Affymetrix probe-set remapping and probe-level filtering leads to dramatic improvements in gene expression measurement accuracy

Mariano Javier Alvarez[a,*], Pavel Sumazin[a,*] and Andrea Califano[a,b]

(a) Joint Centers for Systems Biology, Columbia University, New York, NY, USA.

(b) Department of Biomedical Informatics and Institute for Cancer Genetics and Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA

*These authors contributed equally

1. ABSTRACT

Motivation: Affymetrix GeneChip microarrays are widely used to measure mRNA concentration, but their probe annotation and probe-set construction remain a source of inaccuracy that greatly impacts downstream analysis. We sought to improve gene expression measurement accuracy in large datasets through analysis of coordinated probe-level measurements followed by probe re-annotation and filtering.

Results: Probe remapping and probe-set reconstruction using up-to-date gene annotation has been shown to improve Affymetrix GeneChip microarray accuracy. For commonly used arrays, this re-annotation process affects most probe sets and leads to the loss of many individual probe measurements. We show that probe-set remapping can be significantly improved by analyzing statistical dependencies at the probe level and across large and context-specific datasets, leading to more accurate probe annotation and probe-set construction. In the absence of such analysis, practitioners are reduced to the more ad-hoc probe-set filtering, which leads to loss of informative probe reads and inclusion of data from poor-quality probes. Large-scale expression profile datasets, with forty or more samples profiled on the same platform are now commonplace and allow for more robust probe annotation and filtering. We describe CleanProbeSets, a novel probe-set construction method that leads to significant improvement in gene expression measurement accuracy, leading to higher concordance between analyses both on the same platform and across platforms. With a low and an easy-to-estimate false-positive rate, CleanProbeSets avoids inclusion of flawed probes, accounts for dependence between probes, and addresses measurement variability due to transcript isoforms. CleanProbeSets will be valuable for the analysis of future and existing datasets.

Availability: An implementation of CleanProbeSets is available by request from the authors.

Contact: califano@c2b2.columbia.edu

1. Introduction

Microarray-based genome-wide expression profiling is a powerful and a widely-used tool for studying cell phenotype at the molecular level. Since their inception, however, microarray gene-expression profiling accuracy has been suspect due to poor reproducibility across experiments and across platforms (Tan, Downey et al. 2003; Lossos, Czerwinski et al. 2004; Mecham, Klus et al. 2004). Numerous studies have attempted to improve microarray accuracy by improving analytical and interpretive data processing, and normalization and filtering methods (Page and Coulibaly 2008). Challenges addressed by these methods include individual probe-read quality

(Eisen 1999; Hubbell, Liu et al. 2002 ; Zhang, Miles et al. 2003), the interpretation of probe reads across probe sets (Eisen, Spellman et al. 1998; Irizarry, Bolstad et al. 2003; Gentleman, Carey et al. 2004), and the aggregation of probes into probe sets (Gentleman, Carey et al. 2004; Liu, Zeeberg et al. 2007). Here, we focus on remapping and filtering individual probes in Affymetrix GeneChip microarrays, the most popular genome-wide expression profiling platform. Whether as standalone or within plate sets, Affymetrix GeneChip microarrays are increasingly used for high-volume expression profiling studies (Basso, Margolin et al. 2005; Lamb, Crawford et al. 2006; Cancer Genome Atlas Research Network 2008).

Gautier et al. and Carter et al. were among the first to realize that incorrect Affymetrix probe annotations were major contributors to inconsistencies between repeated experiments across platforms (Gautier, Moller et al. 2004; Carter, Eklund et al. 2005). By redefining probe sets according to probe matches against cDNA libraries they were able to substantially improve cross-platform consistency. Later studies repeatedly showed that re-aggregating and pruning probes based on probe-sequence alignment to up-to-date genome annotation significantly improved expression-profiling accuracy as measured in terms of cross-platform consistency (Gautier, Moller et al. 2004; Dai, Wang et al. 2005; Harbig, Sprinkle et al. 2005). Challenges faced by annotation methods include addressing probes that do not match any transcripts and those that match multiple genes. The most common approach is to discard these problematic probes, and annotate probes that match multiple transcripts of the same gene (Dai, Wang et al. 2005; Liu, Zeeberg et al. 2007). We followed this approach, which, as others have observed, may lead to a loss of up to 71% of the probes in Affymetrix GeneChip microarrays (De Leeuw, Rauwerda et al. 2008).

Probe remapping using up-to-date genomic annotation has been repeatedly shown to improve microarray accuracy, but many challenges in probe-set redesign remain unresolved. Specifically, tissue-specificity of transcript isoforms, post-transcriptional modifications, and allelic variability and mutations, can affect probe accuracy in a context-specific fashion. Because of their complex and poorly understood nature, accounting for these features at the microarray design stage is a prohibitive task. Interestingly, we show that when relatively large gene-expression-profile datasets are available, these issues can be naturally corrected using a relatively simple statistical procedure that identifies highly correlated probe clusters within groups of probes that match the same mRNA targets. By defining probe sets based on information content and probe homogeneity across experiments, we minimized the effect of poorly performing probes in a cell-context specific way and excluded uninformative probes. The latter include probes with poor variability across samples as well as probes that contain special features that lead to cross- or non-specific hybridization.

By harvesting the power of high-volume microarray expression experiments, where forty or more microarray experiments are carried out using the same platform, we showed that probe-set re-annotation and pruning can dramatically improve accuracy, leading to considerable improvements to downstream expression-based analysis. We compared microarray expression consistency when using Affymetrix annotation, AffyProbeMiner annotation (Liu, Zeeberg et al. 2007), and our CleanProbeSets annotation. AffyProbeMiner is a recent effort focused on remapping and re-aggregating probes in Affymetrix GeneChips microarrays. We measured inter- and intra-microarray consistency by computing the correlation between repeated gene profiling

experiments using U133A (Su, Wiltshire et al. 2004), and by comparing gene sets identified as differentially down-regulated in centroblasts relative to naïve B cells using U95A and U133plus2 platforms. We show that annotation by CleanProbeSets dramatically improved the consistency across experiments and platforms, suggesting that re-annotated probe sets will profoundly improve the accuracy of downstream analysis.

2. Methods

2.1 Expression profiles

Gene expression data include 102 B-cell samples profiled on U95A (Basso, Margolin et al. 2005), 152 B-cell samples profiled on U95Av2 (Basso, Margolin et al. 2005), 200 B-cell samples profiled on U133plus2 (unpublished), and 60 samples from 30 human tissues profiled on U133A chips (Su, Wiltshire et al. 2004). B-cell samples were obtained from Gene Expression Omnibus database (Edgar, Domrachev et al. 2002) accession GSE2350. U133A gene expression profiles were obtained from GeneAtlasV2 (Su, Wiltshire et al. 2004), and included analysis of RNA samples from 30 tissues obtained from Clontech and hybridized in duplicates. A Bioconductor-based (Gentleman, Carey et al. 2004) implementation of MAS5 (Irizarry, Bolstad et al. 2003) was used to quantitatively estimate and normalize the intensity levels of probe sets.

2.2 Probe mapping to RefSeq genes

Probe sequences were mapped to the RefSeq sequence database (Pruitt and Maglott 2001) dating December 11th 2008 using ZOOM (Lin, Zhang et al. 2008) and allowing for at most one mismatch per probe (each probe sequence matched at least 24 transcript positions). We matched against the positive orientation of RefSeq transcripts only. Probes that matched multiple genes were discarded, and each location in each matching transcript was annotated.

2.3 Generation of clean probe-sets

Based on probe mapping to RefSeq transcripts and corresponding genes, we constructed gene-focused probe sets after quality control for individual probes, clustering correlated probes, and testing probe-set reliability. Probe sets were used to create CDF files for assigning quantitative probe-set intensity by MAS5.

Probe reliability. We first established the reliability of each individual probe based on its correlation to other probes mapped to the same gene (neighbors) across microarray experiments. The readout obtained from any given probe is informative only if it is significantly correlated with other probes mapping to the same gene. We set the consistency score of each probe to the 90th percentile of computed Spearman correlation coefficients across its neighbors. Before measuring correlation, probe readouts were quantile normalized to abstract away correlation among probes generated by inter-sample systematic bias. Statistical significance was estimated on a gene-per-gene basis using a null distribution generated by computing the correlation between probes mapping to the gene and 1,000 probes selected uniformly at random. Probes with consistency score corresponding to $p>0.01$ were eliminated (Fig. 1A).

Probe clusters. Neighboring probes that map to isoforms that are differentiated by alternative splicing, RNA editing, or non-representative hybridization can produce readouts of different molecular species leading to poor quantitative intensity evaluation for the probe-set. To account

for RNA isoforms, we performed a non-supervised, single-linkage hierarchical clustering of the probes using Spearman correlation coefficient (rho) as a distance measure. First, clusters were formed by iteratively breaking dendrogram edges that were significantly longer than the remainder of the edges in each level according to a one-tail t-test threshold of $p < 10^{-10}$ (Fig. 1B and C). Then, we iteratively merged cluster pairs with distance significance greater than 0.001, were distance between clusters was defined as the distance between the closest elements across clusters, and significance was estimated using a null distribution of 1,000,000 distances between randomly selected probe pairs. For illustration, Figure 1D depicts the two probe sets identified for MAX across three of its known isoforms.

Probe-set reliability. Low information probe sets, composed of few or dependent probes were eliminated to reduce the false discovery rate (FDR). Overlapping probes account for 87.2% of the U95av2 remapped probes and 66.1% of the U133plus2 remapped probes, and they are affected by systematic bias resulting from common technical artifacts and cross-hybridization. These biases artificially improve pairwise correlations across expression vectors and can be estimated by conditioning on probe-overlap size. Figure 1E pictorially demonstrates that pairwise Spearman correlations between probes can be described as an exponential function of their overlap size; we found this significant behavior to hold true across platforms and experiments. To assign consistency scores for probe sets, we derived a score for computing the contribution of each probe according to their overlap with their upstream neighbor. Each probe contributed at most one point to the total score, and the contribution of a probe that overlaps another upstream probe was set according to $s(x)$:

$$s(x) = 1 - \frac{f(x) - k}{1 - k} \quad ; \quad f(x) = c + b \cdot e^{a \cdot x} \quad ,$$

where x is the shortest distance between probe starting positions across isoforms (position shift in Figure 1E); a, b and c are estimated by fitting $f(x)$ to pairwise Spearman correlations for $1 \leq x \leq 24$; and $k = E(f(x))$ for $25 \leq x \leq 50$. We estimated the probe-set FDR for each consistency score using permutation testing, where each CleanProbeSets probe set constructed after permuting experiment labels for each probe was considered a false positive detection. For all experiments reported in this study, we set the minimum probe-set consistency score to $s(x) \geq 3$.

2.4 Differential expression

We used a non-parametric U-test to identify down-regulated genes in centroblasts B cells relative to naïve B cells. Comparisons were made using expression profiles from five biological replicates in each cell type. To identify a representative set of down-regulated genes per platform and annotation method, we used a z-value cutoff of 2.33 ($p<0.01$) for calling differential expression together with a 1.5 fold change requirement. The fold change requirement was used in order to correct for the high expected false discovery rate of this non-parametric 5×5 test across thousands of probe sets (see Figure 5). The 1.5-fold increase from naïve to centroblasts B cells was based on the average intensities of the probe sets after MAS5 normalization and log_2 transformation. Analysis accuracy was measured using permutation testing repeated 20 times per annotation and platform, where the experimental source labels were shuffled for each probe set.

3 Results and discussion

Due to poor mapping to RefSeq genes, poor reliability or low gene expression across experiments, CleanProbeSets discards most of the individual probe measurements in each Affymetrix gene chip. It retains probe-set representations for approximately half of the genes that were originally probed on the array. The loss of data is offset by dramatic improvements in measurement accuracy for the remaining probed genes. Our experiments suggest that the vast majority of data discarded is at best uninformative for downstream analysis, and it may be misleading and reduce its accuracy.

3.1 Probe sets for U95Av2 and U133plus2

CleanProbeSets rejects probes due to poor matches to RefSeq transcripts (see Section 2.2) and poor fit to a consistent probe set (see Section 2.3). Table 1 describes the total number of probes available, the number of discarded probes, and the number of probe sets produced by CleanProbeSets when measuring B-cell expression using U95Av2 and U133plus2. Poor gene matching, due to no- or multiple-gene homology, resulted in a loss of 23% and 47% of the probes for U95Av2 and U133plus2 (Remap in Table 1). Probe and probe-set reliability analysis further discarded 48% and 67% of the RefSeq–mapped probes in U95Av2 and U133plus2. Of the probes discarded due to reliability issues, 55% and 58% were originally mapped to genes containing no consistent probes on U95Av2 and U133plus2 platforms. Over 40% of the discarded probes were mapped to isoforms with consistent probe sets, suggesting that individual probes mapped to expressed genes can be inconsistent due to technical bias (Figure 2). CleanProbeSets eliminated most of the original probes, and represented approximately half of the probed genes by at least one probe set (see Table 1).

The relationship between expression intensity and probe-set reliability depicted in Figure 1G suggests that while low-intensity probe sets are significantly more likely to be eliminated, low intensity on its own is not a sufficient requirement for rejection: many high-intensity probe sets get discarded and low intensity probe sets kept. In addition, we note that imperfectly matching probes (with a single RefSeq homology mismatch) were rejected at a significantly higher rate than perfectly matching probes, and they account for a small portion of the consistent probes. To estimate the sample-size effect on CleanProbeSets analysis we randomly selected subsets from the 152 U95Av2 and 200 U133plus2 microarray experiments in B cells, and estimated the FDR when constructing probe sets with CleanProbeSets; we selected twenty samples per sample size. Results, given in Figure 3 suggest that CleanProbeSets is not effective for analyzing data derived from fewer than 20 microarray experiments. FDR and probe-set sizes showed no significant change for expression sets consisting of 40 or more microarray experiments, suggesting that full statistical power is obtained at this size.

Probe features such as G spots have been shown to disproportionately bias expression measurements (Upton, Langdon et al. 2008). We used DME and motifclass (Smith, Sumazin et al. 2007) to identify patterns that are enriched in sequences of discarded probes relative to sequences of consistent probes. To ensure that discarded probes were truly individually inconsistent and were not discarded due to absent genes, we restricted the study to consistent probe-sets corresponding to genes that had less than 20% probe rejection rates. The most enriched motifs identified were CGGGGG and GGG[G|A][G|C]; both motifs were significantly ($p<0.001$) enriched according to permutation testing, and CGGGGG had sites in 46% of the

inconsistent probes and 20% of the consistent probes. This result suggests that patterns such as G spots are strongly correlated with probe bias, but may not sufficient criteria for probe selection. Twenty percent of the consistent probes included a CGGGGG substring but still showed significant correlation to neighboring G-spot-free probes.

3.2 Agreement across technical replicate experiments

Repeatability of experimental results is one the basic requirements of any technology used for research and product development. We show that experimental repeatability is highly influenced by probe-set annotation quality. We measured Spearman correlation between replicate experiments from GeneAtlasV2 for 30 human tissue samples using Affymetrix, AffyProbeMiner, and CleanProbeSets probe-set annotation. Correlation coefficient distributions are given in Figure 4 and demonstrate that experimental replicates are in better agreement (significantly and dramatically) when using CleanProbeSets probe-set annotation. The CleanProbeSets advantage over AffyProbeMiner is almost entirely due to its ability to eliminate inconsistent probes and construct clean probe sets, and it is not due to an improved probe-sequence mapping to up-to-date genomes. To demonstrate this, we included results taken at an intermediary step of the CleanProbeSets algorithm (Remap-only in Figure 4), where all RefSeq-mapped probes were used for probe-set construction. These results show that in the absence of CleanProbeSets reliability testing, there is no significant difference (at $p<0.05$) between our probe-sequence mapping annotation and AffyProbeMiner annotation. In addition, to demonstrate that probe-set level pruning does not bridge the performance gap between CleanProbeSets and the other annotations, we selected the 1,000 probe sets with highest coefficient of variation (CV) across samples for each annotation and repeated the comparison; CV-based pruning is commonly used to remove poorly-informative probe sets from microarray expression experiments (MAQC 2006). Our results suggest that CleanProbeSets probe sets are significantly more consistent across experiments and that the benefit of its probe-level selection and pruning may not be achieved by probe-set level pruning.

3.3 Cross-platform agreement on differentially expressed genes

Differentially-expressed genes, identified using multiple platforms, are routinely used to quantify cross-platform consistency (Dai, Wang et al. 2005; MAQC 2006). Genes with down-regulated expression during B-cell maturation from naïve to centroblasts B cells may contribute to mature B-cell germinal-center formation. We identified these genes by searching for differentially down-regulated genes in centroblasts B-cell gene-expression profiles relative to naïve B-cell gene-expression profiles. Basso et al. (2005) used Affymetrix geneChip U95A and U95Av2 microarrays to obtain gene expression profiles from five biological replicates of naïve and of centroblasts B cells; the same samples were later analyzed using U133plus2. We measured the consistency of differentially expressed gene sets across platforms, and the FDR of down-regulation calls in each platform under Affymetrix, AffyProbeMiner, and CleanProbeSets probe-set annotation. Figure 5 shows that U133plus2-based analysis was consistently found to be more accurate than U95A-based analysis, and that CleanProbeSets produces significantly and dramatically more accurate intra- and inter-platform results. To identify centroblasts down-regulated genes we used a z-value cutoff of 2.33 ($p<0.01$) for calling down-regulation together with an added 1.5 fold change requirement (see Secion 2.4). Permutation testing estimates for the FDR in the U133plus2 analysis were 10.7%, 5.4% and 3.2% for Affymetrix, AffyProbeMiner, and CleanProbeSets probe-set annotation, respectively. Focusing on genes

probed by both U95A and U133plus2 platforms, we identified 859 and 1234 down-regulated genes in centroblasts by using Affymetrix annotations; 742 and 989 down-regulated genes by using AffyProbeMiner; and 677 and 801 down-regulated genes when using CleanProbeSets. For Affymetrix annotation, 1478 genes were called down-regulated by at least one of the platforms and 615 (41.6%) genes were called down-regulated in both platforms; this ratio improved to 550 of 1181 (46.6%) for AffyProbeMiner, and to 562 of 919 (61.2%) for CleanProbeSets. Note that results are independent of the number of differentially expressed genes identified in each method and each platform; the U133plus2 platform included more probes and more probe sets and was more accurate, while the most accurate results were produced using CleanProbeSets, which defined the fewest probe sets.

4 References

Basso, K., A. A. Margolin, et al. (2005). "Reverse engineering of regulatory networks in human B cells." Nat Genet **37**(4): 382-390.

Cancer Genome Atlas Research Network (2008). "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." Nature **455**(7216): 1061-1068.

Carter, S., A. Eklund, et al. (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**(1): 107.

Dai, M., P. Wang, et al. (2005). "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." Nucl. Acids Res. **33**(20): e175-.

Dai, M., P. Wang, et al. (2005). "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." Nucleic Acids Res **33**(20): e175.

De Leeuw, W., H. Rauwerda, et al. (2008). "Salvaging Affymetrix probes after probe-level re-annotation." BMC Research Notes **1**(1): 66.

Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucl. Acids Res. **30**(1): 207-210.

Eisen, M. B. (1999). ScanAlyze.

Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-14868.

Gautier, L., M. Moller, et al. (2004). "Alternative mapping of probes to genes for Affymetrix chips." BMC Bioinformatics **5**(1): 111.

Gentleman, R. C., V. J. Carey, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biology **5**: R80.

Harbig, J., R. Sprinkle, et al. (2005). "A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array." Nucl. Acids Res. **33**(3): e31-.

Hubbell, E., W. Liu, et al. (2002). "Robust estimators for expression analysis." Bioinformatics **18**(12): 1585-92.

Irizarry, R. A., B. M. Bolstad, et al. (2003). "Summaries of Affymetrix GeneChip probe level data." Nucleic Acids Research **31**.

Lamb, J., E. D. Crawford, et al. (2006). "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease." Science **313**(5795): 1929 - 35.

Lin, H., Z. Zhang, et al. (2008). "ZOOM! Zillions of oligos mapped." Bioinformatics **24**(21): 2431-2437.

Liu, H., B. R. Zeeberg, et al. (2007). "AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets." Bioinformatics **23**(18): 2385-2390.

Lossos, I., D. Czerwinski, et al. (2004). "Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes." N Engl J Med **350**(18): 1828-37

MAQC (2006). "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." Nat Biotech **24**(9): 1151-1161.

Mecham, B., G. Klus, et al. (2004). "Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements." Nucleic Acids Res **32**(9): E74.

Page, G. and I. Coulibaly (2008). "Bioinformatic tools for inferring functional information from plant microarray data: tools for the first steps." Int J Plant Genomics: 147563.

Pruitt, K. D. and D. R. Maglott (2001). "RefSeq and LocusLink: NCBI gene-centered resources." Nucl. Acids Res. **29**(1): 137-140.

Smith, A. D., P. Sumazin, et al. (2007). "Tissue-specific regulatory elements in mammalian promoters." Mol Syst Biol **3**: 73.

Su, A. I., T. Wiltshire, et al. (2004). "A gene atlas of the mouse and human protein-encoding transcriptomes." Proceedings of the National Academy of Sciences of the United States of America **101**(16): 6062-6067.

Tan, P. K., T. J. Downey, et al. (2003). "Evaluation of gene expression measurements from commercial microarray platforms." Nucleic Acids Res **31**: 5676 - 5684.

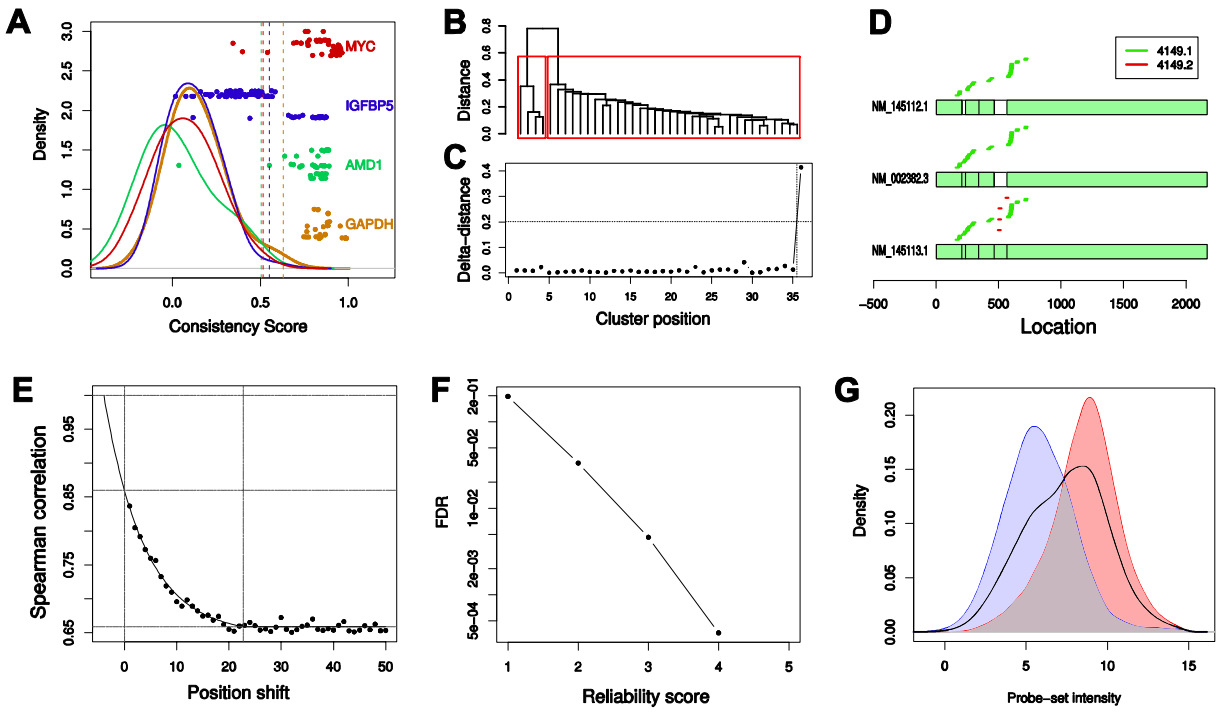
Upton, G. J., W. B. Langdon, et al. (2008). "G-spots cause incorrect expression measurement in Affymetrix microarrays." BMC Genomics **9**: 613.

Zhang, L., M. Miles, et al. (2003). "A model of molecular interactions on short oligonucleotide microarrays." Nat Biotechnol **21**(7): 818-21.

Table 1: Number of probes, probe-sets and genes represented in two popular Affymetrix geneChip microarrays. We report the number of probes, probe-sets and unique entrezIDs for the U95av2 and U133plus2 platforms according to Affymetrix, AffyProbeMiner (APM), Remap, and CleanProbeSets (CPS) annotation. Remap is the first stage of CPS and includes mapping of probe sequences to RefSeq transcripts but no pruning.

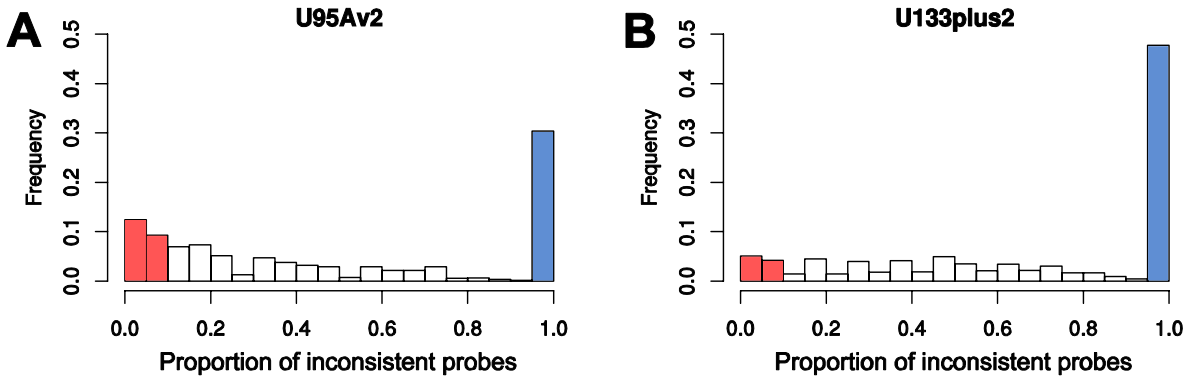
		Affymetrix	APM		Remap		CPS	
U95av2	Probes	201800	163939	(81%)	154485	(77%)	80025	(40%)
	Probe-sets	12625	9130	(72%)	8410	(67%)	4702	(37%)
	EntrezID	8975	8781	(98%)	8410	(94%)	4581	(51%)
U133plus2	Probes	604258	326265	(54%)	318978	(53%)	105449	(17%)
	Probe-sets	54675	19887	(36%)	18303	(33%)	9797	(18%)
	EntrezID	20327	18596	(91%)	18303	(90%)	9564	(47%)

Figure 1: CleanProbeSets algorithm on 152 B-cell samples profiled on U95Av2 chips



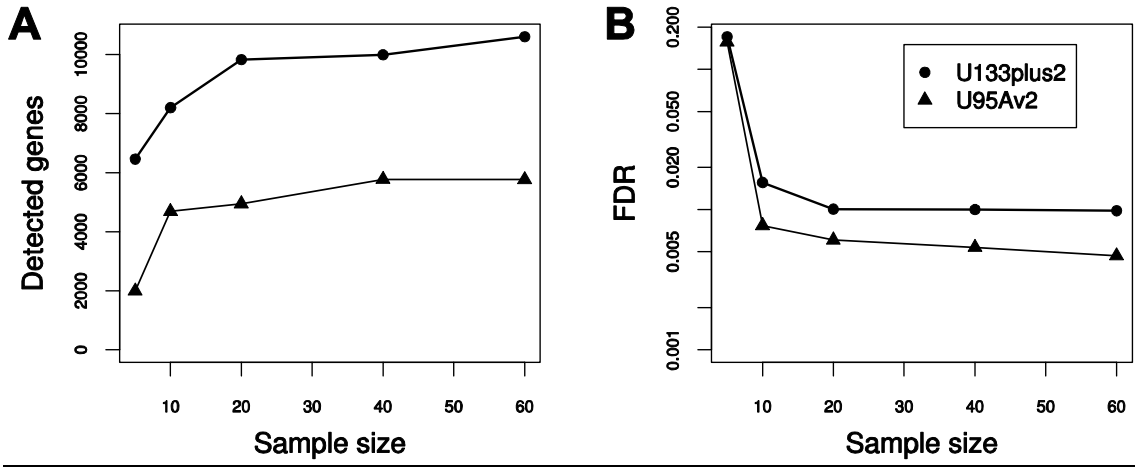
(A) Probe consistency scores for four genes (distinguished by color) and their corresponding null density distributions. Dots represent probes and are plotted according to consistency scores (horizontal axis) and distances from the transcript 3' end (vertical axis). Solid lines depict null density distributions and dotted vertical lines are drawn at their 99 percentile. Consistent probes are to the right of their respective dotted lines. **(B)** Unsupervised hierarchical clustering of probes mapping to MAX (all isoforms); naturally occurring probe clusters are highlight in red. **(C)** Relative distance between each consecutive cluster in the dendrogram in panel B; the right-most point represents the distance between highlighted clusters in panel B. **(D)** Three known isoforms for MAX and the mapping positions of probes belonging to the two probe-sets generated by CleanProbeSets for MAX; probes from probe set 4149.2 are mapped to the splice-variant 5th exon. **(E)** The mean of spearman correlations between overlapping and neighboring probes depends on the distance between them, and it is closely modeled by an exponential function. **(F)** FDR as a function of the reliability score, as estimated by permutation testing; no probe-set with reliability score higher than 4 was identified in permuted sets. **(G)** Probe-set reliability scores are correlated to their MAS5-assigned intensity, as measured before pruning. However, the intersection between distributions for the 4,702 consistent probe sets (red), 3,708 inconsistent probe sets (blue), and all probe sets (black line) suggests that probe-set intensity is not a perfect predictor for probe set reliability.

Figure 2: Inconsistent probe distribution across probe sets



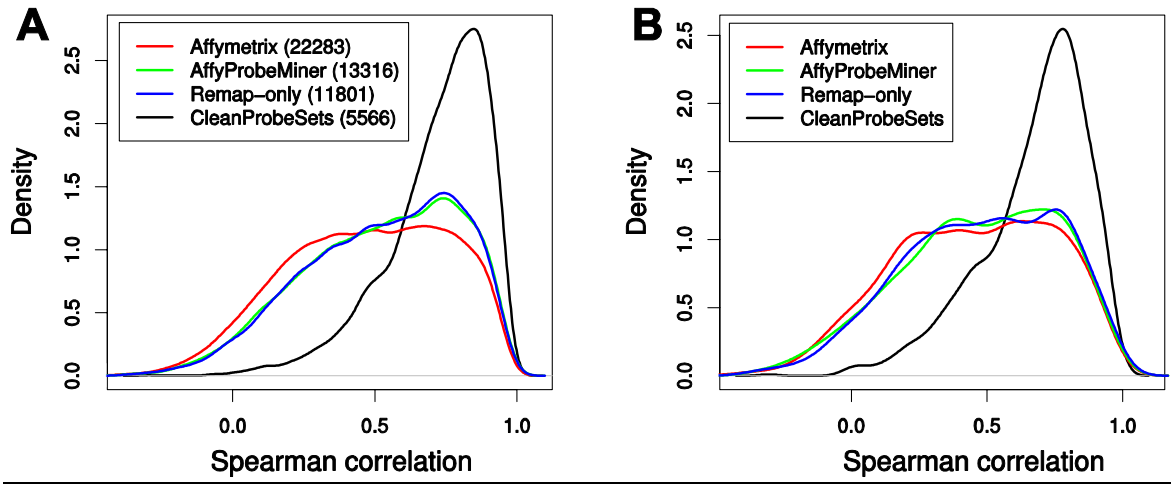
Frequency of probe sets with variable proportion of inconsistent probes as identified by CleanProbeSets for 50 B-cell samples hybridized to **(A)** U95Av2 and **(B)** U133plus2 gene chips. The majority of inconsistent (discarded) probes were mapped to genes with no consistent probes (blue), but 20% and 10% of inconsistent probes in U95Av2 and U133plus2 (red) were mapped to probe sets with over 90% probe consistency rates.

Figure 3: CleanProbeSets estimated false discovery rates



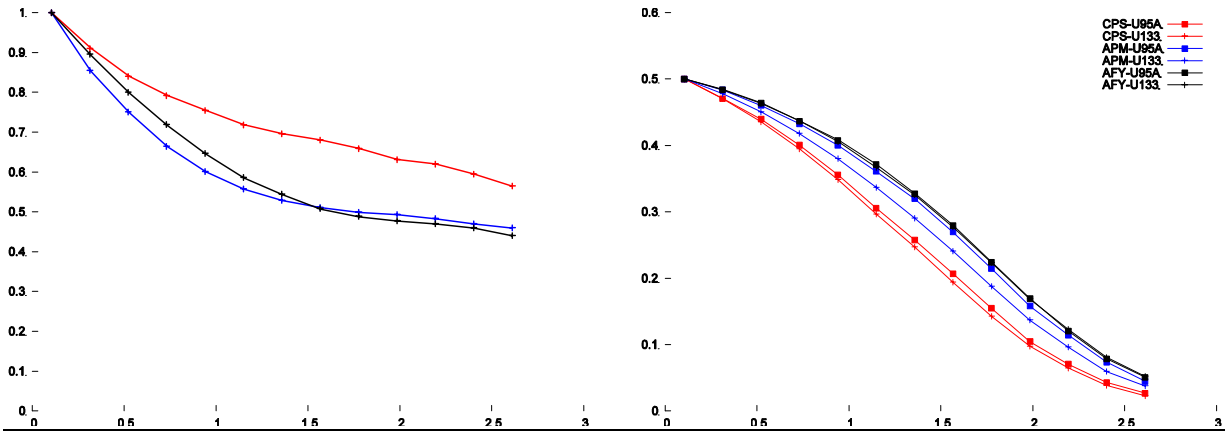
(A) Number of detected probe sets for variable-size randomly-selected subsets of the 152 and 200 microarray expression experiments on the U95av2 and U133plus2 platforms. **(B)** False discovery rate associated with subsets from (A) were estimated by permutation testing.

Figure 4: Consistency across technical replicate experiments



Density distributions for the spearman correlation coefficient across technical replicates considering **(A)** all the probe sets generated by each method (size in parenthesis)**,** and **(B)** only the 1,000 probe sets with the highest sample variation for each method. CleanProbeSets probe sets show dramatically better agreement across technical replicates even after pruning to remove low variability probe sets. Remap-only probe sets were statistically indistinguishable from AffyProbeMiner probe sets; both sets showed significantly better agreement across technical replicates than Affymetrix probe sets in panel A, but the three were statistically indistinguishable after pruning out low sample variation probe sets (panel B).

Figure 5: cross-platform consistency for differential expression analysis



Comparison of cross-platform consistency (U95A vs. U133P2), and estimated individual accuracy of differential expression calls using CleanProbeSets [red], AffyProbeMiner [blue] and Affymetrix [black] annotation. Comparisons are made as a function of the z-value threshold used for identifying differentially down-regulated genes (x-axis). Cross-platform consistency **(left)** was measured as the proportion of genes that are called down regulated by both platforms to genes that are probed by both platforms and are called down regulated by at least one of the platforms. Accuracy of individual experiments **(right)** was measured using FDR estimates from permutation testing, where all probe sets scoring above threshold in the original data are called true positives and all probe sets identified in permutated data are called false positives.