

PANP – a New Method of Gene Detection on Oligonucleotide Expression Arrays

Peter Warren, Deanne Taylor, Paolo G. V. Martini, Jennifer Jackson, Jadwiga Bienkowska

Systems Biology Group

EMD Serono

Rockland, Massachusetts

peter.warren@verizon.net, dtaylor@hsph.harvard.edu

Abstract—The method currently most used for probeset detection calls on Affymetrix GeneChip® Human Genome Arrays is provided as part of the MAS5 software. The MAS method uses Wilcoxon statistics for determining presence-absence (MAS-P/A) calls. However, MAS-P/A is only usable with MAS5 processing, which requires the use of both perfect match (PM) and mismatch (MM) probe data in order to call the resulting probeset present or absent. A considerable amount of recent research has convincingly shown that using MM data in gene expression analysis may be 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 that works with PM-only expression data to establish presence or absence of genes from the probesets in 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 two chip sets: HG-U133A and HG-U133 Plus 2.0. PANP allows the use of any Affymetrix microarray data pre-processing method to generate expression values, including PM-only methods as well as PM and MM methods. We present our results on PANP and its performance using the set of 28 HG-U133A chips from a published Affymetrix Latin squares spike-in dataset as well as an internal TaqMan-validated human tissue dataset on the HG-U133 Plus 2.0 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: RMA, GCRMA, and even MAS5 itself. PANP out-performs MAS-P/A in probeset detection across a full range of concentrations, especially with low concentration transcripts. An R software package has been prepared for PANP and is available in R as part of the Bioconductor package release at <http://www.bioconductor.org>.

Keywords - *microarray, Affymetrix, presence-absence, transcription, probeset detection, gene detection*

I. INTRODUCTION

High-throughput gene expression analysis using Affymetrix GeneChip® oligonucleotide chips is a common method used for studies of transcriptional profiling. These studies often focus on the contrast between expression profiles from two or more samples and the results then used to quantify changes in transcription. For a recent review, see [1].

With chip pre-processing with methods such as RMA [2-4],

GCRMA [5], or MAS5 [6], a measure of signal intensity is recorded at each probe, background is commonly subtracted, and other method-specific normalization procedures are performed. The probe intensities are then summarized into an expression measure of the probeset represented. Probeset intensities across samples are then compared for differential expression using various post-processing methods, such as modeling and clustering [1].

Until recently, there has been little attention paid to the actual process of probeset detection itself. Probeset detection on an Affymetrix GeneChip® Human Genome Array is the process of determining the likely presence or absence of a transcript in a sample. An often-used measure of sample and chip quality is the detection of the numbers of probesets called present, and some post-processing methods use a probeset's present or absent call as a first filtering step in analysis of gene expression.

There is often a lack of information on transcriptional context behind standard differential expression studies. It has long been known that genes can be functionally pleiotropic as a result of multiple roles in different contexts [7]. Comparison of changing gene expression profiles will detect those probesets with discernable change between the two samples, but alone cannot speak to the underlying transcriptional background against which those probesets are detected as changing. Additionally, it will be useful to generate transcriptional profiles across tissues that differ widely in expression. A full sample expression profile will also be useful for systems biology studies of transcription.

To generate a full expression profile on an Affymetrix chip after preprocessing, a chip's probesets must be called as most likely present or absent (P/A) in the sample, optimally as independent of any generalized non-specific hybridization or other background effects remaining after preprocessing. The MAS5 presence-absence (MAS-P/A) method is the most commonly used post-processing method to "call" the presence or absence of a detected probeset signal on an Affymetrix chip [6]. Since the MAS-P/A method works only within MAS5 and requires both PM and MM probes to make the presence-absence (PA) call, PM-only normalization methods such as RMA cannot be used with the MAS-P/A method. A chip intensity profile can be generated from PM-only methods, but

without a P/A call based on internal chip controls. Therefore, PM-only methods such as RMA are typically used for fold-change comparisons.

We were interested in cataloging transcriptional profiles on samples across experiments, platforms, tissues and cell lines. To be accurate, transcriptional profiles must be generated per chip on a set of on-chip controls against non-specific hybridization. Typically it has been the MM probes on Affymetrix chips that have served as those controls [6] but the use of MM probes faces several complications, among them differential intensity based on the internal mismatch identity [8], effects from biotin labeling [9] and the possibility that a mismatch probe may detect some target signal itself. Therefore, a presence-absence method that does not explicitly use MM probes would be desirable.

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, some ESTs are not accurately annotated as to their strand direction. As a result, some Affymetrix 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), following Affymetrix’s annotation convention.

We conceived that a useful per-chip probeset detection method could be based on an intensity distribution of these NSMPs. This distribution can act as a per-sample control for non-specific hybridization. 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 probeset detection on the chip. Such negative controls have been inadvertently included on three human genome chipsets: the Affymetrix HG-U133A, HG-U133B, and HG-U133 Plus 2.0. At this time, PANP has been validated for use with the HG-U133A and HG-U133 Plus 2.0. Unfortunately, the HGU95 chips have too few NSMPs to build a statistically representative set of negative controls. We have limited ourselves currently to study of NSMPs in the human arrays and have not surveyed all of the Affymetrix Genome Arrays for NSMPs.

Affymetrix has provided on their website versions of their chip annotation files that include the annotation of these probesets (<http://www.affymetrix.com>). 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 and 1006 for the HG-U133 Plus 2.0. We concluded that the sets are of sufficient size to provide statistically significant sets of negative controls for non-specific hybridization for the two

chip types. We named the new method “PANP”, for “Presence/Absence calls from Negative Probes”. We generated an R package called “panp” which can use standard Bioconductor data objects to provide presence-absence calls. The panp package has been part of the Bioconductor package set since release 1.8 [10].

In this paper we present the PANP method and compare its performance across several metrics of accuracy and precision against the standard P/A method, MAS-P/A. PANP is evaluated using three preprocessing methods: RMA, GCRMA, and MAS5. Each of these three combinations is then compared with MAS5/MAS-PA. The dCHIP package suite [11] also includes a P/A method that is similar to MAS-PA, so we do not compare it here. While we are aware that there are several other popular methods for Affymetrix array preprocessing, we chose the three methods (GCRMA, RMA, MAS5) to reflect intrinsic differences in normalization methods rather than give an exhaustive analysis of pre-processing methods which can be found elsewhere [12, 13].

A PM-only probeset detection method, called the Half-Price method has recently been developed [14]. 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 comparison between PANP and the Half-Price method can be found under Supplemental Materials at our website [15].

The recently released Affymetrix GeneChip® Exon Arrays have incorporated genomic and antigenomic probes which can be used to estimate hybridization background. Recently there has been a method called GeneBASE [16] developed to exploit these particular probes to determine presence-absence calls on the exon arrays.

II. DATA

A. Spike-in Data

The data used to test the PANP method is a set of 28 HG-U133A chips from the Affymetrix Latin squares spike-in data set [17]. 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 probeset detection methods.

This set of 28 chips is available as part of the Bioconductor package affycomp for comparative analysis of gene expression data processing methods [18].

B. TaqMan Data

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 assayed for quality and hybridized on HG-U133 Plus 2.0 Affymetrix Human Genome Arrays. Fluidic card TaqMan analysis was performed on an ABI 7900HT Sequence Detection System (Applied Biosystems). Manual analysis was also performed on selected genes from these three tissues. Both high-throughput and manual analysis was performed using commercially available probes and primers sets from Applied Biosystems using protocols according to the manufacturer’s instructions.

As a measure of true presence of an mRNA transcript, we used the threshold cycle (ΔCT) values ($CT_{target} - CT_{\beta actin}$) averaged over the two within-card replicates to compare to the same tissue profiled on the Affymetrix chip. We found strong agreement between the within-sample replicates through the ΔCT range. We tabulated all ΔCT scores up to the detection limit of 40 and used a cutoff of 34 for our plots based on common laboratory practice.

We matched the TaqMan probes to the Affymetrix probes through GeneIDs provided by both manufacturers. In the case where there were multiple Affymetrix probes for the same TaqMan probes, we considered each TaqMan-Affymetrix pair individually in the analysis for PA calls.

III. METHODS

A. Generating a list of NSMPs

In order to build a probeset detection method using the NSMPs as a baseline set of negative controls, we selected probesets from the Affymetrix probeset annotation from October 2004 for the three Genome Array 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 gene, which may provide off-target detection signal. We ran BLAT on the NCBI dbEST database to check the resulting negative strand representative probeset consensus sequence. 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. These eight NSMPs also had signal levels associated across all probes in their probesets that were among the highest in the NSMP set with signals greater than 2 standard deviations above the mean probeset signal across the NSMPs in the Latin Square affycomp dataset. These may reflect wrongly annotated NSMPs, cross-hybridization transcripts, or transcripts with unexpected antisense expression. We removed these probesets from our HG-U133A

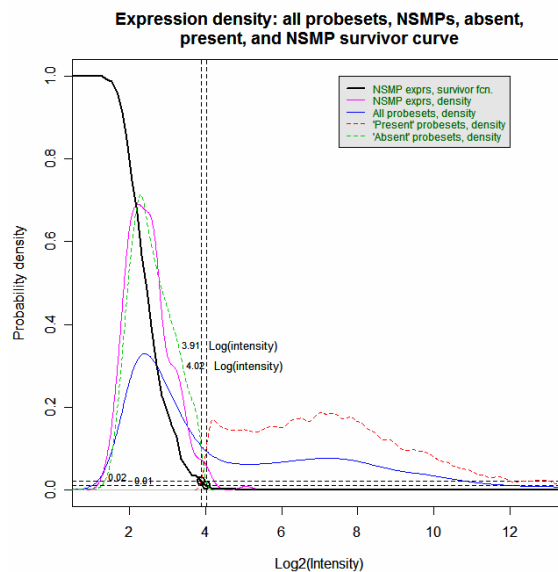


Fig. 1. Expression density plots showing separation of present and absent intensities. Probability density vs. survivor function plots for the combined 28 Latin squares chips, GCRMA preprocessed. Blue is density for all probesets. Magenta is density for the NSMPs (Negative Strand Matching Probesets). 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. Dashed curves indicate densities for probesets called "Present" (red) and "Absent" (green) by PANP as dictated by the cutoffs.

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. These NSMPs therefore became our baseline sets of negative controls.

B. Differential expression p-value from NSMP distribution

We used the empirical distributions of the NSMP signal intensities to devise a p-value of "distance" from the negative probeset distribution. This measure represents the likelihood that a particular probeset’s expression value is 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, GCRMA, or MAS5), for instance using the affy package [19] in Bioconductor [10].

Next, using the panp package, the probability distribution of the signal intensities of the NSMPs is calculated for each chip 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: probesets 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 ($n \times 100$) of the negative controls will be called present. This translates directly to a false positive (FP) rate of n .

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. Then by definition 99% of the negative strand probesets are lower than that expression level and would receive "absent" calls. The remaining 1% is erroneously called "present", therefore 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. With 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.

TABLE I
EQUIVALENCE OF PANP CUTOFFS WITH MAS P/A

False Pos. Rate	PANP cutoff	MAS P/A Alpha
0.00	0.00	0.00
0.01	0.01	0.004
0.02	0.02	0.012
0.03	0.03	0.02
0.04	0.04	0.026
0.05	0.05	0.032
0.06	0.06	0.037
0.07	0.07	0.042
0.08	0.08	0.049
0.09	0.09	0.056
0.10	0.10	0.061

Approximate equivalence is derived by aligning to resulting FP rates.

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, as 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 1.

C. Benchmarking performance of probeset 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 true positive (TP) rate (sensitivity) against the false positive (FP) rate (1-specificity), showing graphically the cost

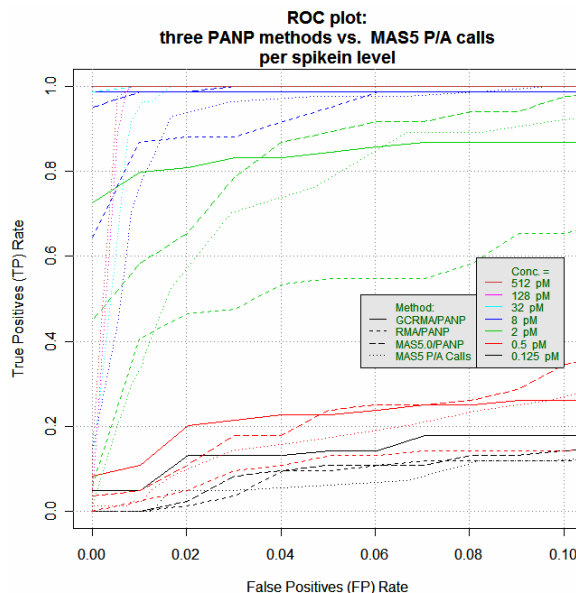


Fig. 2. ROC plots of rates (TP (sensitivity) vs. FP (1-specificity)) comparing three different preprocessing methods using PANP to MAS5 P/A calls across the different spike-in concentration levels.

of increasing sensitivity at the expense of decreasing specificity. 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 method to summarize accuracy and make it non-parametric by calculating the area under each ROC curve and plotting it for each method per spike-in concentration level. This makes for a straightforward comparison of 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-spike-in genes in the 28 HG-U133A chips in the Latin squares set from the affycomp Bioconductor package [17]. We define the majority call as the most prevalent call made for each probeset, whether it is present (P), absent (A), or marginal (M). We determine the percentage variation of the most prevalent call for a given probeset. A perfectly consistent call is 28/28, or 0% variation. Once this percent variation has been calculated for each non-spike-in probeset 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 0 and the tighter the distribution around the mean, the more consistent the method is in making calls.

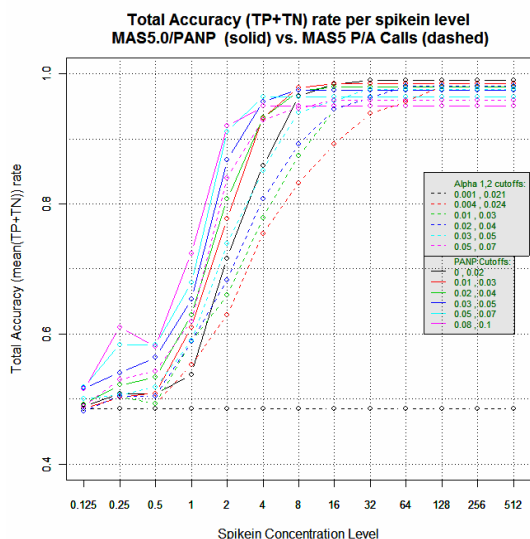
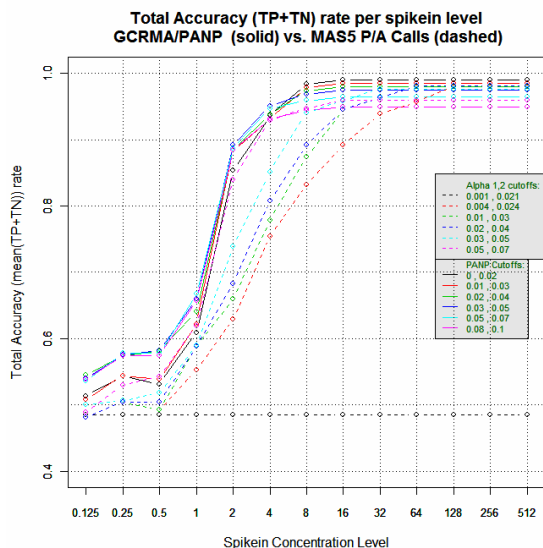


Fig. 3. Total Accuracy plots, defined as mean (TP+TN rates) vs. spike-in level. Solid lines denote PANP, while MAS-P/A uses dashed lines. Results using several equivalent cutoffs are shown (equivalence via FP rates, as in Table 1). Same colors are used for equivalent cutoffs. (a) GCRMA preprocessing was used for PANP vs. MAS5 P/A. (b) MAS5 used for PANP vs. MAS5 P/A.

D. Preprocessing parameters

We compared performance along these metrics of accuracy and precision between MAS-P/A on the one hand, and PANP with three preprocessing methods on the other. We used the implementations in the Bioconductor packages "affy" version 1.5.8 (for RMA, MAS5 and MAS-P/A) [19] and "gcrma" [5] version 1.1.3 for GCRMA. Default parameters were used in all cases for background correction, normalization and summarization.

A. Performance of PANP on the Affymetrix spike-in data

To establish whether PANP shows improvement in probeset 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 selected approaches: PANP/RMA (PANP preceded by RMA preprocessing); PANP/GCRMA; PANP/MAS5; 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 1 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 probeset detection.

Receiver operating characteristic (ROC) plots are useful in providing comparisons of sensitivity (TP) versus 1-specificity (FP). 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/PANP clearly outperform MAS-P/A for every concentration 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/PANP, followed by GCRMA/PANP. However, for reasonable FP rates of < 0.1 (Fig. 2), GCRMA/PANP and MAS5/PANP perform nearly identically, with MAS5/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 figures 3a and 3b. These figure show mean (TP+TN rates) vs. spike-in concentration levels for six equivalent cutoff pairs spanning the FP range of interest, 0 to 0.1. Fig. 3a is for GCRMA/PANP vs. MAS-P/A, while Fig. 3b is for MAS5/PANP vs. MAS-P/A.

As cutoffs are increased towards higher (less stringent) p-values, TPs increase, and TNs decrease. Figs. 3a and 3b 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 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 alpha reaches .004 (red) that the first real gains in total accuracy are seen.

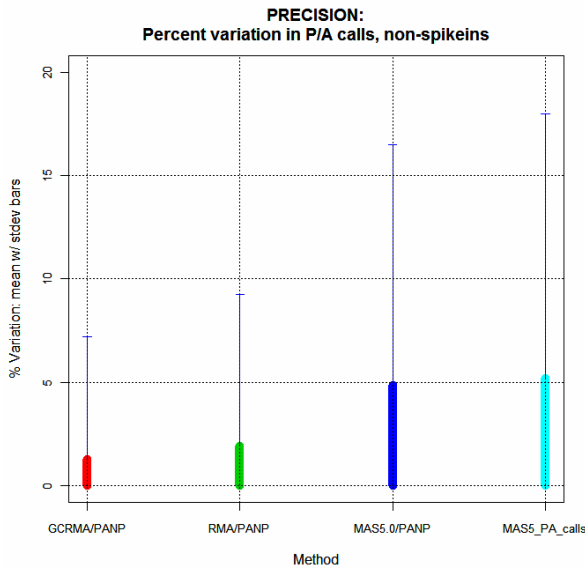


Fig. 4. Plots of means (bars) and standard deviations (error bars) of percent variation in per-gene majority calls across all non-spike-in genes on the 28 HG-U133A Latin squares chips. The four methods are shown from left to right. Highest precision is indicated by means closest to 0 % and smallest standard deviation error bars.

A total accuracy plot of RMA/PANP vs. MAS-P/A (not shown) is similar to the two shown, with RMA/PANP performing better than MAS-P/A except for FP rates near the 0.1 limit of the range evaluated. It is especially interesting to note that MAS5/PANP significantly outperforms MAS5/PA at all concentrations (See Fig. 3b). This shows that when PANP and MAS-P/A are both applied after the same MAS5 preprocessing method, PANP gives significantly better results. By the metric of total accuracy, where true negatives are included as part of the picture, PANP is clearly shown to outperform MAS-P/A calls for reasonable FP rates, regardless of which preprocessing method is used.

Figure 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 percent variation in majority calls (P, M or A) per probeset across all non-spike-in genes, as described in the Methods section, using equivalent cutoffs from Table 1. The closer the mean is to 0 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-spike-in probesets than the RMA and GCRMA methods. This is to be expected, due to the cross-chip normalization used in the RMA methods. It is evident that precision is more dependent on the pre-processing method used than on the probeset detection method. Given that fact however, PANP with MAS5 preprocessing shows slightly better precision than MAS-P/A calls. PANP does not in itself confer significantly better precision; rather, it allows probeset detection to be performed with more-precise preprocessing methods than MAS5.0.

Fig. 5 shows a summary of accuracy by plotting the area under ROC curves per spike-in level for the useful range of FP rates up to 0.1. This shows that PANP, with either GCRMA or MAS5 preprocessing, outperforms MAS-P/A calls for every spike-in 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 most likely reflects a bias induced by RMA which has been noted elsewhere and is assumed to be due to the compressing effect of RMA's multichip normalization and background correction [20]. RMA is known to trade off some accuracy for greatly improved precision, and as part of the GCRMA method, that bias has been corrected, which may be the reason for

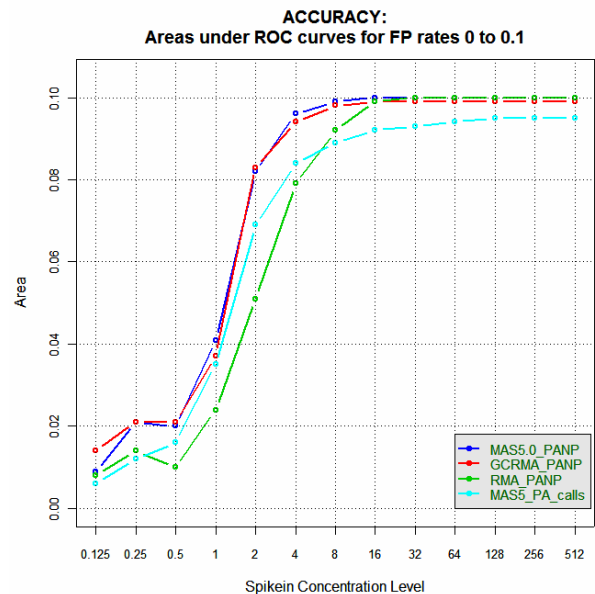


Fig. 5. Summary accuracy plots: areas calculated under the ROC curves that appear in Fig. 2, per spike-in concentration level.

GCRMA's obvious increase in accuracy over RMA with PANP.

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.

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

We used PANP on three Affymetrix HG-133 Plus 2.0 chips,

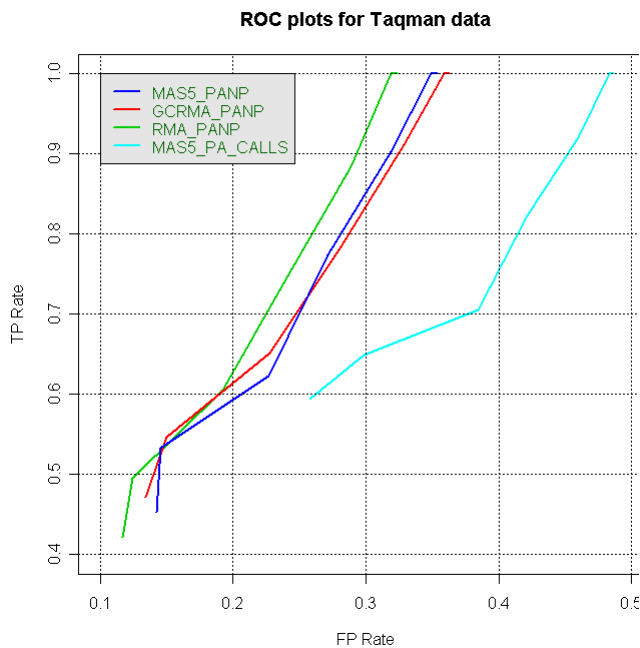


Fig. 6. ROC plot showing TP rate (sensitivity) vs. FP rate (1-specificity) for MAS5 P/A calls compared to PANP with four different preprocessing methods, using equivalent cutoffs.

hybridized to human tissue mRNA. True positives were those detected by TaqMan for fluidic card data for ΔCT count thresholds ranging from 22 to 34. False positives were those called present by PANP or MAS5 P/A calls, but not detected by TaqMan with $\Delta CT < 34$ as discussed in Methods. Fig. 6 shows how well PANP and MAS5 P/A calls performed in making presence calls that match the TaqMan results. Using equivalent P-value cutoffs between PANP and MAS5 P/A calls as described above, we find that all three methods with PANP perform equally well in detecting positive expression using TaqMan validation, and that all result in significantly improved sensitivity over MAS5 P/A calls for any given specificity. Because a measure of TaqMan ΔCT on our samples cannot be compared to a few deliberately spiked-in RNA species in another experiment, we find it unsurprising that there are some variations in the performance of the three pre-processing methods in the TaqMan versus the spike-in data. In the spike-in data, GCRMA performs better with PANP against RMA and MAS5. However, in the TaqMan data, RMA slightly outperforms both MAS5 and GCRMA. Most importantly, it is very clear that any method using PANP has a significantly higher sensitivity for any given specificity than MAS5 P/A.

V. CONCLUSION

PANP has been demonstrated to be a simple, effective, and flexible new probeset 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. 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. In the spike-in data, we have found that MAS5/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 and MAS-P/A in terms of precision. We conclude that PANP delivers the best accuracy and precision overall when compared to MAS-P/A.

VI. ACKNOWLEDGMENT

We would like to acknowledge helpful input from colleagues Gregg McAllister and Robert Campbell at EMD Serono, and Andrew Hill at Wyeth Research.

REFERENCES

- [1] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nat Rev Genet*, vol. 7, pp. 55-65, 2006.
- [2] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, pp. 185-193, January 22 2003.
- [3] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucl. Acids Res.*, vol. 31, p. e15, February 15, 2003.
- [4] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostat*, vol. 4, pp. 249-264, April 1 2003.
- [5] Z. Wu, R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer, "A Model-Based Background Adjustment for Oligonucleotide Expression Arrays," *Journal of the American Statistical Association*, vol. 99, pp. 909-917, 2004.
- [6] Affymetrix, Statistical algorithms reference guide, Technical Report, 2001.
- [7] X. He and J. Zhang, "Toward a Molecular Understanding of Pleiotropy," *Genetics*, vol. 173, pp. 1885-1891, August 1 2006.
- [8] H. Binder and S. Preibisch, "Specific and Nonspecific Hybridization of Oligonucleotide Probes on Microarrays," *Biophys. J.*, vol. 89, pp. 337-352, July 1 2005.
- [9] F. Naef and M. O. Magnasco, "Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays," *Physical Review E*, vol. 68, p. 011906, 2003.
- [10] R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, p. R80, 2004.
- [11] C. Li and W. Hung Wong, "Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application," *Genome Biology*, vol. 2, pp. research0032.1 - research0032.11, 2001.
- [12] R. A. Irizarry, Z. Wu, and H. A. Jaffee, "Comparison of Affymetrix GeneChip expression measures," *Bioinformatics*, vol. 22, pp. 789-794, April 1 2006.
- [13] S. Zakharkin, K. Kim, T. Mehta, L. Chen, S. Barnes, K. Scheirer, R. Parrish, D. Allison, and G. Page, "Sources of variation in Affymetrix microarray experiments," *BMC Bioinformatics*, vol. 6, p. 214, 2005.
- [14] Z. Wu and R. A. Irizarry, "A Statistical Framework for the Analysis of Microarray Probe-Level Data," in *Johns Hopkins*

University, Dept. of Biostatistics Working Papers, Working Paper 73. <http://www.bepress.com/jhubiostat/paper73>, March 2005.

- [15] P. Warren, D. Taylor, P. G. V. Martini, J. Jackson, and J. Bienkowska, PANP website at Brandeis University, <http://www.brandeis.edu/~dtaylor/PANP/>, 2005.
- [16] K. Kapur, Y. Xing, Z. Ouyang, and W. Wong, "Exon arrays provide accurate assessments of gene expression," *Genome Biology*, vol. 8, p. R82, 2007.
- [17] L. M. Cope, R. A. Irizarry, H. A. Jaffee, Z. Wu, and T. P. Speed, "A benchmark for Affymetrix GeneChip expression measures," *Bioinformatics*, vol. 20, pp. 323-331, February 12 2004.
- [18] R. A. Irizarry and Z. Wu, Affycomp ver. 1.4.3, Bioconductor package for R, Bioconductor Project <http://www.bioconductor.org>, 2006.
- [19] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "affy--analysis of Affymetrix GeneChip data at the probe level," *Bioinformatics*, vol. 20, pp. 307-315, February 12 2004.
- [20] Z. Wu and R. A. Irizarry, "Preprocessing of oligonucleotide array data," *Nature Biotechnology*, vol. 22, pp. 656-658, 2004.