Comparison of different probe-level analysis techniques for oligonucleotide microarrays

Barbara Rosati, Frederic Grau, Anneke Kuehler, Samantha Rodriguez, and David McKinnon

State University of New York at Stony Brook, Stony Brook, NY, USA

BioTechniques 36:316-322 (February 2004)

Three different software packages for the probe-level analysis of high-density oligonucleotide microarray data were compared using an experiment-derived data set that was validated using real-time PCR. The efficiency with which these three programs could identify true positives in this data set was assessed. In addition, estimates of false-positive and false-negative rates were determined. The performance of the programs using very small data sets was also compared, and recommendations for use are suggested.

INTRODUCTION

Three different techniques for the probe-level analysis of high-density oligonucleotide microarray data have been implemented in different software packages that are available either from Affymetrix or as freeware (1–5). These different programs have the potential to provide genuinely different approaches to the analysis of the same set of experimental data. Alternatively, it is possible that one of these approaches is significantly superior to the others, rendering use of the others irrelevant. To examine these possibilities, we have compared results obtained with the three different software packages using a typical experimental data set, in which both the number and identity of differentially expressed genes was originally unknown.

This study provides an estimate of the efficiency with which each program can identify true positives compared to the rate at which false positives are called. Positives were confirmed or rejected by use of real-time PCR. In addition, by combining the results from the different techniques, an estimate of the false-negative rate can be obtained for each analysis technique.

The approach described in this paper is different than the typical approaches taken to validating these programs, which have generally relied on data sets obtained from experimental samples into which probes of known concentrations have been spiked. Our study closely mimics the typical experimental situation, in which relatively little is known about the properties of the starting data sample.

In its simplest formulation, the concept behind the design of the Affymetrix oligonucleotide microarrays is to obtain a value for sample hybridization to a set of perfect match oligonucleotides. This, in principle, provides a measure of a gene-specific signal plus a nonspecific hybridization signal. A second value, for sample hybridization to a corresponding set of oligonucleotides that have a single base mismatch, provides an estimate of the nonspecific hybridization signal. The gene-specific signal can then be obtained by subtraction of the mismatch value from the perfect match value. The disadvantage of using both the perfect match and mismatch data sets is that subtraction of the mismatch signal adds a considerable amount of noise to the results, particularly for probes that identify genes with low expression levels (5).

Three different, commonly available, software packages were used to analyze the probe-level data. (*i*) The Affymetrix Microarray Suite (version 5.0) software package (MAS 5.0) uses a combination of rules-based data manipulation and statistical algorithms to provide estimates of several parameters. These include 95% confidence intervals of the ratios of pairwise comparisons of two data sets and presence/absence calls, which indicate whether or not a gene is considered to be expressed in that particular sample (1). (*ii*) The RMA approach as implemented in the *affy* software package (4,5) uses only the perfect match data set to determine expression values. A background correction procedure is used to reduce experimental background signal. (*iii*) The DNA-Chip Analyzer (dChip) program (2,3) uses a statistical data modeling approach to calculate a model-based expression index (MBEI) after eliminating outliers using rule-based procedures. It uses both the perfect match and mismatch data in the determination of the MBEI. The dChip software also provides a perfect match only data analysis option, but this was not used in our analysis.

All three programs were used to analyze the same data set, which was a pairwise comparison of gene expression in rat cardiac epicardium and endocardium (6).

MATERIALS AND METHODS

RNA Preparation and Analysis

Tissue samples were isolated from the epicardial and endocardial surfaces of the left ventricular free wall of rat heart and quick frozen in liquid N_2. Total RNA was prepared using RNeasy® Maxi columns (Qiagen, Valencia, CA, USA). RNA samples were tested by the SUNY Stony Brook Microarray Facility using three Rat Genome U34 (RG-U34) A, B, and C microarrays (Affymetrix, Santa Clara, CA, USA). Three independent replicates of the sample pairs were analyzed. For purposes of comparison, two sample pairs were also hybridized to the Rat Expression Set 230 (RAE230) chip set (Affymetrix).

For data presentation, expression values were transformed to log base 2, if not provided in this way by the software. Data were then presented using an Average Difference plot, which is a plot of the difference between the normalized intensity values of the two sample sets versus the mean of the normalized intensity values for each probe set.

Software and Data Analysis

MAS 5.0. The data were culled to include only those probes in which

at least one Present call was given on one of the six different arrays that contained that probe set. This procedure reduced the number of probe sets analyzed from 26,202 to 12,630. Default tunable parameter values for the Detection and Change call algorithms were used. Two different sets of criteria were used to select putative positive probes: (*i*) Threshold Only: Either the Signal Log Ratio Low was greater than 1 or the Signal Log Ratio High was less than -1 on a minimum of two of the three pairwise comparisons. These criteria resulted in the selection of 25 probe sets. Of these, 22 were tested successfully using real-time PCR. (*ii*) Difference Call and Threshold: Either the Change call was an Increase on a minimum of two chips and the Signal Log Ratio Low was greater than 0.7 on the same chips or the Change call was a Decrease on a minimum of two chips and the Signal Log Ratio High was less than -0.7 on the same chips. These criteria resulted in the selection of 28 probe sets. Of these, 25 were tested successfully using real-time PCR.

RMA. The robust multiarray average (RMA) procedure has been implemented in the *affy* package (5), which was obtained from http://www.bioconductor.org as an add-on package for the statistical software language R (7). The *affy* package can use multiple different analysis approaches, but the RMA approach for the measurement of expression values was used exclusively in this study. Expression values from all the probe sets were included in the analysis since there is no Present/Absent call feature in the software. Presumably, if a gene is differentially expressed in comparisons of the two samples, its cognate mRNA is present in at least one of the samples. The threshold was set at a difference score (in $\log_2$ values) of greater than 0.90. These criteria resulted in the selection of 21 probe sets. All of these were tested successfully using real-time PCR.

DChip. The DNA-Chip Analyzer (dChip) Version 1.2 program (2,3) was obtained from http://www.dchip.org. The expression data produced by this program were culled to include only those probe sets for which at least one Present call was given by the software on one of the six different arrays that contained that probe set. This reduced the number of probe sets analyzed from 26,202 to 13,224. A 1.7-fold change threshold criterion with a 90% confidence interval was used to select putative differentially expressed probe sets. This criterion resulted in the selection of 22 probe sets. Of these, 19 were tested successfully using real-time PCR.

Real-Time PCR

Oligonucleotides for PCR amplification were designed using the target sequences associated with positive probes, unless this sequence was too short, in which case a longer overlapping sequence was used. Real-time PCR was performed using the SYBR® Green PCR Kit (Qiagen) with a DNA Engine Opticon (MJ Research, Waltham, MA, USA). The cycle threshold (C_t) value was converted to an expression value by comparison with a standard curve using the Opticon Monitor (v.1.04) program. Experiments were performed in triplicate using three independent pairs of RNA samples. A particular mRNA was called differentially expressed if the ratio between expression levels, expressed as average of the ratios from the three independent replicates, was more than 1.5-fold. This was an absolute threshold; no statistical criteria were used. Real-time PCR products were sequenced to confirm that the amplicons were from the mRNA of interest. A total of 41 different probe sets were successfully tested using real-time PCR. Of these, 32 were true positives and 9 were false positives.

RESULTS

Average Difference plots of the expression values produced by each program were prepared (Figure 1). The MAS 5.0 software produces a distribution of expression values whose variance increases very markedly for probe sets that identify poorly expressed genes (Figure 1, A and B). The MAS 5.0 expression value distribution is very noisy even though it has been averaged ($n = 3$), and the most variable data points, those called Absent, have

been eliminated. In contrast, the RMA approach produced a much less noisy distribution of expression values, with less and more even variance across the distribution, even though all of the data points are included in the analysis (Figure 1C). The dChip program produced an intermediate result after elimination of the data points called Absent (Figure 1D).

Identification of True and False Positives

To produce a fair comparison between the different algorithms, threshold criteria for each program were set to identify an approximately equal number of putative differentially expressed genes (see Materials and Methods).

Generally, all three programs performed satisfactorily. There was never a complete consensus between any two selected sets of probe sets. Typically, each program identified most or all of the most differentially expressed genes but identified a varying set of the genes that were less differentially expressed between the two tissues.

Table 1. Comparison of True-Positive Calls and False-Positive Calls

	MAS 5.0		RMA	dChip
	Threshold only	Difference call and threshold		
True-positive rate	82% (18/22)	80% (20/25)	100% (21/21)	89% (17/19)
False-positive rate	18% (4/22)	20% (5/25)	0% (0/21)	11% (2/19)

The MAS 5.0 protocol performed well using either the threshold only or difference call and threshold selection criteria. In both cases, approximately 80% of the calls were true positives (Table 1). Consequently, false-positive rates were approximately 20% for both procedures. For the threshold only selection criterion, the false positives were primarily found at low signal levels, reflecting the much greater noise

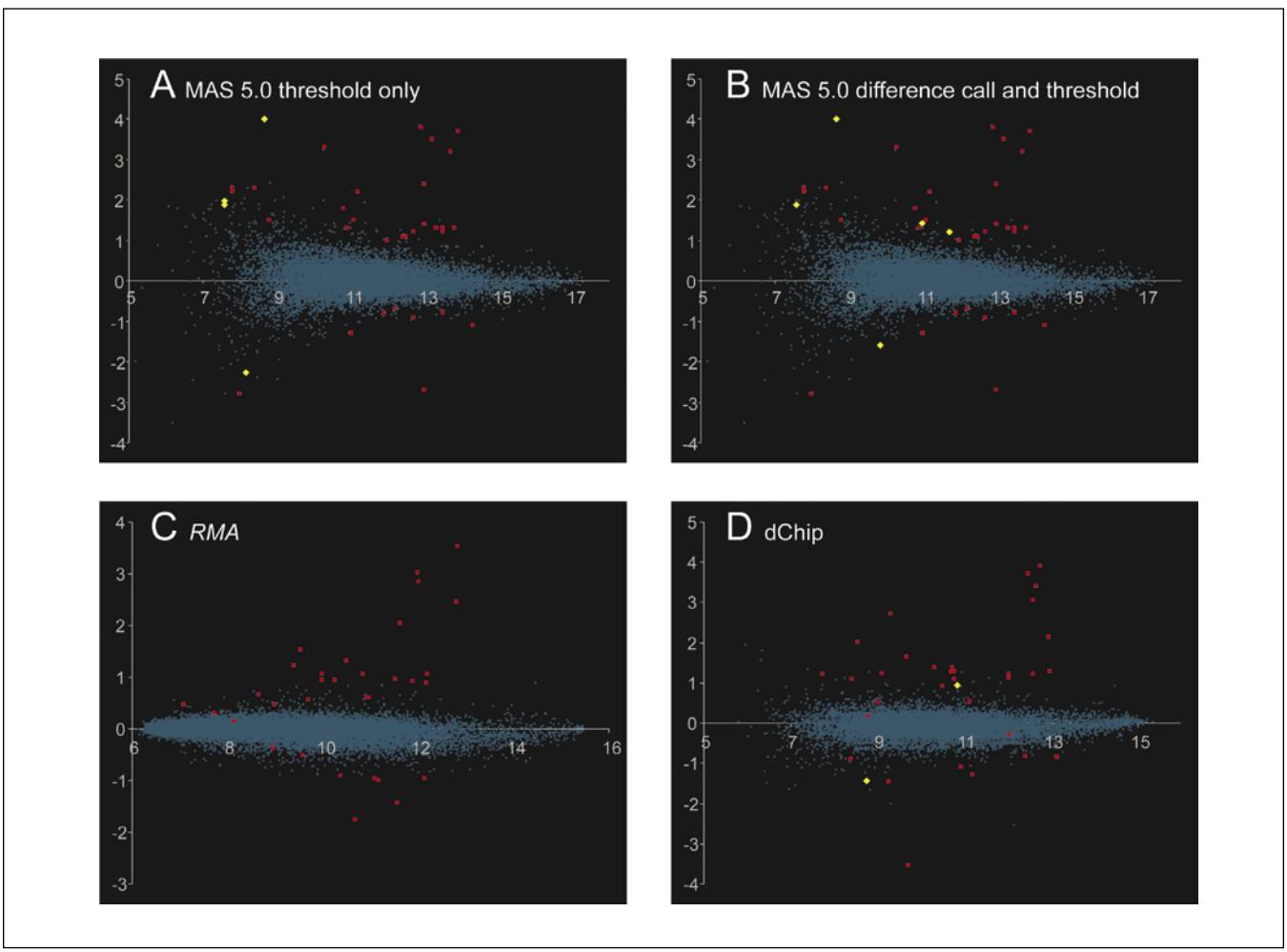

Figure 1. Average Difference plots of gene expression in rat heart. Average Difference plots of expression values produced by (A and B) the MAS 5.0, (C) RMA (*affy*), and (D) dChip programs. All data points are shown as blue dots. The data values for the MAS 5.0 and dChip programs were selected to include only those probe sets that received at least one Present call. True positives are identified as red squares in each distribution. False positives for (A) MAS 5.0 Threshold only, (B) MAS 5.0 Difference call and threshold, (C) RMA (*affy*), and (D) dChip programs are identified as yellow diamonds. The expression values are transformed to $\log_2$. The x-axis represents the average of the two sample values and the y-axis represents the difference between the two values.

in this region of the data (Figure 1A). For the difference call combined with threshold level criteria, the false positives were more evenly spread throughout the distribution (Figure 1B).

The RMA approach proved to be very conservative, identifying 100% true positives for the top 21 probe sets that it selected (Table 1).

The dChip program produced a better level of true positive calls than the MAS 5.0 software and a correspondingly lower false-positive rate (Table 1).

The rate of false positives is not an absolute value since it is strongly dependent on the threshold values (8). The threshold values that were used for the three programs cannot be directly compared because the distribution of expression values produced by each program is quite different. Setting thresholds to produce an arbitrary number of positive calls (in this case approximately 22) is a fair way to compare the performance of the programs; however, the false-positive rate for each program will obviously rise as the threshold values are made increasingly more liberal.

False Negatives

There are two classes of false negatives, those for which the Affymetrix chip is capable of detecting gene expression but the analysis program fails to detect a true difference in gene expression, and those for which the chip lacks sufficient sensitivity to detect gene expression and the data is consequently ignored.

False negatives due to the analysis procedure. To assess the false-negative rate for each program, all the probe sets whose differential expression was confirmed using real-time PCR were included in the analysis. The expression values produced by each program were plotted on an Average Difference plot, and the true positives were identified within this distribution (Figure 1). True positives that were located in the extremes of the distribution of difference values were considered to have a reasonable probability of being identified, assuming that a moderately large number of probe sets were tested using real-time PCR. True positives that were located within the main part of the difference value distribution were considered unlikely to be identified by any statistical criteria. These data points were classified as false negatives for that particular distribution.

The MAS 5.0 program was very efficient at identifying true positives. The distribution of true positives within the complete distribution of the data points generated by this program (Figures 1, A and B) shows that almost all of the true-positive data points lie in regions of the distribution that could be reasonably selected using statistical criteria. This suggests that the false-positive rate for this procedure is very low, close to

zero, given a willingness to endure an increasing rate of false negatives.

The RMA approach clearly misses some true positives (Figure 1C). Approximately 13% (4/32) of the true-positive probe sets would be missed using even very liberal selection criteria.

The dChip program also misses some true positives, although it appears to perform slightly better than the RMA procedure in this respect.

False negatives due to lack of probe set sensitivity. The ability of individual probe sets to detect a given class of transcripts is strongly dependent upon the relative abundance of the transcripts. High and moderate abundance transcripts are likely to be detected with a much greater reliability than low abundance transcripts. To assess the performance of the oligonucleotide microarrays with a class of low abundance transcripts the rate of correct expression calls for potassium channel genes was determined. Potassium channel genes are generally expressed at low levels in rat heart, even though the expressed channels are critical determinants of the electrical properties of the heart. We have previously made a quantitative study of potassium channel gene expression in rat heart using RNase protection assays (Reference 6 and unpublished results).

A total of nine potassium channel genes that are expressed in heart are represented by 15 probe sets on the RG-U34 chips (there is some redundancy in the chip design). The performance of these probe sets was quite poor, with only 7% of the expression calls being correct (i.e., positive). Only 3 of 15 (20%) of the probe sets recorded one or more positive calls using the MAS 5.0 software. The dChip software gave a marginally better result (27% correct).

The relatively poor performance of the RG-U34 chips prompted us to test a new rat chip set (RAE230) that is claimed to provide a modest improvement in sensitivity over the RG-U34 chip set. Two of the same sample pairs that were used on the RG-U34 chips were hybridized to the RAE230 chip sets. For the same subset of potassium channel genes, the RAE230 chip set yielded a 35% correct call rate using the MAS 5.0 software, resulting in 58% of the genes receiving at least one Present call. This is a significant improvement over the older chip design, although it is still less than optimal. This particular test exaggerates the increased sensitivity of the new chip design because it focuses exclusively on a set of genes that are expressed close to the detection threshold for both chip sets, thereby amplifying any differences in sensitivity.

Use of Small Sample Sizes

There are many experimental situations where there is a very limited data set available, either due to limited starting material and cost considerations, or the experiments are preliminary, designed to test feasibility. In preliminary experiments using a single replicate it was found that the MAS 5.0 software had a very high rate of false-positive calls. This was due, in large part, to the high noise levels for low abundance transcripts (Figure 2A). With the MAS 5.0 software, it is necessary to average a minimum of three replicates to obtain acceptable performance (Figure 2B). In contrast, the RMA approach can provide useful results with a single replicate (Figure 2C) because of the much lower noise levels. The elimination of the mismatch data significantly improves the signal-to-noise ratio, particularly in the lowest range of signal strengths, making this program the best choice in this situation.

DISCUSSION

Given the relatively high expense of oligonucleotide microarray experiments for a typical small laboratory, it is essential to ensure that the vast majority of differentially expressed genes are detected in comparisons of a given set of unknown samples. On the other hand, it is important to use an analysis technique that minimizes the number of false positives because it is both time-consuming and expensive to follow up false leads. It was with this in mind that the different analysis techniques were compared.

There were modest differences in the results produced by the three programs. Not unexpectedly, there was a tradeoff in the rates of false-positive and false-negative calls. For example, the RMA approach had a very low rate

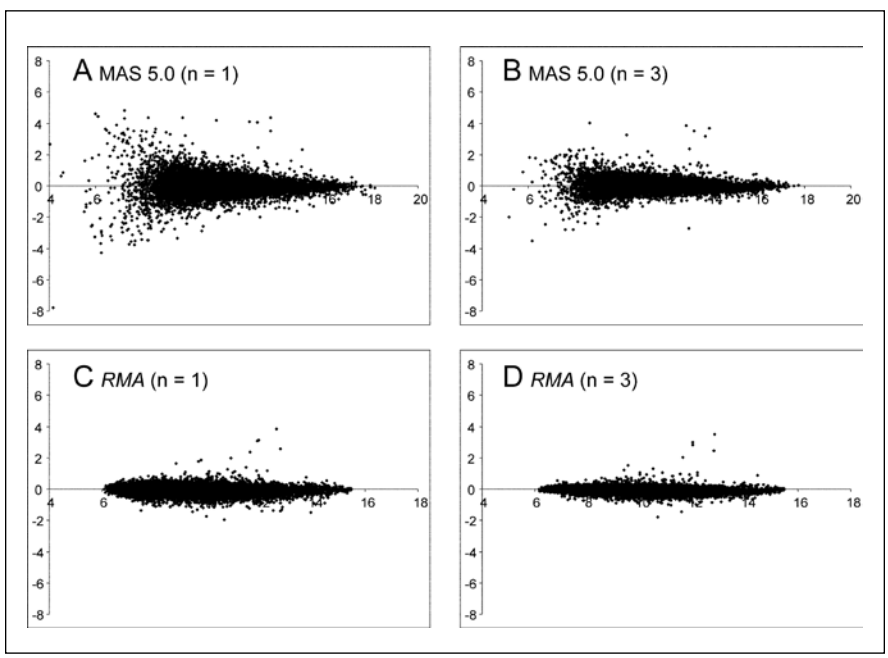

Figure 2. Distribution plots of gene expression values in rat heart. The effect of sample size on the distribution of expression values produced by (A and B) the MAS 5.0 and (C and D) RMA (*affy*) programs are compared. Either one pair of samples (A and C) or three pairs of samples (B and D) were included in the analysis. Data are presented on Average Difference plots with the x-axis representing the average of the sample values and the y-axis representing the difference between the values.

of false-positive calls. On the other hand, its rate of false-negative calls was higher than for the other two programs, and there were clearly some true-positive probe sets that it could not identify. In this respect, it had the most difficulty with probe sets that identified relatively low abundance mRNAs. In contrast, the MAS 5.0 program, which identified all of the true positives tested, produced quite a few false-positive calls, especially in the low abundance region of the distribution. This false-positive rate increases rapidly as the number of probe sets selected is increased by the use of more liberal selection criteria, gradually mitigating the advantage of this program. The dChip program produced intermediate results and provided no obvious advantage over the other two programs in this respect.

The list of top ranked probe sets was somewhat different for each program, with no procedure producing a completely overlapping set of positive calls. For this reason, the easiest approach to achieve the best yield of true positives may be to simply use all three programs and to skim off the top ranking probe sets produced by each program.

Although the three programs were generally similar in their performance, the RMA approach as implemented in the *affy* program had one distinct advantage. Because of the low noise in the expression value distribution produced by the RMA approach, due to the elimination of the mismatch probe data, this program is particularly useful when only a single data set is available (9). Although the RMA approach has the disadvantage of missing more true positives than the other approaches, this is more than compensated for by the superior signal-to-noise ratio in the absence of averaging. If only a single comparison can be used, for reasons of cost or sample availability, or when performing a trial experiment, then this is the best procedure to use to first assess the data. The MAS 5.0 program produces an unacceptably high level of false positives with a single replicate. This advantage of the RMA approach over the other two programs is gradually lost with an increasing number of replicates because averaging of three or more independent replicates significantly improves the signal to noise ratio for the other two programs.

Although the elimination of the mismatch data by the RMA approach has some benefits, it also introduces one significant disadvantage. Some true positives are completely missed, particularly in the low abundance region of the distribution. The mismatch data adds useful information, at least for a fraction of the probes in this region. Examination of individual probe sets suggests that the effect of abundant, nonspecific transcripts incorrectly binding to the perfect match probes is eliminated at least some of the time by the MAS 5.0 program, presumably because the same transcript binds with similar or higher avidity to the mismatch probe. In the absence of the mismatch data, the only way to edit out the effect of perfect match probes that introduce a spurious signal due to binding of nonspecific transcripts is by statistical criteria, which may be difficult to calibrate accurately for every case.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for several useful suggestions. This study was supported by grants HL-28958, NS-29755, and AHA-0235467T.

REFERENCES

1. **Affymetrix Technical Report.** 2001. Statistical algorithms reference guide. Affymetrix, Santa Clara, CA, USA.
2. **Li, C. and W.H. Wong.** 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc. Natl. Acad. Sci. USA *98*:31-36.
3. **Li, C. and W.H. Wong.** 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biology *2*:research0032.1-0032.11.
4. **Bolstad, B.M., R.A. Irizarry, M. Astrand, and T.P. Speed.** 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*:185-193.
5. **Irizarry, RA, L. Gautier, and L. Cope.** 2003. An R Package for Analyses of Affymetrix Oligonucleotide Arrays, p. 45-52. *In* G. Parmigiani, E.S. Garrett, R.A. Irizarry, and S.L. Zeger (Eds.), The Analysis of Gene Expression Data: Methods and Software. Springer, Berlin.
6. **Dixon, J.E. and D. McKinnon.** 1994. Quantitative analysis of potassium channel mRNA expression in atrial and ventricular muscle of rats. Circ. Res. *75*:252-260.
7. **Ihaka, R. and R. Gentleman.** 1996. R: A language for data analysis and graphics. J. Comput. Graph. Stat. *5*:299-314.
8. **Irizarry, R.A., B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed.** 2003. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. *31*:e15.
9. **Rosati, B., F. Grau, S. Rodriguez, H. Li, J.M. Nerbonne, and D. McKinnon.** 2003. Concordant expression of KChIP2 mRNA, protein and transient outward current throughout the canine ventricle. J. Physiol. *548*: 815-822.

Received 22 April 2003; accepted 8 October 2003.

Address correspondence to David McKinnon, Department of Physiology and Biophysics, BST Room 124, Level 6, SUNY, Stony Brook, NY 11794-8661, USA, e-mail: dmckinnon@notes.cc.sunysb.edu