Nucleic Acids Research, 2006, Vol. 34, No. 12 ***e83***
doi:10.1093/nar/gkl404

Sequence biases in large scale gene expression profiling data

Asim S. Siddiqui, Allen D. Delaney, Angelique Schnerch, Obi L. Griffith, Steven J. M. Jones and Marco A. Marra*

Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Research Centre,
British Columbia Cancer Agency, Vancouver, BC, Canada

Received February 17, 2006; Revised May 7, 2006; Accepted May 15, 2006

ABSTRACT

We present the results of a simple, statistical assay that measures the G+C content sensitivity bias of gene expression experiments without the requirement of a duplicate experiment. We analyse five gene expression profiling methods: Affymetrix GeneChip, Long Serial Analysis of Gene Expression (LongSAGE), LongSAGELite, 'Classic' Massively Parallel Signature Sequencing (MPSS) and 'Signature' MPSS. We demonstrate the methods have systematic and random errors leading to a different G+C content sensitivity. The relationship between this experimental error and the G+C content of the probe set or tag that identifies each gene influences whether the gene is detected and, if detected, the level of gene expression measured. LongSAGE has the least bias, while Signature MPSS shows a strong bias to G+C rich tags and Affymetrix data show different bias depending on the data processing method (MAS 5.0, RMA or GC-RMA). The bias in the Affymetrix data primarily impacts genes expressed at lower levels. Despite the larger sampling of the MPSS library, SAGE identifies significantly more genes (60% more RefSeq genes in a single comparison).

INTRODUCTION

Gene expression profiling methods are used extensively for quantitative transcriptome analysis (1–6). Array based methods, such as Affymetrix GeneChip, rely on the hybridization of labeled transcript-derived sequences to oligonucleotide probes synthesized on the microarray. Probe hybridization intensity values are used to measure the expression levels of transcripts.

Sequencing methods, such as MPSS and SAGE generate a short sequence tag for transcripts. The gene expression level is measured by counting these tags. For 3′ tag-based methods, the 'tag' identified is a short stretch of nucleotides adjacent to (and including) the 3′ most site of a specific restriction enzyme in a transcript. The CAGE method is a variant that detects the 5′ most site (7). By generating tag sequences derived from an mRNA population and mapping these tag sequences to transcript databases, one can infer the abundance of transcripts in the original population. Longer tag sequences provide better specificity of mapping tags to genes. SAGE [14 bp tags; (1)] and its variants LongSAGE [21 bp tags; (2)], LongSAGELite [21 bp tags from nanogram quantities of RNA; (8)], SuperSAGE [26 bp tags; (9)] produce concatemers of tags which are cloned and sequenced using standard Sanger di-deoxy terminator sequencing chemistry. The total number of tags sequenced determines the accuracy of gene expression quantization (10) and the ability to detect rare transcripts. Recently, MPSS [20 bp tags; (4)], a sequencing technology able to generate millions of short sequences in parallel, has been used to create digital gene expression profiles at apparently lower costs than typically incurred for LongSAGE libraries sequenced using capillary sequencers and the Sanger chemistry.

New applications of gene expression profiling methods [MPSS, 5′ SAGE, poly(A)± Affymetrix tiling arrays (6,7)] are providing rich views of the transcriptome. Many papers have compared the results of different gene expression experiments across platforms and assessed platform reproducibility (11–18) without providing an explanation for the observed differences. Mecham *et al.* (19) and Carter *et al.* (20) demonstrated that incorrect assignation of Affymetrix probes to genes as one source of variability and suggested sequence-based mapping as a corrective measure. Kluger *et al.* (21) showed that the physical location of probes on the Affymetrix GeneChip is correlated with gene expression levels. The effect of G+C content on the stability of hybridized sequences is well known with higher G+C content corresponding to more stable DNA duplexes (22). Kuo *et al.* (23) found that several probe-specific factors, including G+C content, were associated with the degree of correlation between gene expression levels of the same mRNA sample

*To whom correspondence should be addressed at Genome Sciences Centre, Suite 100, 570 West 7th Avenue, Vancouver BC, Canada V5Z 4S6. Tel: 604 877 6082; Fax: 604 877 6085; Email: mmarra@bcgsc.ca measured by different microarray technologies. Margulies *et al.* (24) identified G+C content bias in 10 bp SAGE libraries and suggested a statistical analysis to assess G+C bias before deep sequencing was initiated. They gave several experimental steps for reducing bias and suggested that longer LongSAGE ditags may not suffer from the same degree of bias. Dinel *et al.* (25) demonstrated that SAGE replicate data was highly reproducible if one accounted for differences in concatemer and ditag replication, the number of sequenced tags and double PCR amplification of ditags.

Here, we compare and analyse publicly available gene expression data generated using one microarray method and four sequencing methods. By using a large number of publicly available experiments, we are able to measure the variation of the G+C sensitivity of all five methods by measuring the G+C sensitivity of an individual experiment without the requirement of a matched sample for comparison. We demonstrate that the variation results in an intrinsic limit on experimental reproducibility within a method (with current technology and methodology) and the ability to compare results across methods. Furthermore, we demonstrate that the differences in genes observed for both intra-method and inter-method comparisons can be explained directly in terms of the relative G+C sensitivity of the experiments.

MATERIALS AND METHODS

Mapping of tags and probes to genes

Tags were mapped to the 3′ most restriction enzyme site of transcripts using RefSeq NM genes for human and mouse (26) and the TAIR annotated gene set for *Arabidopsis* (27). The restriction enzyme used for the LongSAGE and LongSAGELite experiments was NlaIII (tags begin with CATG). MPSS utilized DpnII (tags begin with GATC). References to the entire set of 3′ most transcripts refer to either the set of NlaIII tags or DpnII tags, whichever is appropriate for the experiment under study.

Probes were mapped to Refseq genes using the annotation files provided by Affymetrix for the HGU 133 GeneChip (using both the A and B chips or using the A chip alone as dictated by the available data). Gene expression levels for the Human ES data were derived using both the MAS 5.0 software and GC-RMA. The gene expression values for the set of 993 experiments were derived from processed by variety of software including MAS (14). Present and absent calls from the MAS 5.0 software were used as a measure of gene presence or absence. For GC-RMA processed data, a gene expression level of 3.32 or greater was required for presence. This cutoff called present the same average percentage of genes on the chip as the MAS 5.0 software. Likewise for RMA processed data, a gene expression level of 22.1 or greater was required for presence.

Measuring G+C bias

For SAGE and MPSS, the bias is calculated from the premise that, for any experiment, the G+C content of the observed set of 3′ most tags should be equivalent to the G+C content of a random sampling of 3′ most tags from all transcripts. The calculation is performed in the following manner. The G+C content of the observed tags that map to the set of 3′ most tags of transcripts is determined (minus the four letter restriction site prefix). From that the deviation of the observed G+C content relative to the G+C content of the entire set of 3′ most tags is derived. To provide a fair basis of comparison that accounts for differences in the number of transcripts observed in each experiment (sampling size) as well as possible differences in the NlaIII versus DpnII tags, for each experiment, we calculate the standard deviation of the G+C content of randomly sampled tags (equivalent in number to the mapped tags that the experiment yields) relative to the entire set of 3′ most tags. The random sampling is performed 1000 times with the G+C content of the sampled tags and its deviation from the G+C content of the entire set measured each time. The standard deviation is derived from the 1000 data points. The measured deviation of the observed G+C content is divided by the standard deviation to give the number of standard deviations by which the observed G+C content varied from the expected mean. It is this measure, the number of standard deviations by which the measured mean G+C content deviates from the expected G+C content, that is used to evaluate G+C bias.

A similar approach is used for Affymetrix data. The analysis is performed at the level of the probe set. The premise is that, for any experiment, the G+C content of the observed set of probe sets that are annotated as mapping to RefSeq transcripts should be equivalent to the G+C content of a random sampling of all probe sets on the chip that are annotated as mapping to RefSeq transcripts. The rest of the calculation is performed the same way as it is for SAGE and MPSS.

Affymetrix intensity values were normalized by taking the natural logarithm of the intensity value and then subtracting the median and dividing by the interquartile range for the experiment. Signals that were reported as marginal or absent by the Affymetrix software were ignored.

RESULTS

Assessment of gene coverage

We began by assessing whether the different methods identified comparable numbers of genes. In Table 1, the first comparison shown is between a LongSAGE library of mouse liver RNA and an MPSS library derived from the same RNA sample. The MPSS library yielded 1 724 799 tags, whereas the SAGE library was sampled to only 108 117 tags. Despite the shallower depth of sampling, the SAGE library identified over 60% more RefSeq genes (3593 versus 2226; Materials and Methods). The total number of RefSeq genes that can be identified by LongSAGE with tags constructed using NlaIII (15 054) is slighter greater than that, which can be identified by MPSS with tags constructed using DpnII (14 685; Table 2), but this difference is not large enough to explain the discrepancy in the number of genes identified especially considering the large difference in sampling depth. A comparison of a SAGE library and MPSS library both created using RNA purified from different adult mouse kidneys gave a similar result. In this experiment, 7291 genes were identified by SAGE and only 3547 by MPSS despite the deeper sampling by MPSS (883 305 SAGE tags

Table 1. Differences in G+C content lead to differences in observed gene sets

Species	Tissue	Additional information	Technology	Library identifier	Sampling depth	Genes identified	Genes identified per sampled tag	Number of genes identified by both methods	Genes identified by this method only	A+T/C+G library bias[a]	A+T/C+G bias of tags/ probes of genes missed by this method[b]
Inter-method comparisons											
M.musculus	Liver	Same mRNA used in both experiments	LongSAGE	SM100	108 117	3593	0.033	1425	2168	−4.31	−0.88
			MPSS (signature)	GSM17243	1 724 799	2266	0.0013		841	−18.91	8.7
	Kidney	P84, male	LongSAGE	SM104	883 305	7291	0.0083	2924	4367	4.49	−5.22
		Adult male, cortex	MPSS (signature)	GSM34298	2 230 467	3547	0.0016		623	−34.06	25.24
	Visual cortex	P27, visual cortex, male	LongSAGE	SM029	115 803	5473	0.047	3713	1718	−0.32	−7.58
			LongSAGE-Lite	SM040	137 562	4702	0.034		989	−11.84	10.89
A. haliana	Callus	Biological replicate	MPSS (classic)	CAF	1 637 407	9732	0.0059	6106	3626	3.46	−13.2
			MPSS (signature)	CAS	1 433 143	7075	0.0049		969	−26.31	29.78
	Inflorescence	Biological replicate	MPSS (classic)	INF	1 455 847	9922	0.0068	7904	2018	3.65	−13.43
			MPSS (signature)	INS	2 516 138	9943	0.004		2039	−19.39	25.95
	Leaves	Biological replicate	MPSS (classic)	LEF	2 457 736	10 311	0.0042	7592	2719	3.22	−11.22
			MPSS (signature)	LES	2 752 425	9071	0.0033		1479	−23.16	30.58
	Root	Biological replicate	MPSS (classic)	ROF	3 002 218	9900	0.0033	7244	2656	10.01	−20.68
			MPSS (signature)	ROS	2 047 569	9193	0.0045		1949	−26.38	38.27
	Silique	Biological replicate	MPSS (classic)	SIF	1 673 908	9715	0.0058	6157	3558	7.04	−13.05
			MPSS (signature)	SIS	1 869 453	7373	0.0039		1216	−22.51	30.63
H.sapiens	ES	H7 p22	Affymetrix (A/B)	WA07 a22b24	n/a	6423	n/a	3643	2780	22.44	−16.76
			LongSAGE	SHE13	272 422	6440	0.024		2797	−2.08	1.42
		H9 MEFs p38	Affymetrix (A/B)	WA09 a1b2	n/a	6014	n/a	3773	2241	25.68	−21.22
			LongSAGE	SHES2	466 042	7358	0.016		3585	2.57	1.11
		H1 p54 matrigel	Affymetrix (A/B)	WA01 a20b21	n/a	6101	n/a	3375	2727	25.74	−18.11
			LongSAGE	SHE16	218 169	5894	0.027		2519	−4.03	3.13
		H14 p22	Affymetrix (A/B)	WA14 a23b25	n/a	6028	n/a	3326	2702	24.35	−17.29
			LongSAGE	SHE14	212 136	6020	0.028		2694	2.56	2.50
		HES4 p36	Affymetrix (A/B)	ES04 a87b91	n/a	5940	n/a	3262	2678	26.21	−19.0
			LongSAGE	SHE11	209 177	6134	0.029		2872	−0.36	3.77

Table 1. *Continued*

Species	Tissue	Additional information	Technology	Library identifier	Sampling depth	Genes identified	Genes identified per sampled tag	Number of genes identified by both methods	Genes identified by this method only	A+T/C+G library bias[a]	A+T/C+G bias of tags/probes of genes missed by this method[b]
Intra-method comparisons											
M.musculus	Heart—atria	Theiler stage 14	LongSAGE	SM006	106 604	4600	0.043	3618	982	2.9	−4.22
	Heart—bulbus cordis	Theiler stage 14		SM005	107 297	4761	0.044		1143	−3.22	6.9
	Liver	Left lobe	MPSS (signature)	GSM32357	1 900 569	2910	0.0015	1580	1330	−17.48	−10.46
		Right lobe		GSM36724	1 810 280	2014	0.001		434	−21.76	−3.48
	Visual cortex, P27	Biological replicate	LongSAGE-Lite	SM073	109 117	4495	0.041	3527	968	−10.28	1.1
				SM040	137 562	4702	0.034		1175	−11.84	4.3
H.sapiens	ES	H1 p54 matrigel	Affymetrix (A/B)	WA01 a26b30	n/a	6285	n/a	5827	458	27.91	−3.47
		H1 p54 matrigel	Affymetrix (A/B)	WA01 a21b22	n/a	6102	n/a		275	25.74	2.01

[a] The A+T/C+G bias of the library is measured in units of the number of standard deviations by which the observed bias deviates from neutral. A positive number indicates that the dataset is A+T rich relative to an unbiased sample. Within each pair of experiments the A+T rich dataset is always given first.

[b] This column provides the A+T/C+G bias of the tags of genes that were not observed. Consistently, the A+T rich dataset misses C+G rich tags and vice-versa.

versus 2 230 467 MPSS tags). For both comparisons, the number of genes observed in common, 1425 genes for the first comparison and 2924 for the second comparison, were substantially reduced from the total number observed in an individual experiment, indicating that a large fraction of genes are observed in only one experiment in each comparison.

Further comparisons are provided in Table 1. Interestingly, MPSS [classic; (5)] generally identified more genes than MPSS [signature; (4)]. The numbers of genes identified by Affymetrix and LongSAGE are comparable. From Table 1 we see that, in general, the number of genes shared by a pair of experiments is only somewhat larger and often smaller than the sum total of unique genes identified by each experiment in a comparison.

These comparisons, though limited in number, demonstrated that there were differences in the number of transcripts identified using each of the methods that could not be explained by sampling depth alone. Our initial survey of transcripts observed by one method and missed by the other implicated a bias in detection sensitivity, perhaps related to the G+C content of the transcripts. Therefore, we sought to assess the G+C content bias for each experiment.

Table 2. The number of NM RefSeq genes that can be identified by each method

Experiment	Species	Number of genes that can be identified
LongSAGE (NlaIII)	Human	14 129
MPSS (DpnII)	Human	13 555
Affymetrix U133A/B	Human	11 700 (14 186 probe sets)
Affymetrix U133 A	Human	11 193 (13 173 probe sets)
LongSAGE (NlaIII)	Mouse	15 054
MPSS (DpnII)	Mouse	14 685
MPSS (DpnII)	*Arabidopsis*	22 381

For SAGE and MPPS, unique mappings are required.

Assessment of G+C content bias and detection sensitivity

The G+C detection sensitivity (G+C DS; Materials and Methods) was calculated for a series of gene expression profiling experiments. These were 67 Signature MPSS *Mus musculus* experiments [from http://www.ncbi.nlm.nih.gov/projects/geo/info/mouse-trans.html and downloaded from the Gene Expression Omnibus (28)], 83 LongSAGE *M.musculus* experiments (29), 21 LongSAGELite *M.musculus* experiments (29), 5 Classic MPSS *Arabidopsis thaliana* experiments (30), 12 Signature MPSS *A.thaliana* experiments (30), 28 Affymetrix *Homo sapiens* experiments (from www.transcriptomes.org), 16 LongSAGE *H.sapiens* experiments (from www.transcriptomes.org) and 993 Affymetrix *H.sapiens* experiments [from the Gene Expression Omnibus (28)]. The G+C DS is a measure of the deviation of G+C content of observed tags or probe sets from neutral and indicates increased sensitivity to G+C or A+T rich tags or probe sets.

Figure 1 shows the distribution of G+C DS among the experiments comprising each experimental group. The mean and standard deviation of G+C DS for each series of experiments is shown in Table 3. The LongSAGE data shows the least bias with a G+C DS of 0.15 ± 3.75. LongSAGE-Lite has some bias in detection sensitivity

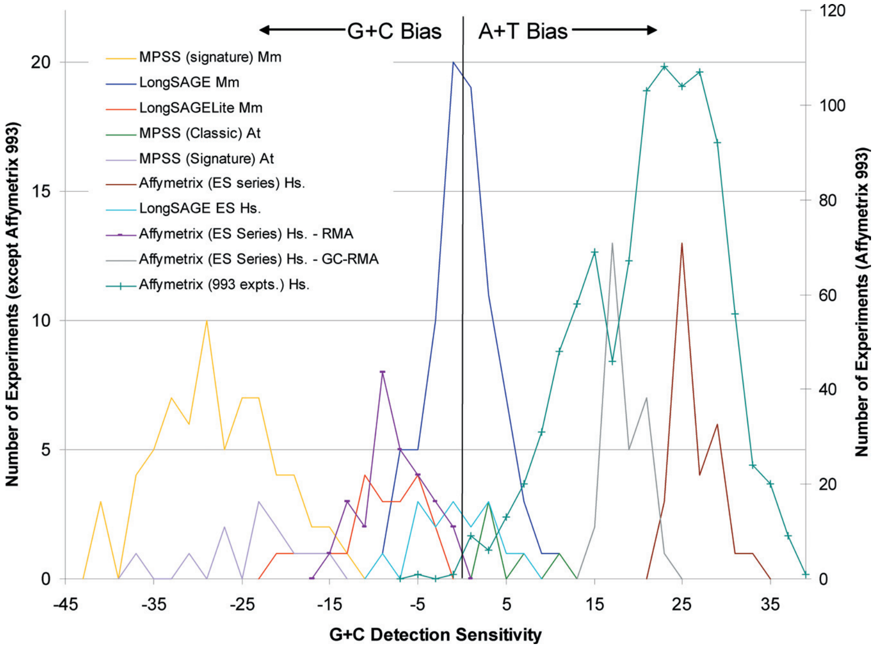

Figure 1. Histogram showing the distribution of bias among five experimental methods. The A+T/C+G bias of an individual experiment is measured in units of the number of standard deviations by which the observed bias deviates from the expected bias. The biases of the individual experiments comprising each series are plotted as a histogram. The position of the peak and the width of the distribution are different for each method and illustrate differences in systematic and random error with respect to A+T/G+C bias.

Table 3. Mean and standard deviation of the distributions of G+C DS for each experimental series (Figure 1)

Experiment series	Number of experiments	Mean G+C DS	Standard deviation of G+C DS	Number of replicates required to achieve a standard error in the mean of 1.0
MPSS (Signature) Mus.	67	−27.76	6.48	42
LongSAGE *Mus.*	83	0.15	3.75	15
LongSAGELite Mus.	21	−9.48	4.89	24
MPSS (Classic) Arab.	5	5.48	2.98	9
MPSS (Signature) Arab.	12	−23.7	5.79	35
Affymetrix (ES series) U133A/B	28	26.36	2.32	6
Affymetrix (ES series) U133A RMA	28	−7.56	3.54	13
Affymetrix (ES series) U133A GC-RMA	28	18.35	1.96	4
Affymetrix (993 expts.) U133A	993	21.43	7.62	59
LongSAGE Human ES	16	−0.57	4.07	17

towards G+C rich tags (−9.48 ± 4.89). For both *A.thaliana* and *M.musculus*, signature MPSS shows a strong G+C bias in detection sensitivity, although, for *A.thaliana*, classical MPSS shows a smaller bias towards A+T. This difference between the two MPSS methods may occur as a result of differences in the method of preparation of the samples for sequencing (30). If so, it is possible that the strong G+C bias in signature MPSS data arises in that step and not during sequencing itself. The Affymetrix data demonstrate a preferential detection of A+T rich probe sets when processed by GC-RMA and MAS 5.0. The GC-RMA software (31) uses the G+C content of probes to correct the gene expression level and Figure 1 demonstrates that Affymetrix data processed using GC-RMA has only ∼70% of the bias of MAS 5.0 processed data. Unlike GC-RMA, the RMA software (32) does not correct the data for the G+C content of probes. It introduces less bias than either MAS 5.0 or GC-RMA and the bias is towards G+C rich probesets.

To ensure that these results were not due solely to rare transcripts, we repeated the experiment restricting the analysis to transcripts observed at a level of 100 tags per million or greater for SAGE and MPSS (Supplementary Figure 1). This reduced the number of transcripts to ∼17% of the original number. The smaller number of transcripts led to a larger standard deviation and hence to a smaller, measured bias for all experiments. The distribution of G+C DS does not change and therefore, we conclude that the G+C DS is affecting transcripts at all expression levels for these technologies. The same experiment was repeated for the Affymetrix GeneChip data restricting the analysis to transcripts observed with a normalized signal intensity of 100 or greater (Supplementary Figures 1 and 2), a cutoff that reduces the number of transcripts detected to ∼15% of the original number. The mean bias is largely reduced (Supplementary Figure 1). We conclude that the bias in the Affymetrix GeneChip data are restricted to genes expressed at lower levels.

The observed differences in G+C DS for the different experimental platforms and conditions raised the possibility that these differences might explain why some genes were observed (those with tags or probe sets with favourable G+C contents relative to the G+C DS of the experimental method) while others were not (those with tags or probe sets with an unfavourable G+C content relative to the G+C DS of the experimental method). Having established that differences in G+C DS existed, we turned our attention to evaluating the impact of these differences on data interpretation and identification of expressed genes.

Assessment of the impact of G+C detection sensitivity

To assess the impact of G+C DS on gene identification, we analysed pairs of gene expression profiling experiments performed on similar samples. Where possible, we utilized pairs of experiments that constituted biological replicates, or even better, experiments performed on the same RNA sample. In a few cases, where a biological replicate was not available, it was necessary to substitute a closely matched experiment (e.g. mouse heart atrium versus mouse heart bulbus cordis at the same developmental stage). Though the lack of replicate experiments could be perceived as a limitation of our analysis, a strong, consistent pattern is demonstrated that strengthens our confidence in the observations.

Comparison of the experimental results allowed us to identify genes with detected expression using one experimental method, but not detected by the second experimental method. We refer to the undetected genes as ‘missed genes’. For intra-method comparisons, missed genes were simply those that the second experimental method failed to detect. For inter-method comparisons, the initial list of missed genes were filtered to remove those genes that the second method was unable to detect because it lacked the probe set or tag. The tags or probe sets (using the second experimental method’s tags or probe sets), corresponding to the missed genes were identified.

Using the same approach as before, we evaluated the G+C DS of the probe sets or tags associated with the missed genes (Table 1, last column). The results consistently demonstrated that if one experimental method had a more negative G+C DS relative to the second experimental method, then the tags or probe sets of genes missed by the first experiment would be A+T rich (have a positive G+C DS) and those missed by the second experiment would be G+C rich (have a negative G+C DS). For example, the first matched pair of libraries is a comparison between LongSAGE and MPSS liver libraries. The data from the LongSAGE library with a G+C DS of −4.31 are A+T rich relative to the data from the MPSS library which have a G+C DS of −18.91. Both are G+C rich relative to zero. The NlaIII tags of genes missed by LongSAGE are G+C rich (−0.88 G+C DS) and the DpnII tags of genes missed by MPSS are A+T rich (8.7 G+C DS). This result is reproduced consistently for each matched pair for both inter- and intra-method comparisons. This confirms our earlier hypothesis that differences in G+C DS results in different genes observed.

Our general interpretation of these results is that both experimental methods correctly identify tags or probe sets representing transcripts in the sample; however, each method exhibits a different sensitivity to the G+C content of those transcript regions detected by it. Within each experimental method, there is a range of G+C DS, perhaps related to differences in experimental conditions. Hence, the variation in G+C DS arises from two sources: a method dependant bias (mean G+C DS; Table 3) and a method dependent variability (standard deviation of G+C DS; Table 3). The method dependant variability is related to the reproducibility of a particular method.

Looking at the LongSAGE ES Hs and Affymetrix ES Hs series (Figure 1), we see that there is variability in G+C DS for replicate samples from the same tissue and that the mean bias is similar to the mean bias of the larger distributions when they are compared to LongSAGE Mm and Affymetrix Hs series, respectively. This lends further evidence to support the assertion that the observed variance in G+C DS is a product of the technology being used to measure gene expression levels and not reflective of the real gene expression levels. The width of the larger Affymetrix series is much wider than the Affymetrix ES series. This difference is addressed in the next section.

The data in Table 1 provide additional confirmation of the G+C bias in the Affymetrix data. The comparison of LongSAGE and Affymetrix demonstrates that genes observed by LongSAGE and missed by Affymetrix have G+C rich probe sets. This orthogonal observation support the notion that the A+T bias in the Affymetrix MAS processed data reduces sensitivity of gene expression detection for genes with G+C rich probe sets.

Variability of affymetrix results among experimental series

Figure 1 and Table 3 show that the Affymetrix results from www.transcriptomes.org, though encapsulated within the distribution of the larger group of 993 experiments, have a much smaller standard deviation than the larger group. The ES Affymetrix data were generated by a single laboratory; hence, suspecting that different laboratories, utilizing different scanners, protocols and batches of chips, would produce data with a greater variation in G+C DS than a single lab, we divided the data into groups of experimental series as specified by GEO. Each of the chosen experimental series comprised of repeat gene expression experiments performed on the same tissue, but extracted from different individuals with the exception of GSE1296 which was comprised of gene expression experiments performed on the same tissue processed using different collection protocols.

Figure 2 provides a representative sample of the series and demonstrates that the results from each series have their own distribution. Although, there are limited data points in the individual series, the data appear to confirm our hypothesis that the different laboratories produce data with different levels of bias. The standard deviations of these distributions are smaller and hence experiments performed as part of the same series will have a higher degree of reproducibility than when compared between series. Therefore, as demonstrated previously by others (33) we expect the results of replicate Affymetrix experiments performed on the same material at different laboratories to have greater variability than those performed at the same laboratory.

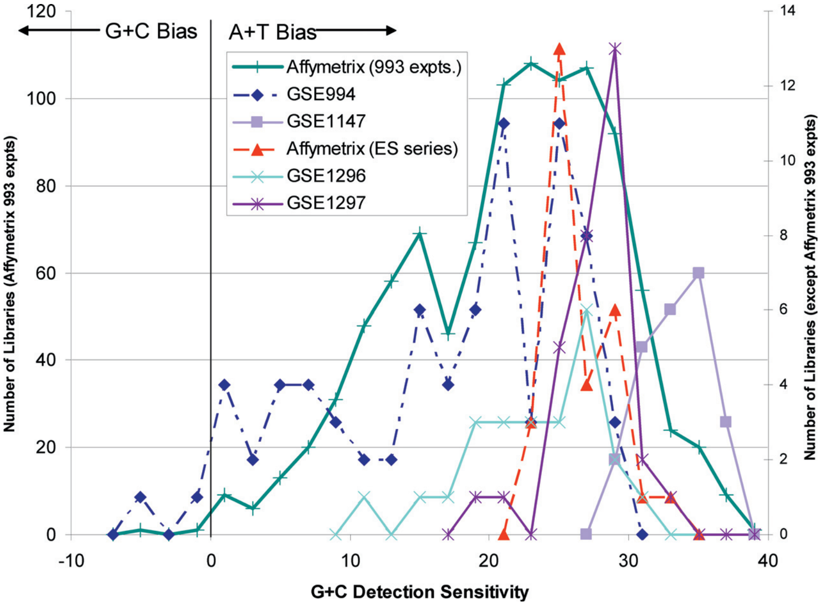

Figure 2. Histogram showing the distribution of bias among individual series of Affymetrix experiments. For each series, the position of the peak and the width of the distribution are different. The figure illustrates individual laboratories have their own bias and variation in bias. Thus, combining experiments from different laboratories leads to a greater width in the summed distribution (Affymetrix 993 experiments). The series identifiers in the figure (e.g. GSE994) are GEO dataset identifiers.

DISCUSSION

The mean and standard deviation of the G+C DS (Table 3) can be discussed in terms of accuracy and reproducibility, respectively or systematic error and random error, respectively. A method with a larger standard deviation in observed G+C DS has more variation in experimental results. This variation can be corrected through the use of replicates, though methods with a larger standard deviation will require additional replicates to obtain a more precise measure of gene expression levels. However, replicates will not improve the accuracy of the measurement if there is a systematic error underlying the experimental method as measured by the deviation of the mean G+C DS from zero. Hence, the better method of measuring gene expression has both a G+C DS mean and standard deviation of zero. In reality, each method has its own strengths and weaknesses. While LongSAGE has the lowest mean G+C DS, replicates are relatively expensive at this time. Affymetrix has a similar standard deviation in G+C DS to LongSAGE and replicates are less expensive. For Affymetrix, the mean G+C DS is strongly related to the data processing method (MAS, RMA and GC-RMA). Although LongSAGELite demonstrates a G+C DS bias, the method allows the creation of SAGE libraries from nanogram quantities of RNA. Signature MPSS shows the strongest G+C DS bias, but allows deep profiling of RNA samples. In contrast to the results for Signature MPSS, Classical MPSS shows a much smaller bias.

The Affymetrix data show different biases with each of the data processing methods. RMA does not utilize the probeset sequence to correct the gene expression levels and that data show a G+C bias. GC-RMA attempts to correct for stronger hybridization to G+C rich probesets. Our results indicate that GC-RMA may be overcorrecting the data and an intermediate correction may yield better results.

Series of experiments performed across different platforms often utilize a reference experiment against which other experiments are normalized. This strategy is an attempt at removing the measurement biases inherent in each platform. The effectiveness of this strategy is limited by experimental realities: unknown non-linear signal response, signal saturation and background noise. For example, if a gene is observed at levels comparable to the background noise, its true expression level cannot be recovered. We have shown that the difference in detected genes is extremely large. For example, in Affymetrix MAS processed data and LongSAGE comparisons only approximately one-half of the observed genes are common to both experiments. These differences cannot be resolved by normalization and we present, for the first time, an explanation for these differences.

This study has also demonstrated that the impact of the G+C bias is not merely restricted to detection but affects the gene expression level itself. Hence, one cannot infer that gene A is more highly expressed than gene B from the gene expression intensity value assigned to each gene. While relative expression changes between individual experiments within a single platform are valid, we have shown, for the first time, that certain genes will be harder to observe (i.e. have lower expression intensity values) in certain platforms (e.g. MPSS is poor at detecting genes with high A+T content probe sets).

We have demonstrated that differences in sensitivity directly impact the genes that are observed. This observation presents an explanation and means of quantifying one source of experimental error in gene expression experiments. Of the methods studied, LongSAGE performs the best, while MPSS shows a strong bias to G+C rich tags and Affymetrix shows different biases as a function of the data processing method. The Affymetrix bias is partially corrected by the RMA software and a similar approach may improve MPSS data. Still, large experimental biases remain and an understanding of them is necessary for correct interpretation of gene expression data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

M.A.M. and S.J.M.J. are Scholars of the Michael Smith Foundation for Health Research. M.A.M. is an NCIC Terry Fox Young Investigator. O.L.G. is supported by the Canadian Institutes of Health Research (CIHR). We thank Narayan K. Bhat (SAIC) for providing the RNA for the construction of the LongSAGE liver library matched to the MPSS library. We thank Tarun Nayer for data processing support. Funding for this study was provided by Genome Canada.

Conflict of interest statement. None declared.

REFERENCES

1. Velculescu,V.E., Vogelstein,B. and Kinzler,K.W. (2000) Analysing uncharted transcriptomes with SAGE. *Trends Genet*., **16**, 423–425.
2. Saha,S., Sparks,A.B., Rago,C., Akmaev,V., Wang,C.J., Vogelstein,B., Kinzler,K.W. and Velculescu,V.E. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol*., **20**, 508–512.
3. Boon,K., Osorio,E.C., Greenhut,S.F., Schaefer,C.F., Shoemaker,J., Polyak,K., Morin,P.J., Buetow,K.H., Strausberg,R.L., De Souza,S.J. *et al*. (2002) An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11287–11292.
4. Brenner,S., Johnson,M., Bridgham,J., Golda,G., Lloyd,D.H., Johnson,D., Luo,S., McCurdy,S., Foy,M., Ewan,M. *et al*. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol*., **18**, 630–634.
5. Meyers,B.C., Vu,T.H., Tej,S.S., Ghazal,H., Matvienko,M., Agrawal,V., Ning,J. and Haudenschild,C.D. (2004) Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing. *Nat. Biotechnol*., **22**, 1006–1011.
6. Cheng,J., Kapranov,P., Drenkow,J., Dike,S., Brubaker,S., Patel,S., Long,J., Stern,D., Tammana,H., Helt,G. *et al*. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–1154.
7. Harbers,M. and Carninci,P. (2005) Tag-based approaches for transcriptome research and genome annotation. *Nature Meth*., **2**, 495–502.
8. Peters,D.G., Kassam,A.B., Yonas,H., O'Hare,E.H., Ferrell,R.E. and Brufsky,A.M. (1999) Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite. *Nucleic Acids Res*., **27**, e39.
9. Matsumura,H., Reich,S., Ito,A., Saitoh,H., Kamoun,S., Winter,P., Kahl,G., Reuter,M., Kruger,D.H. and Terauchi,R. (2003) Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proc. Natl Acad. Sci. USA*, **100**, 15718–15723.
10. Audic,S. and Claverie,J.M. (1997) The significance of digital gene expression profiles. *Genome Res*., **7**, 986–995.
11. Mah,N., Thelin,A., Lu,T., Nikolaus,S., Kuhbacher,T., Gurbuz,Y., Eickhoff,H., Kloppel,G., Lehrach,H., Mellgard,B. *et al*. (2004) A

comparison of oligonucleotide and cDNA-based microarray systems. *Physiol. Genomics.*, **16**, 361–370.

12. Lu,J., Lal,A., Merriman,B., Nelson,S. and Riggins,G. (2004) A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. *Genomics*, **84**, 631–636.
13. Gnatenko,D.V., Dunn,J.J., McCorkle,S.R., Weissmann,D., Perrotta,P.L. and Bahou,W.F. (2003) Transcript profiling of human platelets using microarray and serial analysis of gene expression. *Blood*, **101**, 2285–2293.
14. Griffith,O.L., Pleasance,E.D., Fulton,D.L., Oveisi,M., Ester,M., Siddiqui,A.S. and Jones,S.J.M. (2005) Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses. *Genomics*, **86**, 476–488.
15. Huminiecki,L., Lloyd,A.T. and Wolfe,K.H. (2003) Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics*, **4**, 31.
16. Jarvinen,A.K., Hautaniemi,S., Edgren,H., Auvinen,P., Saarela,J., Kallioniemi,O.P. and Monni,O. (2004) Are data from different gene expression microarray platforms comparable? *Genomics*, **83**, 1164–1168.
17. Rogojina,A.T., Orr,W.E., Song,B.K. and Geisert,E.E.,Jr (2003) Comparing the use of Affymetrix to spotted oligonucleotide microarrays using two retinal pigment epithelium cell lines. *Mol. Vis.*, **9**, 482–496.
18. Asyali,M.H. and Alci,M. (2005) Reliability analysis of microarray data using fuzzy c-means and normal mixture modeling based classification methods. *Bioinformatics*, **21**, 644–649.
19. Mecham,B.H., Klus,G.T., Strovel,J., Augustus,M., Byrne,D., Bozso,P., Wetmore,D.Z., Mariani,T.J., Kohane,I.S. and Szallasi,Z. (2004) Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res.*, **32**, e74.
20. Carter,S.L., Eklund,A.C., Mecham,B.H., Kohane,I.S. and Szallasi,Z. (2005) Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics*, **6**, 107.
21. Kluger,Y., Yu,H., Qian,J. and Gerstein,M. (2003) Relationship between gene co-expression and probe localization on microarray slides. *BMC Genomics*, **4**, 49.
22. Southern,E., Mir,K. and Shchepinov,M. (1999) Molecular interactions on microarrays. *Nature Genet.*, **21**, 5–9.
23. Kuo,W.P., Jenssen,T.K., Butte,A.J., Ohno-Machado,L. and Kohane,I.S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.
24. Margulies,E.H., Kardia,S.L. and Innis,J.W. (2001) Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res.*, **29**, E60–E60.
25. Dinel,S., Bolduc,C., Belleau,P., Boivin,A., Yoshioka,M., Calvo,E., Piedboeuf,B., Snyder,E.E., Labrie,F. and St-Amand,J. (2005) Reproducibility, bioinformatic analysis and power of the SAGE method to evaluate changes in transcriptome. *Nucleic Acids Res.*, **33**, e26.
26. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003) NCBI Reference Sequence project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.
27. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al.* (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
28. Barrett,T., Suzek,T.O., Troup,D.B., Wilhite,S.E., Ngau,W.C., Ledoux,P., Rudnev,D., Lash,A.E., Fujibuchi,W. and Edgar,R. (2005) NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res.*, **33**, D562–D566.
29. Siddiqui,A.S., Khattra,J., Delaney,A.D., Zhao,Y., Astell,C., Asano,J., Babakaiff,R., Barber,S., Beland,J., Bohacec,S. *et al.* (2005) A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc. Natl Acad. Sci. USA*, **102**, 18485–18490.
30. Meyers,B.C., Tej,S.S., Vu,T.H., Haudenschild,C.D., Agrawal,V., Edberg,S.B., Ghazal,H. and Decola,S. (2004) The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. *Genome Res.*, **14**, 1641–1653.
31. Wu,Z. and Irizarry,R.A. (2005) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J. Comput. Biol.*, **12**, 882–893.
32. Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.
33. Piper,M.D., Daran-Lapujade,P., Bro,C., Regenberg,B., Knudsen,S., Nielsen,J. and Pronk,J.T. (2002) Reproducibility of oligonucleotide microarray transcriptome analyses. An interlaboratory comparison using chemostat cultures of *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **277**, 37001–37008.