# PANP - a New Method of Gene Detection on Oligonucleotide Expression Arrays

Peter Warren[1,2], Jadwiga Bienkowska[1, 3], Paolo G. V. Martini[1], Jennifer Jackson[1], and Deanne M. Taylor [1,2]

1 Systems Biology Group, Serono Research Institute, Rockland MA, USA. 2 Bioinformatics Program, Rabb Graduate School, Brandeis University, Waltham MA, USA. 3 Department of Biomedical Engineering, Boston University, Boston MA.

peter.warren@verizon.net, deanne.taylor@serono.com

**Abstract.** Currently, the method most used for gene detection calls on Affymetrix oligonucleotide arrays is provided as part of the MAS5.0 software. The MAS method uses Wilcoxon statistics for determining presence-absence (MAS-P/A) calls. It is known that MAS-P/A is limited by its need to use both perfect match (PM) and mismatch (MM) probe data in order to call a gene present or absent. A considerable amount of recent research has convincingly shown that using MM data in gene expression analysis is problematic. The RMA method, which uses PM data only, is one method that has been developed in response to this. However, there is no publicly available method outside of MAS-P/A to establish presence or absence of genes from microarray data. It seems desirable to decouple the method used to generate gene expression values from the method used to make gene detection calls. We have therefore developed a statistical method in R, called Presence-Absence calls with Negative Probesets (PANP) which uses sets of Affymetrix-reported probes with no known hybridization partners on three chip sets: HG-U133A, HG-U133B, and HG-U133 Plus 2. This method uses a single, empirically-derived means to generate p-values used for cutoffs, which can reduce errors that can be introduced by using fitted models. In fact, PANP allows a user to utilize any Affymetrix microarray data pre-processing method to generate expression values, including PM-only methods as well as PM-MM methods. Cutoffs are generated in terms of the data on each chip, so even pre-processing methods that do not normalize across chipsets can be used. Additionally, a user can specify an acceptable p-value cutoff threshold. We present our results on PANP and its performance against the set of 28 HG-U133A chips from a published Affymetrix Latin squares spike-in dataset as well as an in-house TaqMan-validated human tissue dataset on the HG-U133 Plus 2 chipsets. We find that using these datasets, PANP out-performs the MAS-PA method in several metrics of accuracy and precision, using a variety of pre-processing methods: MAS5.0, RMA, and GCRMA. PANP out-performs MAS-P/A in gene detection across a full range of concentrations, especially of low concentration transcripts. An R software package has been prepared for PANP and is available as a BioConductor package at the Bioconductor Web site, http://www.bioconductor.org

## Introduction

High-throughput gene expression analysis using Affymetrix oligonucleotide chips is a common method used for transcriptional profiling. These studies often focus on the contrast between expression profiles from two or more samples. However, generating full expression profiles are important for several purposes, including clustering of gene expression data or populating system or network maps. To generate an expression profile, a method of "calling" a gene's expression must be used to distinguish a probeset's target sequence as most likely present or absent (PA) in the sample independent of any generalized non-specific hybridization or background effects remaining after pre-processing.

Currently, the MAS5.0 presence-absence (MAS-P/A) method is the commonly used post-processing method to "call" the presence or absence of a detected gene signal on an Affymetrix chip (Affymetrix 2001). Since the MAS-P/A method requires both PM and MM probes to make the presence-absence (PA) call, PM-only normalization methods cannot be used with MAS5.0. A chip intensity profile can be generated from PM-only methods, but without a PA call based on probability of detection.

Many Affymetrix probesets are designed based on EST matches in the public databases. Normally, these can provide good target matches to predicted protein-coding genes. However, occasionally ESTs are poorly annotated as to their strand direction. As a result, some probesets have been designed in the reverse complement – in the "sense" direction against their own transcripts. That is, these probesets cannot hybridize to the true (intended) EST target, but would hybridize instead to the reverse complement if it was transcribed. We decided to call these Negative Strand Matching Probesets (NSMPs).

We conceived that a useful per-chip gene detection method could be based on an intensity distribution of these NSMPs. The number of probesets would need to be large enough to provide a robust probability distribution defining "absence". Given such a set of NSMPs, some measure of distance from their intensity distribution would need to be devised to indicate the p-value of "presence" or "absence" of a gene detection on the chip. On the Affymetrix chips, there are some individual negative control probesets, but these are far too few to provide the robust distribution needed for a gene detection method. However, much larger sets of negative controls have fortuitously, and inadvertently, been included on three human genome chipsets: the Affymetrix HG-U133A, HG-U133B, and HG-U133 Plus 2. At this time, PANP can only be used with these three chipsets. Unfortunately, the HGU95 chips have too few NSMPs to build a statistically representative set of negative controls.

Affymetrix has recently (October, 2004) released versions of their chip annotation files that include the annotation of these probesets. Based on our analysis, we believe the annotated NSMPs provide a fortuitous, ready-made set of negative controls for use in quantifying a distribution of non-specific hybridization signal.

Our final sets contained 300 NSMPs for the HG-U133A, 363 for the HG-U133B, and 1006 for the HG-U133 Plus 2. We concluded that the sets are of sufficient size to provide statistically significant sets of negative controls for each of the three chips. We named the new method "PANP", for "Presence/Absence calls from Negative Probes".

A PM-only gene detection method, called the Half-Price method, has recently been developed, but was not publicly available at the time of writing of this paper (Wu and Irizarry 2005). Software for this method was kindly provided to us by the authors, and we analyzed it along with our PANP method and MAS-P/A. The results of comparison between PANP and the Half-Price method can be found under Supplemental Materials at our website (http://www.brandeis.edu/~dtaylor/PANP/).

## Data

The data used to test the PANP method is a set of 28 HG-U133A chips from the Affymetrix Latin squares spike-in dataset (Cope, Irizarry et al. 2004). In this dataset, 42 genes are spiked in at 14 concentration levels: 0.0, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256 and 512 pM. The Latin squares arrangement ensures that each of the 14 concentration levels has 84 expression data points (28 chips times 3 replicates per gene per concentration). The genes are spiked into a known sample drawn from the HeLa cell line, and the spike-in genes are known to be normally absent from this sample. To increase the representation of the 0 concentration level, we have added the NSMPs to the small set of negative controls in the spike-in data set. Since all spike-ins are of a known concentration, it is possible to evaluate the effectiveness of a given method at detecting the spike-in genes. This gives a picture of a method's accuracy. Also, since all non-spiked genes in the background sample are applied equally on all 28 chips, it is possible to determine the precision (or variability) of each method at calling these genes present or absent, even though their actual state is not known. These two measures together are necessary to provide a basis for comparing gene detection methods.

This set of 28 chips is the set used in the Bioconductor package affyCOMP version II, for comparative analysis of gene expression data processing methods, and has considerable credibility in the research community. (Cope, Irizarry et al. 2004; Gentleman, Carey et al. 2004)

### TaqMan dataset

Three human tissues were prepared for expression studies according to an in-house RNA extraction protocol. The RNA was prepared and hybridized according to recommended Affymetrix protocols. The tissue RNA was hybridized on HG-U133_Plus2.0 Affymetrix microarrays.

Fluidic card TaqMan analysis (ABI 7900HT) and manual TaqMan analysis were performed on selected genes from these three tissues, using commercially available probes and primers sets from Applied Biosystems according to the manufacturer’s instructions.

## Methods

In order to build a gene detection method using the NSMPs as a baseline set of negative controls, we selected probesets from the Affymetrix probeset annotation for the three chip versions, searching for the phrase "negative strand matching probes" associated with the probeset's own gene identifier, and eliminating any probesets with "cross hyb" annotation, which indicates a match to another, off-target gene. We ran BLAT on the NCBI dbEST database to check the resulting negative strand probesets. For example, on the HG-U133A, we found eight NSMPs whose target sequences had correct matches to greater than 5 independently reported EST transcripts matching the probeset consensus sequence and had signal levels associated across all probes that were among the highest in the NSMP set with signals greater than 2SD above the mean probeset signal across the NSMPs in the Latin Square dataset. These may reflect wrongly annotated NSMPs, cross-hybridization transcripts, or transcripts with unexpected antisense expression. We removed these probesets from our HG-U133A NSMP set as we assumed these were probesets that may be strongly hybridizing to intended target sequences. We were satisfied that the vast majority of our remaining NSMPs reflected true negative strand matching probes with no cross-hybridization partners.

### Benchmarking performance of gene detection methods

We concentrate on two key dimensions of performance: accuracy and precision. We first focus on a useful measure of accuracy, the receiver operating characteristic (ROC) curve. This plots the true positive (TP) rate against the false positive (FP) rate, showing graphically the cost (increased FP rate) of increasing the TP rate. This cost can be used as a measure of performance: the lower the cost, the better the method is at making TP calls for a given FP rate. To extract this information from the ROC curves, we then make use of a standard way to summarize accuracy and make it non-parametric: we take the area under each ROC curve and plot it for each method, per spike-in concentration level. This makes it straightforward to compare how the methods perform across the full range of concentration levels.

We also use a measure we call total accuracy, defined as the mean of the combined true positive (TP) and true negative (TN) rates: mean (TP+TN) rates. We can then calculate this total accuracy per concentration level to further quantify the comparative performance of the methods, showing how well each does in making accurate present *and* absent calls for a given p-value cutoff.

To evaluate precision, we look at the variability in majority calls for all the 22,258 non-spikein genes in the 28 HG-U133Achips in the Latin squares set. The majority call is defined as the most prevalent call made for each gene, whether it is present (P), absent (A), or marginal (M). We determine the fraction of the most prevalent call for a given gene. A perfectly consistent call is 28/28, or 1. Once this fraction has been calculated for each non-spikein gene across the 28 arrays, then the mean and standard deviation of those values is calculated. This becomes a measure of consistency, or precision, in making calls. The closer the mean is to 1, and the tighter the distribution around the mean, the more consistent the method is in making calls.

**Differential expression p-value from NSMP distribution**

We used the empirical distributions of the NSMPs to devise a p-value of "distance" from the negative probeset distribution. The measure represents the likelihood that, for a present gene, its expression value was sufficiently higher than the bulk of the negative controls. This has several advantages, including simplicity and accuracy.

The data from a chip or set of chips are first pre-processed using any desired method, such as RMA (Irizarry, Hobbs et al. 2003), GCRMA (Wu, Irizarry et al. 2004), and MAS5.0 (Affymetrix 2001). Each method is available as part of the *affy* package in Bioconductor (Gautier, Cope et al. 2004).

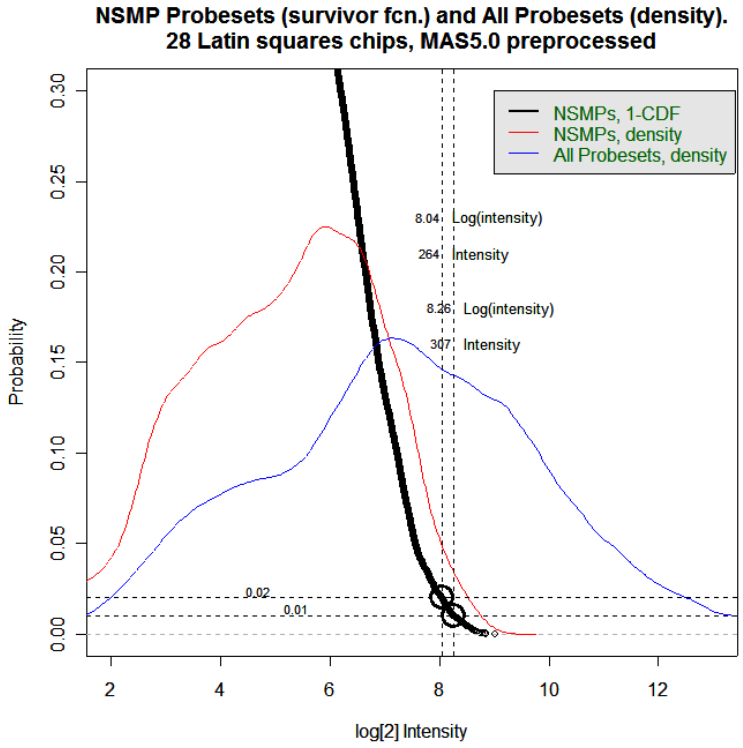


**Fig. 1.** Probability density vs. survivor function plots for the combined 28 Latin squares chips. Blue is density for all probesets. Red is density for the NSMPs (Negative Strand Matching Probesets). Black is empirical survivor function (1-CDF) points for NSMPs. Intercepts are shown for two example p-value cutoffs: 0.01 and 0.02, and for the interpolated intensities at those points.

For each chip in an analysis, the probability distribution of the NSMPs is calculated and then used to generate the cumulative distribution function (CDF). To derive a

cutoff intensity at any given p-value cutoff, we use the survivor distribution (1-CDF), as illustrated in **Fig. 1**. A selected p-value cutoff (Y axis) is interpolated on the survivor curve into a corresponding intensity (X axis). This intensity provides the expression level cutoff used to make presence/absence calls: genes with intensities higher than the cutoff are more likely to be present; those lower than the cutoff are likely to be absent. As always with p-values, lower numbers indicate increased significance. Note that the PANP p-value cutoff is a direct, empirically derived number. For cutoff *n*, a percentage of (*n* x 100) of the negative controls will be called present. This translates directly to a false positive (FP) rate of *n* (as shown in **Table S1)**

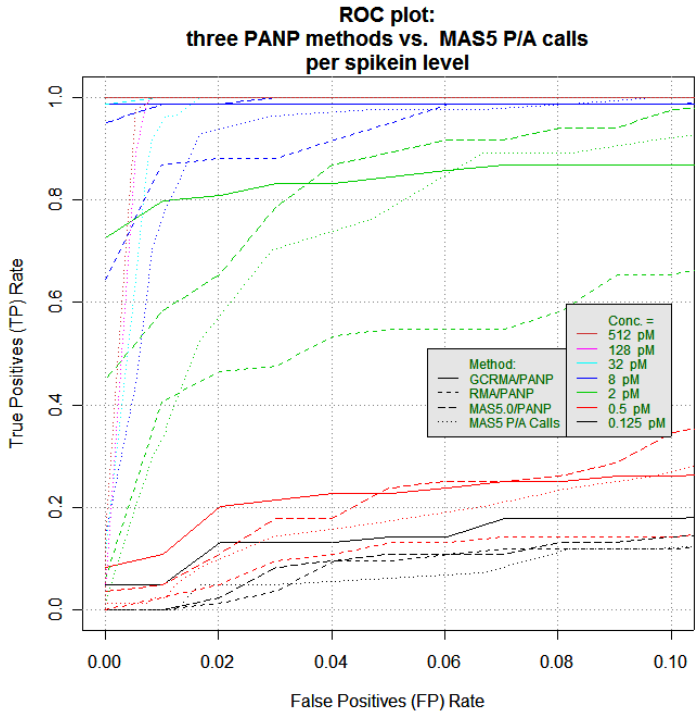


**Fig. 2**. ROC plots (TP vs. FP rates) comparing PANP with three different preprocessing methods to MAS5 P/A calls.

For example, assume a p-value of 0.01 is selected by a user as a significance threshold, and the expression level at that point is interpolated from the survivor function. By definition, then, 99% of the negative strand probesets are lower than that expression level, and would receive "absent" calls. The remaining 1% are erroneously called "present", so the FP rate for the negative strand probes in this case is 0.01.

The PANP method is based on the assumption that the great majority of all true negative probesets will fall within the distribution of this set of NSMPs. Therefore,

for the lower cutoff of p = 0.01, the overall FP rate should be close to 0.01. To determine an appropriate cutoff p-value using PANP, the user should decide on an acceptable FP rate, and choose appropriately. For example, **Fig. S1** shows the true positive CDF curves for all 28 Latin Squares chips resulting from the use of a 0.05 cutoff.

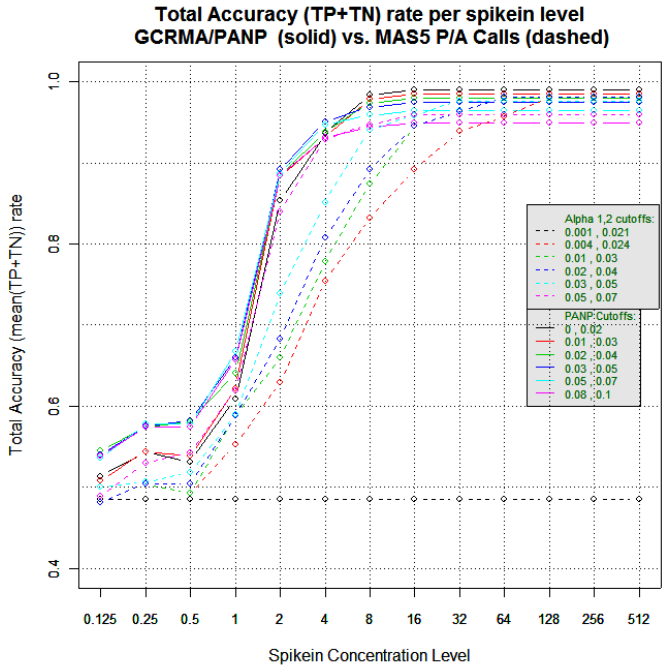


**Fig. 3.** Total Accuracy plots, defined as mean (TP+TN rates) vs. spikein level. Solid lines denote PANP, while MAS-P/A uses dashed lines. Results using several equivalent cutoffs are shown (equivalence via FP rates, as in **Table S1**), for FP rates <= 0.1. PANP was preprocessed with GCRMA.

In order to compare PANP to MAS-P/A two cutoff values are used, which we designate as "tight cutoff" (more stringent) and "loose cutoff" (less stringent). Like MAS-P/A, values below the tight cutoff indicate "presence"; values above the loose cutoff indicate "absence"; and values between the two cutoffs are considered "marginal". However, PANP's cutoff values cannot be directly compared to MAS-P/A's alpha 1 and alpha 2 cutoffs. (The latter have different meaning: they are cutoffs for Wilcoxon ranking p-values assigned to each probeset's intensity.) Therefore, we determined equivalence between the cutoffs by aligning them to false positive (FP) rates resulting from using each cutoff pair. This alignment is shown in **Table S1**.

## Results and Discussion

To establish whether PANP shows improvement in gene detection over the widely used MAS-P/A method, we evaluated and compared performance in the two key areas of accuracy and precision (as defined in the Methods section). Furthermore, to incorporate an assessment of the impact of choice of pre-processing method on PANP's performance, we compared the performance of four approaches: PANP/RMA (PANP preceded by RMP preprocessing); PANP/GCRMA; PANP/MAS5.0; and MAS-P/A itself.

Because of the inequality between the p-values of PANP and MAS-P/A, we chose equivalent p-values using the resulting FP rate, presented in **Table S1** as detailed in Methods. For all our plots of results, we focused on a reasonable range of FP rates from 0 to 0.1 (that is, up to 10%), above which results become increasingly useless for practical gene detection.

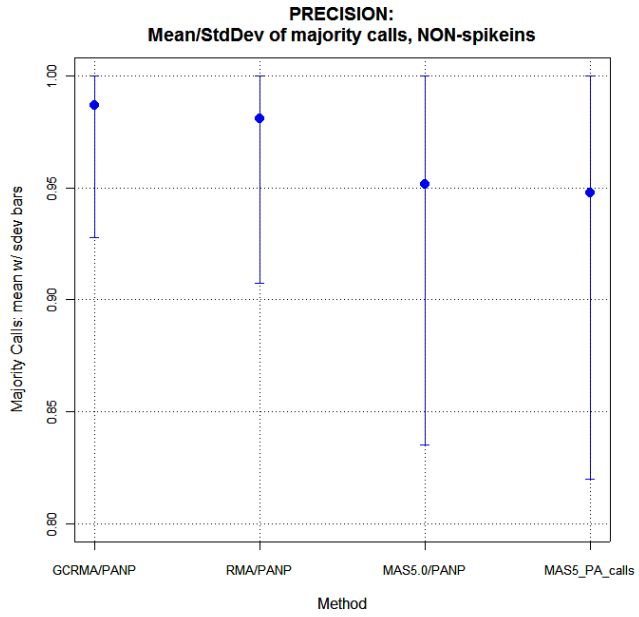


**Fig. 4**. Plots of mean and standard deviation of majority calls of all non-spike-in genes on the 28 HG-U133A Latin squares chips. The four methods are shown from left to right. The closer the mean and standard deviation are to 1, the higher the precision.

Receiver operating characteristic (ROC) plots are useful in providing comparisons of accuracy. **Fig. 2** compares ROC curves for the four approaches. These curves plot true positives (TP) rates against false positives (FP) rates for different cutoff pairs, thus showing the tradeoff between sensitivity (TP) and error rates (FP). It is evident that GCRMA/PANP and MAS5.0/PANP clearly outperform MAS-P/A for every con-

centration level. RMA/PANP, however, shows a clear advantage only for some concentrations.

Over the entire range of false positive (FP) rates (not shown), the overall best performer in the four approaches is MAS5.0/PANP, followed by GCRMA/PANP. However, for reasonable FP rates of < 0.1 (**Fig. 2**), GCRMA/PANP and MAS5.0/PANP perform nearly identically, with MAS5.0/PANP holding a slight edge.

The accuracy picture is not complete without considering true negative (TN) as well as true positive rates. It can be just as important to have high accuracy for absence calls as for presence calls. Therefore, we evaluated total accuracy, which combines both TP and TN rates as detailed in Methods. Total accuracy results are shown in **Figs 3** and **S2**: mean (TP+TN rates) vs. spike-in concentration levels for six equivalent cutoff pairs spanning the FP range of interest, 0-0.1. **Fig. 3** is for GCRMA/PANP vs. MAS-P/A, while **Fig. S2** is for MAS5.0/PANP vs. MAS-P/A.

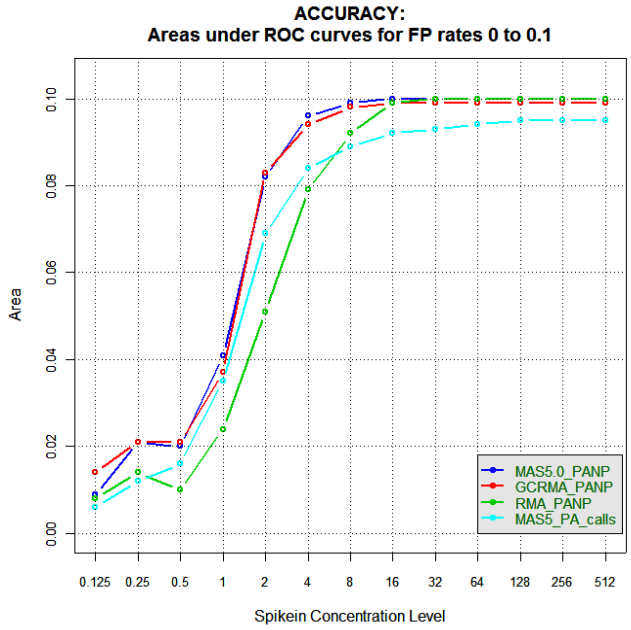


**Fig. 5.** Summary accuracy plots: areas calculated under the ROC curves that appear in **Fig. 2**, per spike-in concentration level. The four methods are color-coded.

As cutoffs are increased towards higher (less stringent) p-values, TPs increase, and TNs decrease. **Figs. 3** and **S2** shows this tradeoff: for MAS-P/A calls, the ratio of TNs lost to TPs gained as cutoffs increase is clearly larger than for PANP with either GCRMA or MAS5.0 preprocessing. This is evident for all the curves shown. Note also that for the first alpha pair, the MAS-P/A total accuracy rate remains minimal across all concentration levels, and it is not until the parameter alpha1 reaches .004 (red) that the first real gains in total accuracy are seen.

A total accuracy plot of RMA/PANP vs. MAS-P/A (not shown) is similar to the two shown, with RMA/PANP performing better that MAS-P/A except for FP rates near the 0.1 limit of the range evaluated. It is especially interesting to note that MAS5.0/PANP significantly out-performs MAS5.0/PA at all concentrations (See **Fig. S2**) which allows for a direct comparison of PANP to MAS-P/A using the same preprocessing method.

By the metric of total accuracy, where true negatives are included as part of the picture, PANP is clearly shown to out-perform MAS-P/A calls for reasonable FP rates, regardless of which preprocessing method is used.

**Fig. 4** shows the variability (precision) of the four approaches over all non-spike-in genes in all 28 arrays, where these arrays have all been processed together. **Fig. 4** shows the mean and standard deviation of the majority calls (P,M or A) fractions, as described in the Methods section, using equivalent cutoffs from **Table S1**. The higher the mean and the smaller the standard deviation, the better the precision. Both MAS-P/A calls and MAS5.0/PANP show significantly more variability (i.e., poorer precision) in P/M/A calls on the non-spikein probesets than the RMA and GCRMA methods. This is to be expected, due to the cross-chip normalization used in the RMA methods. However, MAS5.0/PANP seems to perform equally well if not slightly better than MAS-P/A. The increased variability in p-values in MAS5.0 implies less precision in presence-absence calls, although it is evident that precision is strongly dependent on the pre-processing method used.

**Fig. 5** shows a summary of accuracy by plotting the area under ROC curves per spikein level for the useful range of FP rates up to 0.1. This shows that PANP, with either GCRMA or MAS5.0 processing, outperforms MAS-P/A calls for every spikein level. RMA/PANP outperforms MAS-P/A for concentration levels below 0.5pM and above 4pM. However, RMA/PANP generally lags behind the other two preprocessing methods. This result correlates with issues regarding RMA's overall accuracy which have been noted elsewhere, and are assumed to be due to the compressing effect of RMA's multichip normalization and background correction. RMA is known to trade off some accuracy for greatly improved precision. Clearly, GCRMA has improved accuracy, apparently due to its use of probe-specific binding affinities derived from GC-content information.

The areas under MAS-P/A's ROC curves in **Fig. 5** never reach the 0.1 asymptote because even at the highest concentration levels MAS-P/A always results in some FP rate for any non-zero TP rate, even for the smallest alpha cutoffs, as seen in **Fig. 2**. Conversely, PANP achieves 100% TP rate with 0 FP rate for all concentrations above 8 pM.

It may be interesting to note that in the vast majority of NSMPs, there is a relatively low variance and low probeset intensity between Latin Square data samples in the NSMP set, perhaps reflecting the fact that non-specific hybridization does not contribute to the NSMP probeset signals in a strongly differential fashion. The small amount of variance in the signal may reflect an accurate picture of non-specific hybridization on whole Affymetrix probesets.

Because of the profile of NSMPs on the HG-U133B, we find there may be a greater number of NSMPs available on that chip than are currently annotated, and therefore this chip may need further annotation work before being used in PANP.

**Performance against human tissue oligonucleotide chips vs. TaqMan verification**

We used PANP on three Affymetrix HG-133 Plus 2 chips, hybridized to human tissue mRNA. Using a strict P-value cutoff of 0.01 or 0.02, we find that all three methods with PANP perform equally well in detecting positive expression using TaqMan validation, especially when the 0.02 cutoff is used. This is particularly noticeable in detection of low-expressing genes, such as GPCRs (see **Table 1**). Because we assume that TaqMan validation of certain genes is a detection of true negatives or true positives, we only record the true positive rates here. Using equivalent MAS-P/A alpha cutoffs of 0.002 and 0.003, **Table 1** also shows that PANP with all three pre-processing methods significantly out-performs MAS-P/A.

| Total TP, P only, detected (cut-off/alpha) | RMA PANP | GCRMA PANP | MAS5 PANP | MAS 5PA |
|---|---|---|---|---|
| Fluidic card TaqMan data (0.01/0.002) | 24 | 22 | 27 | 16 |
| Manual GPCR Taqman data (0.1/0.002) | 44 | 49 | 60 | 35 |
| Totals for TP detected (0.01/0.002) | 68 | 71 | 87 | 51 |
| TP for expanded cutoffs (0.02/0.003) | 102 | 105 | 101 | 83 |

**Table 1.** True positives, returned from TaqMan results for two equivalent cutoff/alpha values. True positives are detected for fluidic card data, followed by GPCR data, followed by totals for both datasets. Detailed results for the expanded 0.02 cutoff are not shown, just the total result.

## Conclusions

PANP has been demonstrated to be a simple, effective, and flexible new gene detection/calling method that outperforms the current standard, MAS-P/A calls, by several key metrics of accuracy and precision. This has been demonstrated on both spike-in data sets and TaqMan validation of non-spike-in data sets of low concentration. PANP allows one to use any Affymetrix pre-processing method to generate expression values; it can be used with PM-only pre-processing methods, as well as methods that use both PMs and MMs. For making presence-absence calls, PANP's cutoffs are tailored to each chip: PANP uses each chip's unique negative probeset expression distribution to generate a p-value cutoff. This can reduce errors that can be introduced by using fitted models. While this is not as significant for the RMA or GCRMA methods that normalize across chips, it can be useful in methods that normalize gene expression levels on a per-chip basis.

MAS5.0/PANP and GCRMA/PANP are nearly tied in terms of accuracy, but GCRMA/PANP is clearly superior in precision. Although RMA/PANP did not do as well as the others in accuracy, it still out-performed MAS5.0 and MAS-P/A in terms of precision. PANP delivers the best accuracy and precision overall when compared to MAS-P/A.

## References


Affymetrix (2001). Statistical algorithms reference guide. Technical report, Affymetrix.

Cope, L. M., R. A. Irizarry, et al. (2004). "A benchmark for Affymetrix GeneChip expression measures." Bioinformatics **20**(3): 323-331.

Gautier, L., L. Cope, et al. (2004). "affy--analysis of Affymetrix GeneChip data at the probe level." Bioinformatics **20**(3): 307-315.

Gentleman, R., V. Carey, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biology **5**(10): R80.

http://www.brandeis.edu/~dtaylor/PANP/ PANP website at Brandeis University.

Irizarry, R. A., B. Hobbs, et al. (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostat **4**(2): 249-264.

Wu, Z., R. Irizarry, et al. (2004). "A Model Based Background Adjustment for Oligonucleotide Expression Arrays." Johns Hopkins University Dept. of Biostatistics Working Paper Series(1001).

Wu, Z. and R. A. Irizarry (2005). "A Statistical Framework for the Analysis of Microarray Probe-Level Data." Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 73 http://www.bepress.com/jhubiostat/paper73.


Supplemental figures, tables, and information is available from the PANP website at Brandeis University: http://www.brandeis.edu/~dtaylor/PANP/ and the Bioconductor website at http://www.bioconductor.org/