sets that did not meet the "detection filter", the false positive rate was approximately 1% at a p-value of 0.01. This set of false positives was equally balanced between increases and decreases (53 increases and 58 decreases in the two treatment groups), indicating random noise and not an effect of using 2 μg instead of 10 μg of RNA. The fold changes ranged from 1.2 to 12.5. The large fold changes result from very small denominators used in the fold change calculations for these probe sets.

Discussion

This experiment demonstrated that technical variability is much smaller than biological variation when using the standard protocol. The number of differences between samples from the same animal that resulted from using the lower amount of starting material was much smaller than the biological variation between animals treated the same and labeled by the standard protocol. The standard

protocol worked well with samples as small as 2 µg, producing results very similar to those of the 10 µg sample from the same animal. Although in this experiment, starting with less than 1 µg of total RNA did not produce enough biotinylated cRNA for hybridization under the normal protocol, a minor change (using vacuum evaporation to concentrate samples before cRNA synthesis, and mixing the minimum 200 µl hybridization volume) should allow use of 1 µg samples with the standard protocol. This extends the range of samples that can readily be analyzed on Affymetrix GeneChips® using the standard protocol.

We have demonstrated here that a single cycle of amplification sufficed to produce cRNA from samples as low as 0.1 µg of total RNA. The amplification protocol uses the cRNA from the initial protocol as starting material for a second round of cDNA synthesis and *in vitro* transcription. We hypothesized that each round of cDNA synthesis would lead to some truncation of the molecules, due both to the possibility of priming the second strand from an internal site and to cleavage of the relatively labile RNA. For this reason, we limited our amplification to a single round, rather than using two cycles as had previously been reported [10]. The hypothesized shortening was observed: the cRNA extended to about 850 nt, compared with about 2000 nt for the standard protocol. This shortening was also demonstrated by differential loss of signal from probe sets further away from the 3' end of the RNA, as shown with the Affymetrix control probe sets (Tables 1 and 6), and by the progressive loss in probe sets detected as a function of their distance from the 3' end (Figure 4). The amplification systematically affected over 700 probe sets in at least 7 of 8 samples for both the 0.1 µg and 0.5 µg groups. Indeed, of the probe sets with consistent decreases across the 0.1 µg and 0.5 µg samples about 90% have target distances greater than 400 nucleotides from the 3' end of the measured transcript, with 26–28% in the 400–600 range and 60–62% over 600. Because the effect of the loss of the 5' end is systematic, it renders a group of probe sets designed from sequences further from the 3' end undetectable. Therefore, we recommend using the standard protocol instead of using the amplification strategy for samples down to 1 µg of total RNA, and our simplified amplification protocol for smaller samples, at least down to 0.1 µg. This extends the range of samples that can be usefully analyzed by oligonucleotide microarrays. This experiment used the Affymetrix RGU34A GeneChip®, which was designed using version 34 of Uni Gene for the rat, November 1998 For newer arrays such as the new human U133 GeneChip®, designed using better sequence information and improved probe designs (Affymetrix technical report Array Design for the GeneChip® Human Genome 133 Set), we expect that the problem with loss of the 5' end of the transcript. will be lessened,

but not eliminated, because the targets are more likely to be near the real 3' ends of the mRNAs The truncation that results from amplifying samples means that the same protocol should be used for all samples in a given study.

New protocols that increase the production of cDNA [14,15] using primers attached to the 5' end of the transcript could increase cDNA yields from small amounts of RNA But Iscove [14] states "only a few hundred bases of extreme 3' sequence" are amplified by their method This procedure would be expected to greatly exacerbate the loss of signal due to shortened transcripts.

There was additional variation in the amplified samples (Table 5) that is probably due to the extra steps required in the protocol. This extra noise is partially responsible for the decrease in the number of probe sets that differ at a p-value ≤ 0.01. This, coupled with the decrease in the percent of probe sets present, reduces the ability to find transcripts that differ significantly in expression when using the amplification protocol. The problem is exacerbated for transcripts with small differences in expression (low fold changes). Fold changes ≥ 2 were much more likely to be identified in the amplified samples. It may be especially helpful to increase the number of arrays used in these amplification experiments to get more power to detect changes

Our experiment also provided a false positive estimate of technical variability for the t-test with 14 found in each treatment group when comparing the 10 µg samples to the 2 µg samples (0.4% of the present probe sets.). These false positives came disproportionately from the genes expressed at lower levels. Genes expressed at these low levels often show high fold-changes because the denominator is so low (often near background), this points out the danger in emphasizing high fold-changes, rather than reproducible changes. Therefore, for genes expressed at lower levels, it might be reasonable to use a more restrictive p-value, none of the false positives had a p-value less than 0.001. Restricting probe sets to those minimally present (our "detection filter") dramatically decreases the number of false positives, from an average of 56 down to 14, restricting analysis to genes called present in a higher fraction of the arrays from one of the comparison groups could further reduce false positives, but at the cost of missing some true positives. The tradeoff can be chosen by an investigator based upon the relative cost of false positives and value of detecting differences in genes expressed at low levels.

The statistical power to detect differences was much reduced when 3 samples per group were analyzed instead of 4. The 2 µg samples and the matched 10 µg samples were able to detect 83–86 differences at p ≤ 0.01 as compared

to 150 differences when using all of the 10 µg samples. This decrease was expected, but illustrates that a 25% decrease in expense may result in a much greater loss of information.

Conclusions

This experiment explored the effects of using less than the standard 10 µg of total RNA for an Affymetnx GeneChip® experiment, and examined how biological and technical variation affect the ability to detect biological differences in a typical experiment comparing gene expression in two conditions. The overall conclusions are that (1) small amounts of RNA can be used effectively in the standard protocol, (2) even very small amounts of RNA (0.1 µg) can be used with our simplified amplification protocol to detect differential gene expression, (3) biological variation is larger than technical variation, (4) very low-level signals are prone to false positives and to less reliable fold-changes, false positives can be reduced by filtering out probe sets not reliably detected before statistical comparisons, and (5) using 4 independent biological samples is much better than using 3 samples in allowing detection of consistent changes and reducing false positives.

Methods

Labeling test

Total RNA was extracted from the livers of 4 rats fed a normal diet (untreated) and 4 fed a vitamin-deficient diet (treated) using the RNeasy® kit (Qiagen Inc, Valencia, CA). The RNA was resuspended and re-extracted using the same protocol, to reduce DNA contamination. For the labeling test, two pools were created, one treated and one untreated, by mixing equal aliquots from each of the 4 RNA samples. The final concentration was adjusted to 1 µg/2 µl, and each pool was divided into 4 aliquots of 1 µg each. These pooled samples were used to determine the average cRNA yield from 1 µg of total RNA and to compare the yields of the T7 polymerases from two different *in vitro* transcription kits. Biotinylated cRNA was prepared using the standard Affymetrix protocol [8] except that the Epicentre AmpliScribe™ T7 polymerase (Epicentre, Madison, WI) was substituted for the ENZO T7 polymerase (BioArray, High Yield RNA Transcript Labeling Kit, ENZO Diagnostics, Inc., Farmingdale, NY) for 2 of the treated and 2 of the untreated pooled samples. The yield of cRNA was estimated from absorbance at 260 nm, using an Amersham Pharmacia Biotech Ultrospec 3100 pro spectrophotometer. This part of the experiment was the only time RNAs from different animals were pooled, the cRNAs from these pooled samples were not hybridized to arrays.

Labeling for microarrays

Aliquots of each of the original 8 samples of total RNA (one from each rat) were treated as individual samples for all hybridization experiments. Each sample was serially

diluted to yield 10 µl samples containing 10 µg, 2 µg, 0.5 µg and 0.1 µg total RNA (32 samples). For the 10 µg and 2 µg samples, cRNA was prepared using the standard Affymetrix protocol [8]. We made slight modifications for the 2 µg samples to increase the concentration by decreasing added water where that was possible. Because the cRNA yield from the 1 µg pooled sample test was low, we decided to amplify the smaller samples. The 0.5 µg and 0.1 µg samples were amplified by a modification of the protocol of Ohyama et al. [10], using only a single round of amplification. In short, double-stranded cDNA was synthesized from the total RNA using the SuperScript II kit from Invitrogen and the Affymetrix $T7-(dT)_{24}$ primer, which contains a T7 promoter attached to a poly-dT sequence 5'-GGCCAGTGAATTGTAATACGACTCACTATAG-GGAGGCGG-$(dT)_{24}$-3'. The Epicentre AmpliScribe T7 High Yield Transcription kit was used to produce unlabeled cRNA by *in vitro* transcription with unbiotinylated NTPs. This *in vitro* transcription step was followed by a second round of double-stranded cDNA synthesis, and finally by *in vitro* transcription using the T7 RNA polymerase with the Enzo BioArray, High Yield RNA Transcript Labeling Kit with biotinylated NTPs in the usual manner. Yields of cDNA and cRNA were measured using an Amersham Pharmacia Biotech Ultrospec 3100 pro spectrophotometer in order to adjust concentrations for subsequent steps. Measurements of the amplified samples were made after the second round of cDNA synthesis and *in vitro* transcription, to limit the loss of sample. After the final round of cRNA synthesis, aliquots of biotinylated cRNA from each sample were electrophoresed on 1% agarose gels in TBE buffer to check them for quality and length, the buffer and gel both contained a 500 ng/ml concentration of ethidium bromide. The Invitrogen 1 Kb Plus DNA ladder was used to provide a relative measure of length.

Since it was not feasible to label this many samples at one time, a balanced experimental design was used processing groups of 10 µg and 2 µg samples for the same animal together and always labeling an equal number of normal and vitamin-deficient samples at one time. For the amplification protocol, 0.5 µg and 0.1 µg samples from the same animal were processed together for each step, again with balanced numbers of normal and vitamin deficient samples processed together.

Array hybridization

Each sample was hybridized to a separate Affymetrix RGU34A GeneChip®. For most samples, hybridization cocktails of 300 µl contained 15 µg of fragmented cRNA, 200 µl of the cocktail were injected into the GeneChip® for hybridization. One 2 µg sample had a lower yield so a 200 µl hybridization cocktail containing 10 µg of cRNA was made, to keep the concentration fixed. For the 0.1 µg samples, the cRNA yield was less than 15 µg so 5–7 5 µg

of cRNA was used for the hybridization cocktail. Data from two of the 2 μg GeneChips® (1 from each treatment group) and their hybridization cocktail were unusable due to a bad lot of BSA (used as a blocking factor during hybridization). The analyses for the 2 μg group were completed using the remaining 6 GeneChips®. For some comparisons with this group, only the corresponding 6 GeneChips® from the 10 μg samples were used.

Scanning and analysis

Each GeneChip® was scanned and analyzed using Affymetrix Microarray Analysis Suite (MAS) version 5.0 [11]. Each sample was scaled to a target intensity of 1000 using the "all probe sets" scaling option, this option scales the trimmed mean target intensity to the specified value [11]. The Affymetrix MAS5 expression report provides statistics for each chip that can be used for quality control purposes. Included in this report are noise, background, the percent of probe sets called present, the scaling factor calculated by the absolute analysis algorithm and the ratio of 3' to 5' signal for GAPDH and β-actin. These measures were used to judge the quality and similarity of data from the various RNA sample size groups.

MAS5 "absolute" and "comparison" expression analyses were performed [11]. The RGU34A chip contains 8799 probe sets (series of probe pairs that query parts of the same gene or EST) [7]. For each probe set, absolute analysis generates a signal value (expression level), a detection call of absent, present, or marginal, and a p-value associated with the detection call [11]. Comparison analysis examines 2 GeneChips® and indicates for each probe set whether there is a significant difference in the signal between the two arrays. The output is a change call of increase, marginal increase, decrease, marginal decrease or no change, a p-value associated with the change call, and the magnitude of the difference (as the signal log ratio, the $\log_2$(log base 2) ratio of the signal from first chip to the signal on the second or baseline chip). Comparisons were made between the vitamin-deficient and normal diet animals in each sample size group using the normal diet samples as the baseline for each comparison. All possible comparisons were made, this equaled 16 comparisons each (4 × 4) for the 10 μg, 0.5 μg and 0.1 μg groups and 9 comparisons for the 2 μg samples. Comparisons were also made between each of the 2 μg, 0.5 μg and 0.1 μg samples and the 10 μg sample from the same animal, to look for apparent differences in gene expression that are really due to the differences in the amount of starting material and sample processing. For these comparisons, the 10 μg sample was used as the baseline.

In addition, comparisons were made between the 10 μg arrays within each treatment group to measure variability, both between-animal and technical. This analysis was done for the 0.5 μg group as well. The 0.1 μg arrays were compared to the 0.5 μg arrays from the same animal, to determine how well these two amplified samples corresponded with one another.

Data from each RNA sample size group were exported and each probe set analyzed in the following ways:

1. The number of GeneChips® on which a probe set was detected within each of the treatment groups (normal vs. deficient diet) was calculated. Each present call was assigned the value 1 and each marginal call assigned 0.51. Only probe sets that are present in at least one-half of the samples from one of the treatment groups (either normal or deficient diet) were used for further analysis, we call this the "detection filter". Note that we did not require the probe set to be present in both treatment groups, just in one. For this experiment, we required a score of at least 2 for the 10 μg, 0.1 μg and 0.5 μg groups (4 GeneChips® in each treatment group) and 1.51 for the 2 μg and corresponding 6-sample subset of the 10 μg groups (3 GeneChips® in each treatment group).

2. The sum of the calls from the comparison analyses was calculated, using 1 for increase, 0.5 for marginal increase, 0 for no change, -0.5 for marginal decrease, and -1 for decrease.

3. The mean, standard deviation and coefficient of variation of the signal for each probe set within each treatment group (normal and deficient diet) were calculated.

4. Student's t-test for equal means with the assumption of unequal variance [12] was calculated to test for significant differences in signal (expression level) between the normal and vitamin deficient groups. This test was applied separately to the signals and the $\log_2$ transformation of the signals.

5. The Wilcoxon rank-sum non-parametric test for equal means between normal and deficient diet groups [13] was performed.

6. The $\log_2$ ratio (fold change) of the mean signals for normal *vs.* deficient diet groups was calculated:

$\log_2$ ratio = $\log_2$((mean of signal for normal) / (mean of signal for deficient))

Changes within an RNA sample size group between normal and deficient diet groups were assessed as significant if the probe set passed the "detection filter" (#1 above) and the p-value was less than 0.01 for the t-test using either the original signal or log transformed signal (#4 above).

Scatter plots comparing individual samples were created using Microsoft Excel®. SAS (SAS Institute Inc., Cary, NC) was used for univariate analysis of the signals for each RNA sample size group.

Distance of target from 3' end

To calculate the distance of the target from the 3' end of the interrogated sequence, the target sequence for each probe set (supplied by Affymetrix; http://www.affymetrix.com was BLASTed [16,17] against the nr (non-redundant) database at NCBI http://www.ncbi.nlm.nih.gov. Using the first (best) alignment returned, the distance of the 5' end of the target sequence from the 3' end of the mRNA or EST in the nr database was calculated (Figure 3). A few of the alignments were rejected because the first alignment was to a BAC sequence or to the wrong strand.

Abbreviations

cRNA, complementary RNA

cDNA, complementary DNA

IVT, *in vitro* transcription

MAS5, microarray analysis suite version 5 from Affymetrix

nr, non-redundant (nucleotide database at NCBI)

nt, nucleotides

Authors' contributions

JNM participated in the design of the study, performed statistical analysis and drafted the manuscript. REJ and CRN carried out the biochemistry and aided in optimizing the methods. DWC designed and coordinated the original dietary study and provided the total RNA for this work. HJE conceived the study, and participated in its design, coordination, analysis and writing. All authors read and approved the final manuscript.

Acknowledgements

References

1. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO and Herskowitz I **The transcriptional program of sporulation in budding yeast.** *Science* 1998, **282:**699-705
2. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D and Futcher B **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297
3. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG and Thomas DG **Gene-expression profiles predict survival of patients with lung adenocarcinoma.** *Nat Med* 2002, **8:**816-824
4. Giordano TJ, Shedden KA, Schwartz DR, Kuick R, Taylor JM, Lee N, Misek DE, Greenson JK, Kardia SL and Beer DG **Organ-specific molecular classification of primary lung, colon, and ovarian adenocarcinomas using gene expression profiles.** *Am J Pathol* 2001, **159:**1231-1238
5. Matejuk A, Dwyer J, Zamora A, Vandenbark AA and Offner H **Evaluation of the effects of 17beta-estradiol (17beta-e2) on gene expression in experimental autoimmune encephalomyelitis using DNA microarray.** *Endocrinology* 2002, **143:**313-319
6. Lipshutz RJ, Fodor SP, Gingeras TR and Lockhart DJ **High density synthetic oligonucleotide arrays.** *Nat Genet* 1999, **21:**20-24
7. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H and Brown EL **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nat Biotechnol* 1996, **14:**1675-1680
8. Affymetrix **GeneChip® Expression Analysis Technical Manual.** *Santa Clara, CA: Affymetrix* 2001,
9. Mahadevappa M and Warrington JA **A high-density probe array sample preparation method using 10- to 100-fold fewer cells.** *Nat Biotechnol* 1999, **17:**1134-1136
10. Ohyama H, Zhang X, Kohno Y, Alevizos I, Posner M, Wong DT and Todd R **Laser capture microdissection-generated target sample for high-density oligonucleotide array hybridization.** *Biotechniques* 2000, **29:**530-536
11. Affymetrix **Affymetrix Microarray Suite User's Guide, version 5.0.** *Santa Clara, CA; Affymetrix* 2001,
12. Freund JE **Mathematical Statistics.** *Upper Saddle River, NJ: Prentice Hall* 1992, 462-465
13. Pagano M and Gauvreau K **Principles of Biostatistics.** *Pacific Grove, CA: Brooks/Cole* 2000, 308-312
14. Iscove NN, Barbara M, Gu M, Gibson M, Modi C and Winegarden N **Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA.** *Nat Biotechnol* 2002, **20:**940-943
15. Ginsberg SD and Che S **RNA amplification in brain tissues.** *Neurochem Res* 2002, **27:**981-992
16. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410
17. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402