 Human Molecular Genetics, 2002, Vol. 11, No. 19 2207–2221

Estimating the statistical significance of gene expression changes observed with oligonucleotide arrays†

Andrew D. Strand[1], James M. Olson[1] and Charles Kooperberg[2,*]

[1]Clinical Research Division and [2]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, PO Box 19024, MP 1002, Seattle, WA 98109-1024, USA

Received March 12, 2002; Revised and Accepted July 11, 2002

We present a simple method to assign approximate *P*-values to gene expression changes detected with Affymetrix oligonucleotide arrays and software. The method pools data for groups of genes and a small number of like-to-like comparisons in order to estimate the significance of changes observed for single genes in comparisons of experimental interest. Statistical significance levels are based on the observed variability in the fractional majority of probe pairs that indicate increasing or decreasing differential expression in comparisons of technical replicates. From this reference distribution or error model, we compute the expected frequency for fractional majorities in comparisons for $N \geq 2$. These computed distributions are the source of *P*-value estimates for changes seen in the experimental comparisons. The method is intended to complement the Affymetrix software and to rationalize gene selection for experimental designs involving limited replication.

INTRODUCTION

The papers of the Hereditary Disease Array Group represent a cooperative effort to learn more about neurodegenerative diseases caused by expanded polyglutamine [poly(Q)] tracts (1–4). These studies were begun with the hope that gene expression profiling of mouse and cell line models would reveal the earliest events in neural dysfunction, elucidate the course of developing disease, and stimulate fresh hypotheses. Choosing a common method for selecting differentially expressed genes was an issue that had to be addressed before results in different model systems could be compared.

In the few years that DNA microarrays have been available, no method of identifying differentially expressed genes in two populations has achieved general use and acceptance. Application of standard statistical tests to determine differential gene expression, such as *t*-statistics, has been hampered by the fact that there are often few replicates, owing to constraints on funding and samples. With few replicates, these tests have few degrees of freedom and thus little power to discern differences between groups. In place of statistical criteria, arbitrary thresholds (whereby a gene is identified as differentially expressed if the ratio of expression in one sample relative to the other exceeds a certain magnitude) are commonly applied (5,6). Ratio-based criteria are certainly not optimal because of several well-known flaws. The relative error increases as the signal decreases and fixed thresholds remove changes below the limit from further consideration, thus negating the increased power that one should derive from replicating experiments.

Another approach is to determine thresholds for differential expression based on empirical observation. Here, spiking experiments are performed in order to calibrate differential expression in terms of the microarray's sensitivity and discrimination. Thorough use of this approach is beyond the scope of individual laboratories using commercial arrays, so researchers rely on empirical thresholds as determined by the array manufacturer (7–9). The Affymetrix Microarray Suite GeneChip 4.0 software comparison analysis provides this empirically based assessment of differential expression as a Difference Call. The Difference Call is generated in comparison analysis of two arrays, and consistent calls in comparisons have been found to be a reliable indicator of differential expression (10–12). However, no statistical significance estimate is provided, and it is not clear how one should weigh changes that are called in a fraction of the relevant comparisons.

Statistical inference would select genes for further scrutiny by the likelihood that their apparent changes did not occur by chance. One reasonable way to create gene lists from different experiments prior to comparing and contrasting the lists is to

*To whom correspondence should be addressed. Tel: +1 2066677808; Fax: +1 2066674142; Email: clk@fhcrc.org
†This paper is part of the Microarray Report Special Series. See, Orr H.J. (2002) *Hum. Mol. Genet.*, **11**: 1909–1910.

set a common threshold of statistical significance that each gene must pass before it finds its way on to the final list. An alternative method to use when selecting groups of genes is to control the false-discovery rate (FDR) (13). The FDR can be estimated by comparing the observed number of genes at a given significance level with the number expected by chance. However, standard *t*-statistics lack power in cases of limited replication, which has led to the development of statistical tests combining data on many genes to estimate the significance of changes for individual genes (14–17).

We present a method for estimating the statistical significance of gene expression changes measured with Affymetrix oligonucleotide arrays and software. False-positive rates and *P*-values are based on variability observed in comparisons of replicate samples. This method uses output provided by Affymetrix Microarray Suite (MAS) 4.0 GeneChip software. The method augments the Affymetrix software in that it accounts for replication and is most useful for experiments involving small numbers of replicates, i.e. $N = 2, \ldots, \sim 6$. The approach can be easily adapted in order to accommodate changes in the arrays, hybridization conditions or image analysis software.

RESULTS

Background and rationale

The Hereditary Disease Array Group, a consortium investigating transcriptional dysregulation in neurodegenerative diseases, generated a large number (>200 in the studies considered here) of expression profiles of cerebral cortex, cerebellum, and striatum from different poly(Q) disease mouse models (1–4). The main experimental question of interest for each model was how did the transgenic line carrying the expanded poly(Q) tract differ from its non-transgenic siblings or transgenic lines carrying normal-length poly(Q) tracts. If there were evidence of abnormal expression profiles in different models, it would also be interesting to compare the changes to see what was common and unique to each model or tissue.

While in total there was a very large amount of data, each experiment was done using a small number of replicates. As an alternative to arbitrary thresholds or *t*-statistics, we explored error models of different quantities from the Affymetrix MAS 4.0 output in order to define differentially expressed genes. Again, owing to limited replication, there were not enough data to make truly useful experiment-specific error models; therefore, in order to effectively increase the number of degrees of freedom available to analyze each small experiment, we made tissue-specific error models by pooling data from comparisons of replicates in the different experiments. With such a reference, changes seen in the experimental comparisons of interest could be related to how often similar or larger changes were seen due to technical and normal biological variability.

In comparisons of technical replicates, no experimental variables of interest are involved, and variables such as the age, genotype, tissue and strain of the mice are held constant. To control technical sources of variability, procedures had been standardized. A single person had done all sample preparation, and all arrays were processed at the FHCRC Array Facility. Tissue-specific models can be rationalized, since in each tissue some genes are uniquely expressed, and common genes are expressed at different levels. Differences between mouse brain tissues have also been shown to be more extensive than normal variation within a tissue (18).

Affymetrix arrays and analysis

Affymetrix oligonucleotide arrays, data processing and empirical analysis algorithms have been described in detail elsewhere (8,9). Briefly, each gene is represented by multiple 25-base oligonucleotides called probes, synthesized in square tiles on the surface of a silicon wafer. The probes come in pairs. One probe, the Perfect Match (PM), is designed to be complementary to a reference sequence. The second probe, or Mismatch (MM), is meant to control for cross-hybridization and contains a homomeric mismatch at the central position. A biotinylated cRNA derived from the mRNA population in a single sample is hybridized to each array. The surrogate expression level for a gene, its Average Difference, is measured by summing the differences between the PM and MM signals and dividing that sum by the number of probe pairs used in the calculation (9). On the mouse arrays used in these studies, 20 probe pairs typically represent genes.

The Affymetrix software can consider each array separately or can directly compare two arrays. In a comparison, the reference sample is called the baseline and the sample being compared with the reference is the experiment. The size of the gene expression change may be expressed as the averaged difference in intensity units between the matching probe pairs, the Average Difference Change, a relative Fold Change or, perhaps best, as the logarithm of the ratio of Average Differences. Changes judged significant based on empirical thresholds of sensitivity and selectivity as determined by Affymetrix are indicated by a Difference Call.

The MAS 4.0 empirical Difference Call is based on four elements, three of which are related to the number of a gene's probe pairs that are deemed higher or lower in the experiment than they are in the baseline. The cutoff for the probe pair increase–decrease decision is called the Change Threshold. This value is a dynamic number calculated from the noise or pixel-to-pixel variation in the two images being compared. Spiking experiments performed by Affymetrix informed the magnitude of this threshold as it relates to the noise (7–9).

The Difference Call has been shown to be a reliable indicator of differential gene expression, but it is not associated with any particular significance level, thus limiting its usefulness. We considered and rejected as too granular an error model based on the Difference Calls. In our largest set of replicate comparisons (32 comparisons), 45% of the genes were never called and 25% received only one call. The fact that so few genes receive calls and the binary nature of the Difference Call make it difficult to establish a meaningful distribution for estimating *P*-values. We then explored how the geometric mean Fold Change, mean Average Difference Change and the fractional majority of the probe pairs indicated as 'increased' or 'decreased' in the comparisons related to the Difference Call. Figure 1 clearly shows that the latter has the strongest correlation with the Difference Call.

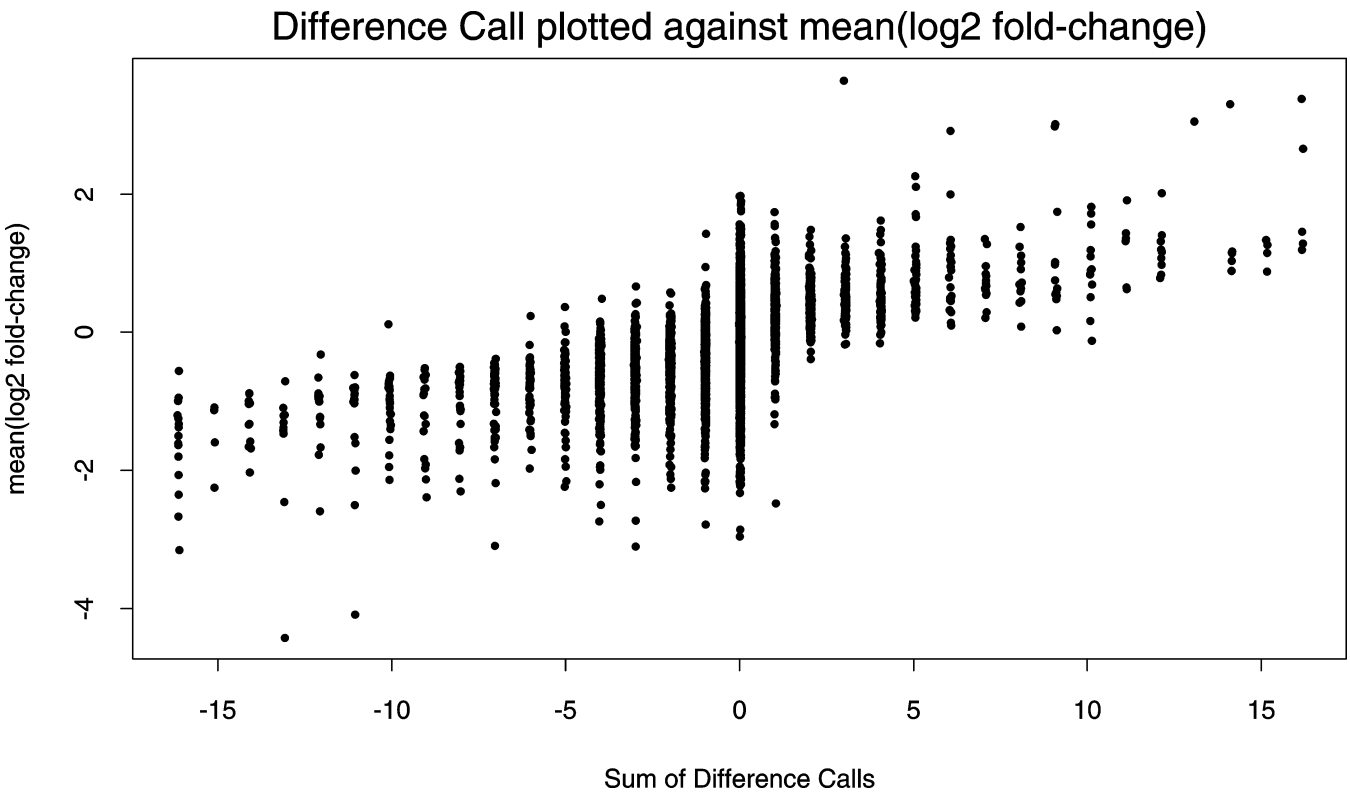

Figure 1. Relationship of the Affymetrix empirical Difference Call to various MAS 4.0 outputs. The number of Difference Calls for each gene from the DRPLA 65Q versus wild-type cerebellum comparisons were tabulated by converting each Decrease call to -1 and each Increase call to $+1$ and summing over the 16 comparisons (data from 3). These are plotted on the horizontal axis of each graph. Averaged values for other comparison analysis quantities are plotted on the vertical axis. (**A**) Difference Calls versus mean ($\log_2$ Fold Change). (**B**) Difference Calls versus mean Average Difference Change. (**C**) Difference Calls versus mean (Increase Ratio − Decrease Ratio).

We illustrate some properties of the Average Difference as determined by MAS 4.0 that also led us to consider other quantities for modeling. Figure 2 shows the 16 PM values for a randomly selected gene from the 36 different arrays, with data from Sipione *et al.* (4). It is clear that some tiles are always high or always low relative to the other tiles. The effect of taking averages, as when computing the Average Difference, is that tiles with low signals are effectively ignored. As the correlation coefficients of the lowest 12 probes with the highest four probes are usually very high (∼0.9 for this gene; data not shown), many weaker probes contain useful information. Further evidence of how the Average Difference tends to be dominated by the strongest probes is shown in Figure 3. The horizontal axis in this plot is the average PM signal for the gene shown in Figure 2 from the same 36 arrays, and shown on the vertical axis is the average signal after randomly allotting to each array the 12 tiles with the lowest values. Scrambling the lower 12 out of 16 tiles has no effect on the ordering of the arrays. In effect, the Average Difference for most genes is based on a fraction of the available information.

Error model

Comparisons between replicate samples were used to assess the variability of the number of probe pairs that increased and decreased in the mouse brain gene expression profiles. Specifically, we were interested in the difference of the Increase and Decrease counts, which is essentially the majority of the probe pairs that detected or 'voted' for a change due to variability in sample preparation, arrays and normal gene expression. In particular, 32 striatum, 18 cerebellum and 6 cortex like-to-like comparisons on Affymetrix Mu11K A and Mu11K B oligonucleotide arrays were generated. These comparisons all involved independent pairwise comparisons, i.e. no sample was considered twice on a particular type of array.

As we wished to make tissue-specific error models, genes were ranked by their mean Average Difference so as to reflect the gene expression profile of each tissue. Technical components of variability, noise at the low end and saturation at the high end, also depend upon signal strength. Figure 4 shows that the number of probe pairs that changed, as determined by the MAS 4.0 default settings, in the like-to-like comparisons depended on the Average Difference. The numbers of Increased, Decreased and Increased minus Decreased probe pairs showed a similar, though less striking, dependence on the Average Difference (not shown). To reduce this dependence, genes were considered in bins. In a Bayesian procedure for regularized *t*-tests for differential expression, genes were also considered in bins after ranking by expression level (16).

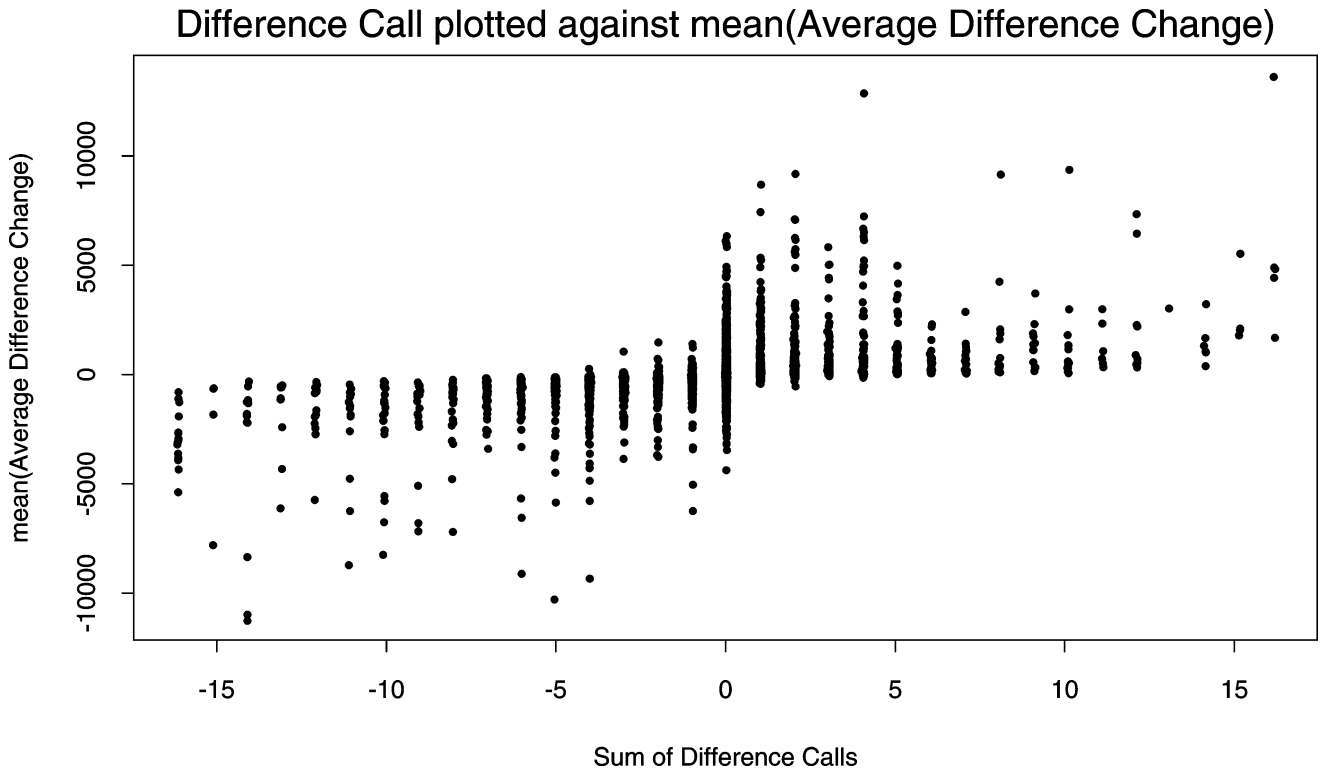

Figure 1B continued.

In binning, it is assumed that few genes normally behave in a highly variable manner. This is an implicit assumption of global image scaling, our standard procedure, and in most spotted cDNA microarray normalizations (19). Microarray studies of normal gene expression variability in yeast and mice suggest this a reasonably valid assumption (20,21). In addition, our analysis of the Difference Call frequency (a crude tissue- and gene-specific error model) indicated that few genes varied in a significant fraction of the comparisons. Most of these variable genes were evenly distributed across the top 40% of genes as ranked by their mean Average Difference. The effect of these genes, if any, is to overestimate the true variation for the majority of genes in the bin by a negligible amount. This in turn leads to slightly conservative *P*-value estimates, so that in the end fewer genes might be called significant.

Choosing the optimal number of bins is a traditional bias–variance tradeoff that is similar to selecting the bandwidth for kernel methods and local polynomial regression (22). Too many bins will lead to highly variable estimates of the *P*-values, while too few will lead to biased estimates. In theory, the number of bins should depend both on the number of technical replicates and on the number of genes on the array. While bandwidths are sometimes selected using cross-validation, such an approach is of little help in our situation. Theoretical rate-of-convergence results suggest that the optimal number of bins would increase only modestly as the number of replicates increases.

P-value threshold curves for the striatum data using various bin numbers are shown in Figure 5, which also provides further motivation for why we did not develop a gene-specific error model. Because 25 bins, each with ~260 genes, did not yield smooth *P*-value estimates, clearly there were not enough data to obtain stable estimates using single genes. A sliding-window bin could have smoothed the *P*-value threshold curve but would have required an order of magnitude more computation. We chose 10 bins as a convenient number and reasonable compromise between bias and variance.

In each bin, all the (Increase − Decrease) values were combined. For the striatum data, this gave ~21 000 realizations per bin, since there are ~660 genes in each decile, and there were 32 striatum like-to-like comparisons. After tabulating the frequencies of the values in the bins, the exact sampling distributions for various *N* were computed. This computation is greatly facilitated by the fact that only a discrete number of changes are possible, since virtually all genes have the same number of probe pairs. To accommodate all genes on the array, the changes were normalized by the usual number of probes per gene so that (Increase − Decrease) becomes (Increase Ratio − Decrease Ratio). The approximate *P*-values assigned to the absolute value of the mean experiment-to-baseline (Increase Ratio − Decrease Ratio) are based on these computed sampling distributions.

The approximate *P*-values associated with (Increase Ratio − Decrease Ratio) averages of different sizes in mouse

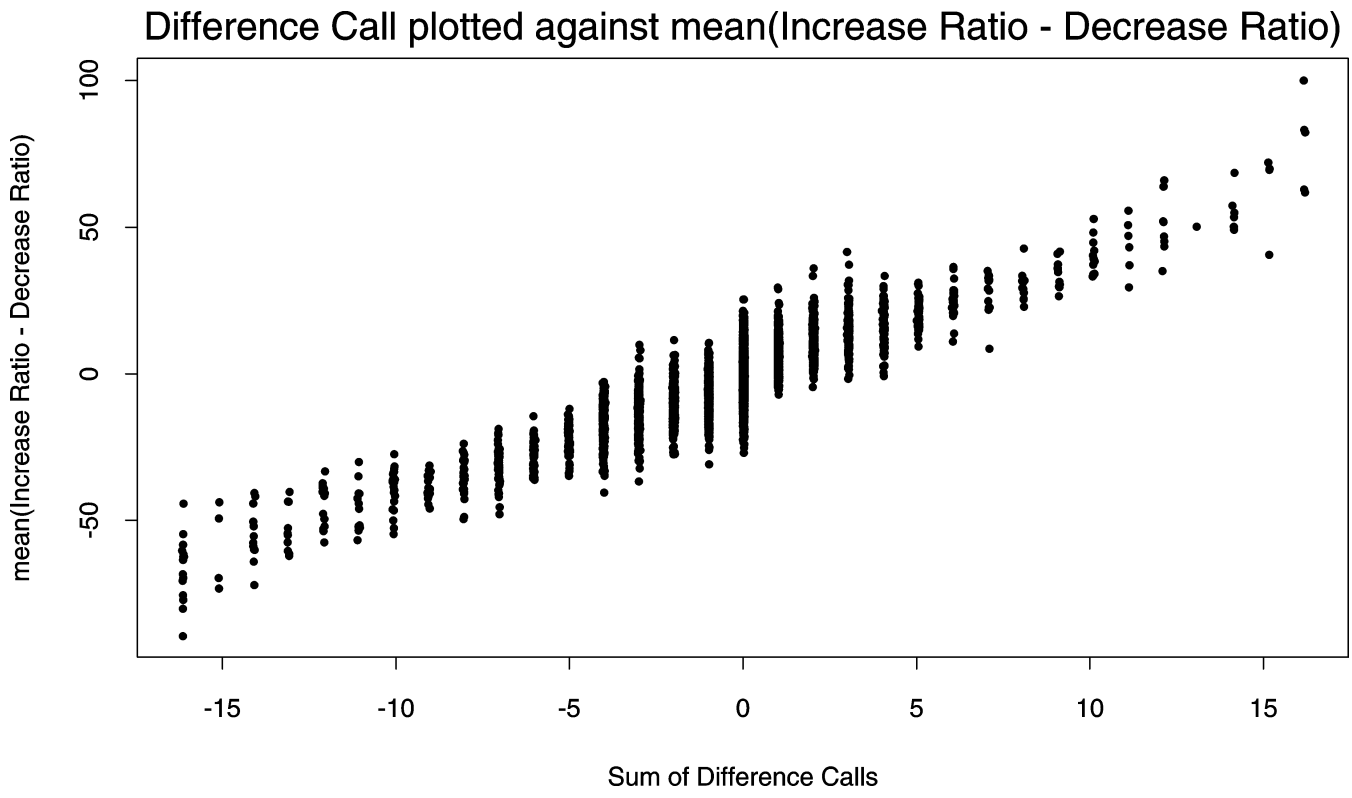

Figure 1C continued.

striatum, cerebellum, and cortex are shown in Table 1. We show values only for $N = 2$ and 3. Tables for $N = 2, \ldots, 6$ are available in the Supplemental Data (www.neumetrix.info).

In the striatum and cerebellum models, which were based on larger samples of technical variability than the cortex model, values within a bin associated with a given significance fall in a narrow range. The similarity of the striatum and cerebellum models along with the effect of increasing N are shown in Figure 6. The more lenient thresholds of the cortex error model reflect that small experiment's low technical variability. The cortex error model thresholds at a given P-value are also somewhat flatter across all bins, while the striatum and cerebellum model thresholds vary with the bin. In those two models, genes with stronger signals have higher thresholds than genes with lower signals.

Correspondence with MAS 4.0

In Figure 1, it was seen that the Difference Call is essentially a function of the fractional majority of probe pairs that indicate a change in a particular direction. The benefits of our method over the Difference Call are that it accounts for replication and provides statistical significance estimates. For example, using data from Luthi-Carter *et al.* (3), in the R6/2 cerebellum comparisons, a P-value ≤ 0.001 identifies 183 genes, and captures 97% of the genes with four calls, 68% of those with three, 10% of those with two, and a small fraction of the rest. It could be argued that *ad hoc* 'three out of four' criteria, which would have identified 170 genes, would have done as well as our method in this situation. However, applying this *ad hoc* cutoff to the DRPLA Q65 cerebellum experiment done at $N = 4$ would seem to be much too conservative. The P-value estimates identify 469 genes at $P < 0.001$. This cutoff captures every gene that received ≥ 9 calls in the16 comparisons and the majority of genes with 6–8 calls, and identifies 388 more genes than a 'three out of four' criterion.

Confirmation of microarray results

Detailed descriptions of confirmation studies are given in (1–4), so only summary data are presented here. Most of the genes that have been selected for follow-up confirmatory analysis by northern blot or quantitative RT–PCR have an assigned significance of $P < 0.001$. Using this as our cutoff, we display in Table 2 the number of probe sets at that level in the different studies and the expected confirmation rate (1 − FDR). Below the expected confirmation rate, the results of confirmatory assays are shown. If P-values were calculated for northern or RT–PCR assays, then $P < 0.05$ had to be met before the gene was considered as confirmed.

In confirmatory studies, ~50 genes were examined in several tissues and models, resulting in >70 tests to see if genes

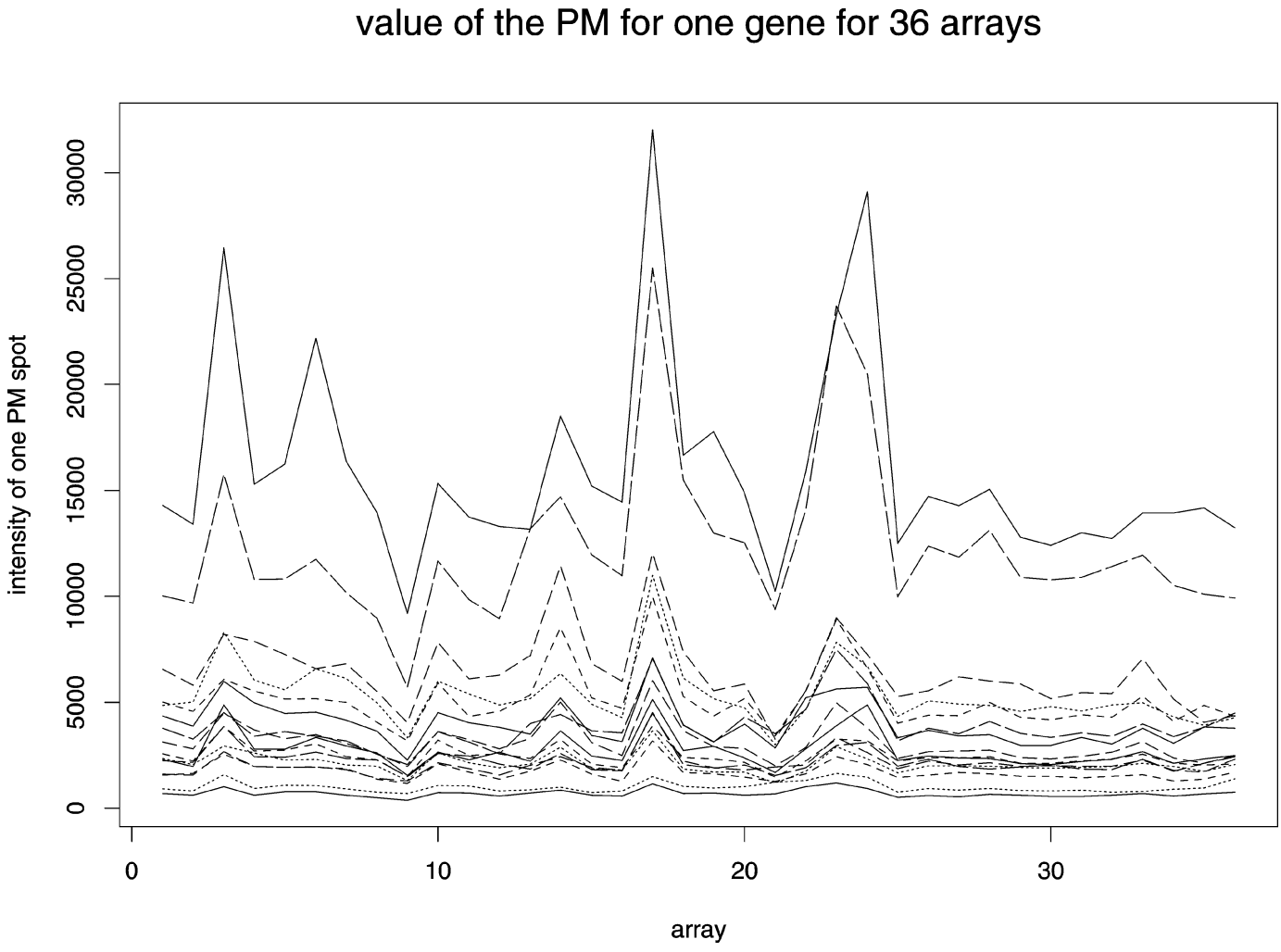

Figure 2. The Perfect Match (PM) values for a randomly chosen gene on 36 different arrays. Each line represents the signal from a particular PM tile on 36 separate arrays (data from 4). Very few of the lines cross. This means that on each of the 36 arrays, the rank order of the PM signals is nearly the same. In addition, each line looks quite similar to the other lines. Signal values for the lower, less strongly reporting PM tiles accurately reflect the more robust PM tiles, and can provide useful information.

detected at $P < 0.001$ by microarray actually varied. The results of these tests represent the most extensive examination of a gene selection method of which we are aware. Genes that were not confirmed as changing by northern blot or RT–PCR usually showed a trend in the same direction suggested by the array data, but failed to meet the $P < 0.05$ threshold. Overall, the confirmation rates within experiments are essentially as expected from the FDR estimates. Confirmation is worst where there are few predicted changes, as in younger R6/2 mice (2) and in the Aronin and YAC data (1). We note that the Aronin and YAC studies and less statistically powerful studies on 12-week-old R6/2 mice predict essentially the same numbers of gene changes. This suggests that gene expression changes in the Aronin and YAC mice are less severe in the examined tissues than changes seen in the R6/2 model, and thus Aronin and YAC changes are likely to be harder to detect by secondary methods.

DISCUSSION

Every reasonable gene selection method captures different parts of the true set of differentially expressed genes, while falsely identifying or missing others. Statistical methods provide means for controlling the effect that these two types of errors have on the final list of genes. The method outlined here was designed to assign approximate significance levels to differences observed in oligonucleotide microarray experiments using statistics supplied by Affymetrix MAS 4.0 software. We developed this method specifically for the microarray experiments of the Hereditary Disease Array Group (1–4), but the approach can be used with any set of Affymetrix microarray data. Researchers with few replicates might pool data using the same array type and biological material in order to make error models. The confirmation studies conducted by the Hereditary Disease Array Group consortium are the most thorough test of

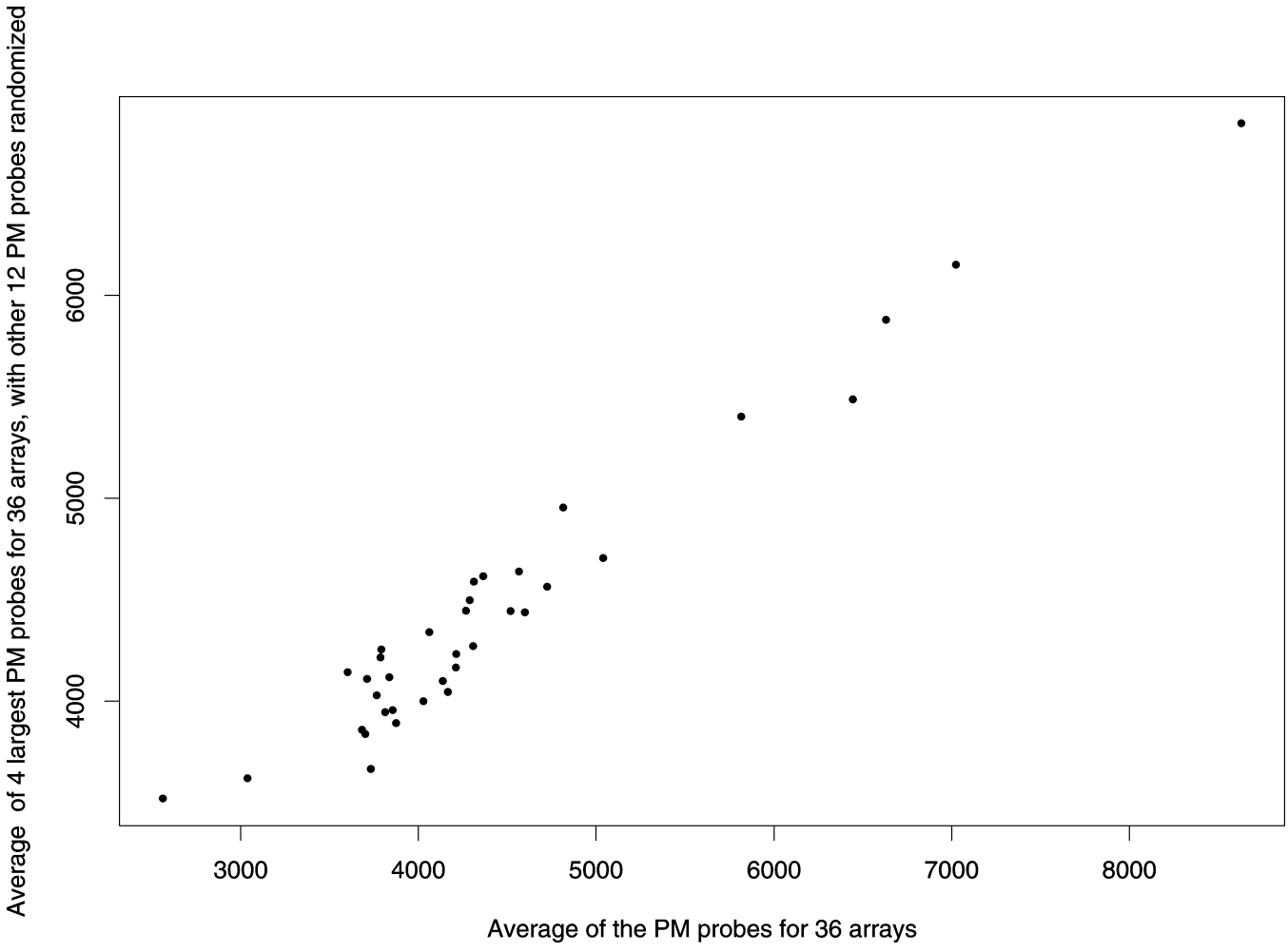

Figure 3. Scatterplot of the average PM signal for a gene on 36 arrays before and after randomizing the 12 lowest signal PM tiles. Using the same gene and 36 arrays from Figure 2 (data from 4), we plot on the horizontal axis the average PM signal from the actual data. On the vertical axis, we plot the average PM signal after mixing the 12 lowest PM signals from all arrays and randomly reassigning them to different arrays. The correlation between the data before and after scrambling is 0.975. Thus the average is dominated by the largest PM values.

a method for selecting differentially expressed genes of which we are aware, and show this method is as successful, if not better, than other methods.

At first glance, our choice of statistic to model, the fractional majority of probe pairs that indicate a change, seems somewhat unusual. The net number of probe pairs that increase or decrease provides little information as to the magnitude of the change, and thus some information is lost. Intuitively, however, the likelihood that a real difference is being detected should strongly correspond with the number of probes that indicate the difference. The most readily available quantities in the GeneChip output that relate to the individual probe pair changes in a comparison are the number that 'increase' or 'decrease' more than the Change Threshold. We have observed that this is the statistic in the MAS 4.0 output that most strongly correlates with the empirical Difference Call. Furthermore, the fractional majority of probes is not easily influenced by one or two outlying probe pairs. When determining differential gene expression, quantities that more equitably combine information from all of the probe pairs would seem preferable to quantities that may be dominated by a few probe pairs. As the strongest probes, both PM and MM, so often dominate the Average Difference, the quality of the Average Difference and the amount of information that we have lost are not entirely clear. Magnitude information can always be considered after differentially expressed genes have been selected.

Our proposal uses three approaches to determining differential gene expression with microarrays that have not previously been combined. First, in choosing which MAS 4.0 statistic to use, we chose the fraction of probe pairs changed, rather than the more commonly used ratio of Average Differences. As noted above, this statistic is closely related to the Affymetrix empirical Difference Call. While the changing probe pair summary measure is non-parametric in spirit, we should stress that we carry out a parametric test for this quantity based on a reference distribution. Second, the reference distribution was constructed by combining the data from several like-to-like comparisons on the same tissue. As

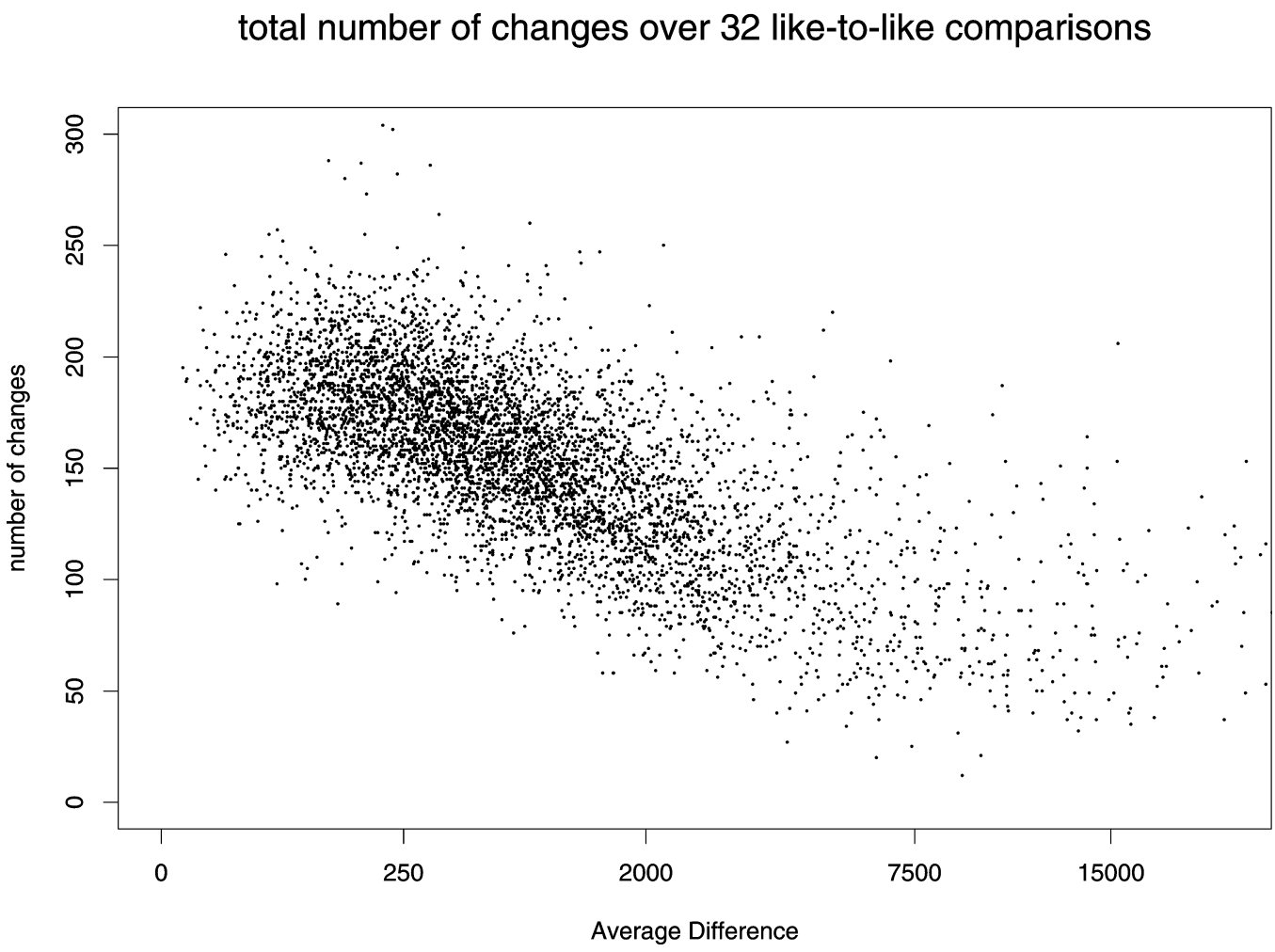

Figure 4. The number of probe pairs that change depends upon the signal strength. The data from the 32 striatum like-to-like comparisons were used to explore the relationship between the Average Difference (signal strength) and the number of probe pairs that change more than the Change Threshold. For all of the ~13 000 Probe Sets or genes, the sum of the number of increasing and decreasing probe pairs in the 32 comparisons is plotted against the Probe Set's mean Average Difference. There is a clear trend in the data. Probe Sets with lower signals tend to have more changing probe pairs. We interpret this to mean that signals with smaller Change Thresholds are more prone to random fluctuations causing a probe pair to be scored as Increased or Decreased.

there were many more such like-to-like comparisons than replicates in each individual experiment, this effectively increased the number of degrees of freedom. Third, by combining genes of comparable expression levels in a binning procedure, we generated more data points for the reference distribution. Having more data points allowed us to avoid assuming normality for our summary measure, an assumption that would have been impossible to verify, since such a verification should be carried out for each gene separately.

Another approach to assessing the significance of gene expression changes measured by Affymetrix arrays is proposed by Li and Wong (23,24). Their dChip software models the individual (PM − MM) differences using a multiplicative model with gene–sample and gene–probe-pair effects. This method also appears very attractive when a small number of replicates are available. As such, we suspect that the dChip approach is much more powerful than, for example, using *t*-statistics on the log(Average Differences). The dChip analysis, however, does not allow one to make use of a large number of like-to-like experiments such as we had, in order to establish a reference distribution. Conceivably, a hybrid of the probe-pair model of Li and Wong with our approach to building a reference distribution could yield an even more powerful approach to testing for differential gene expression.

While this paper was in preparation, Affymetrix released a new version of its analysis software, MAS 5.0. It reports for each gene a 'Change *P*-value' for differences observed in a comparison of two samples. These significance levels are based upon a Wilcoxon signed-rank test of the signal differences from the separate probe pairs (25). Thus, in its latest version of microarray analysis software, Affymetrix also uses information from the separate probe pairs to estimate the statistical significance of the differences between samples. Our choice of the probe pair increase and decrease count to determine significance was made independently of and well before we had knowledge of the new software. As the data on the individual

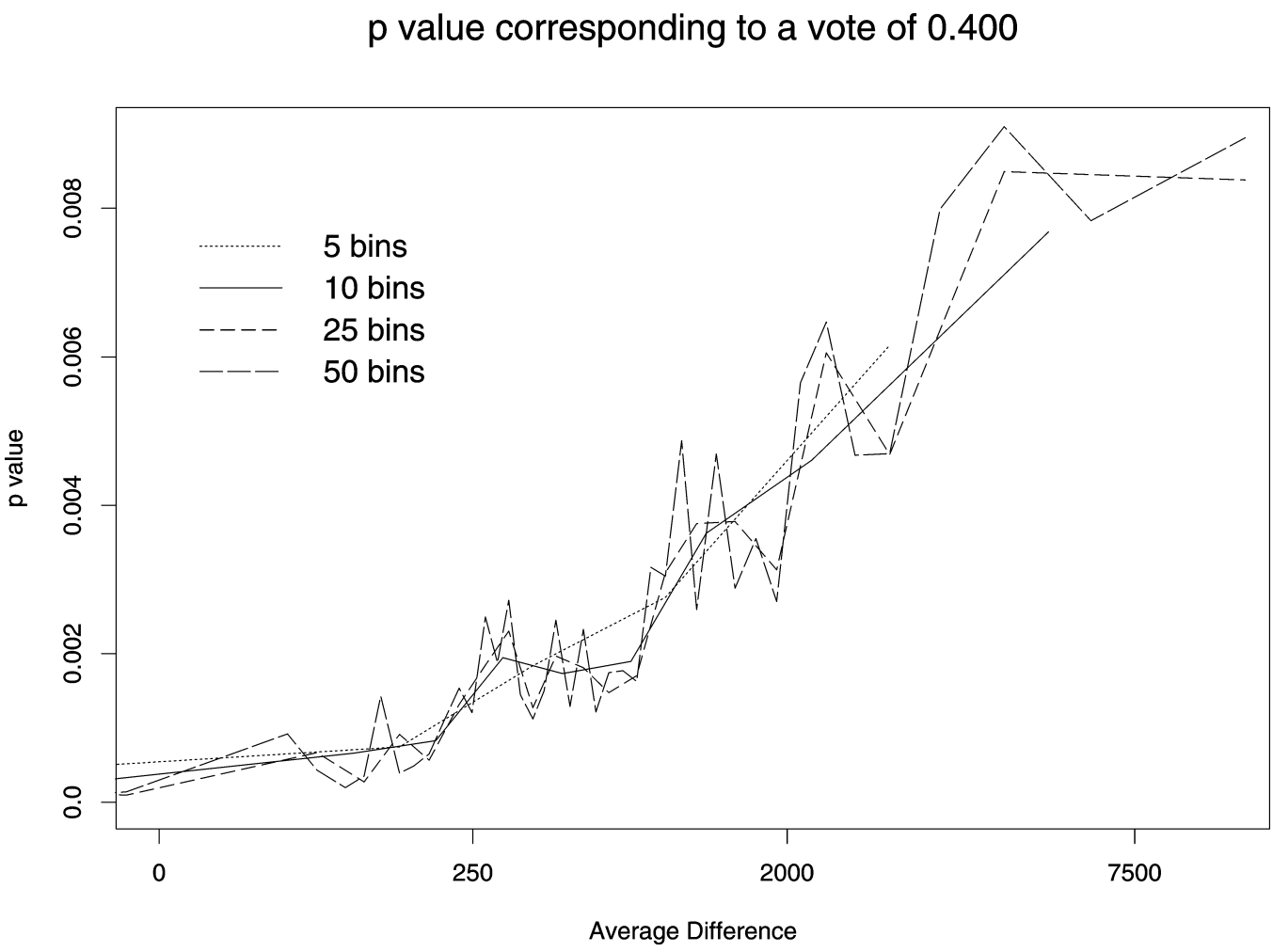

Figure 5. Plots of sampling-distribution-based *P*-values for an Increase Ratio − Decrease Ratio >0.40 with different numbers of bins. The Increase Ratio − Decrease Ratio values for each gene on the Mu11K A array from the 32 striatum like-to-like comparisons were alternately placed into $b = 5$, 10, 25 or 50 bins. Bin assignments were made after ranking the genes by their mean Average Difference on the 64 different arrays. Each bin contains $\sim 32 \times 6600/b$ realizations of (Increase Ratio − Decrease Ratio). For each of the bins, we compute the exact sampling distribution for the mean of N (in this case, $N = 2$). Based upon this sampling distribution, the calculated frequency, or estimated *P*-value, of values exceeding 0.40 is plotted for each bin.

probe pair's signals is not part of the normal MAS GeneChip 4.0 output, we rely on a 'vote' of anonymous probe pairs rather than standard non-parametric tests of ordinal values. Despite this and the extensive differences between the old and new Affymetrix software, our method gives results that are very consistent with those obtained with the new MAS GeneChip 5.0 software (data not shown). We are confident that complete reanalysis of the data using the new software would not fundamentally change our present view of the data if the new Change *P*-value or Difference Call metrics were selected as the standards for determining differential gene expression.

No version of the standard Affymetix analysis software yet accommodates experimental design involving replication. This was a primary motivation behind our error model. In addition, the MAS 5.0 Change *P*-values reported for single comparisons are also often rather extreme, i.e. $P < 0.00001$. This may be because the Wilcoxon test assumes independent measurements while the probe pairs are more accurately considered as repeated measurements of a single sample. Very small Change *P*-values are also caused by the Wilcoxon test being used on a vector of differences comprising PM − MM and a PM − Background correction, which causes the *P*-values to be more extreme owing to the repetition of values. Affymetrix finds through empirical testing that combining both quantities leads to slightly more accurate data at high and low target concentrations (E. Hubbell, Affymetrix, personal communication). Thus the Change *P*-values reported by MAS 5.0 probably should not be literally associated with a false-positive rate due to the non-standard dependence between values.

Issues of statistical significance versus biological significance can arise when statistical criteria are used to select genes. In addition, the ability to detect potential changes must be considered in light of the resolution of the secondary method. For these reasons, it may sometimes be advisable to apply other criteria such as interest and the apparent magnitude of the

Table 1. R_g thresholds associated with approximate *P*-values

Mu 11K A array Striatum $N=2$ *P*-value	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
0.05	0.225	0.200	0.200	0.225	0.225	0.225	0.225	0.250	0.225	0.250
0.01	0.300	0.275	0.275	0.300	0.325	0.325	0.325	0.350	0.350	0.400
0.005	0.325	0.300	0.325	0.325	0.350	0.350	0.375	0.400	0.400	0.450
0.001	0.400	0.350	0.400	0.400	0.450	0.450	0.450	0.475	0.500	0.550
0.0005	0.425	0.375	0.425	0.425	0.475	0.475	0.475	0.525	0.550	0.575
0.0001	0.475	0.425	0.500	0.525	0.550	0.550	0.550	0.600	0.625	0.675
0.00005	0.500	0.450	0.525	0.550	0.600	0.575	0.575	0.625	0.650	0.725
0.00001	0.550	0.500	0.600	0.600	0.650	0.625	0.650	0.700	0.750	0.825
Mu 11K A array Striatum $N=3$ *P*-value	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
0.05	0.167	0.167	0.167	0.167	0.183	0.183	0.183	0.200	0.200	0.200
0.01	0.233	0.217	0.233	0.233	0.250	0.250	0.250	0.267	0.283	0.300
0.005	0.250	0.233	0.250	0.267	0.283	0.283	0.283	0.300	0.317	0.333
0.001	0.300	0.283	0.300	0.317	0.350	0.333	0.350	0.367	0.383	0.400
0.0005	0.333	0.300	0.333	0.333	0.367	0.367	0.367	0.400	0.417	0.433
0.0001	0.367	0.333	0.383	0.400	0.433	0.417	0.417	0.467	0.483	0.517
0.00005	0.400	0.350	0.400	0.417	0.450	0.433	0.450	0.483	0.500	0.550
0.00001	0.433	0.383	0.450	0.467	0.500	0.483	0.500	0.550	0.567	0.617
Mu 11K A array Cerebellum $N=2$ *P*-value	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
0.05	0.200	0.200	0.200	0.200	0.225	0.225	0.225	0.225	0.200	0.225
0.01	0.275	0.250	0.275	0.300	0.300	0.300	0.325	0.325	0.325	0.350
0.005	0.300	0.275	0.300	0.325	0.350	0.325	0.375	0.375	0.375	0.400
0.001	0.375	0.350	0.375	0.400	0.425	0.400	0.475	0.475	0.450	0.500
0.0005	0.400	0.400	0.400	0.425	0.450	0.450	0.500	0.525	0.500	0.525
0.0001	0.450	0.500	0.450	0.500	0.525	0.500	0.575	0.600	0.575	0.625
0.00005	0.475	0.525	0.475	0.525	0.550	0.525	0.600	0.625	0.600	0.650
0.00001	0.525	0.575	0.525	0.575	0.625	0.575	0.675	0.700	0.675	0.775
Mu 11K A array Cerebellum $N=3$ *P*-value	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
0.05	0.150	0.150	0.167	0.167	0.167	0.183	0.183	0.183	0.167	0.167
0.01	0.217	0.200	0.217	0.233	0.233	0.233	0.250	0.267	0.250	0.267
0.005	0.233	0.233	0.250	0.250	0.267	0.267	0.283	0.300	0.283	0.300
0.001	0.283	0.283	0.300	0.317	0.333	0.317	0.350	0.367	0.350	0.367
0.0005	0.300	0.300	0.317	0.333	0.350	0.333	0.383	0.383	0.367	0.400
0.0001	0.350	0.367	0.350	0.383	0.400	0.383	0.433	0.450	0.433	0.467
0.00005	0.367	0.383	0.383	0.400	0.433	0.417	0.467	0.467	0.450	0.500
0.00001	0.417	0.433	0.417	0.450	0.483	0.450	0.517	0.533	0.517	0.567
Mu 11K A array Cortex $N=2$ *P*-value	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
0.05	0.200	0.175	0.175	0.200	0.200	0.200	0.200	0.200	0.200	0.175
0.01	0.250	0.250	0.225	0.250	0.250	0.275	0.275	0.275	0.275	0.275
0.005	0.275	0.275	0.250	0.275	0.275	0.300	0.325	0.325	0.300	0.325
0.001	0.350	0.300	0.300	0.325	0.350	0.350	0.400	0.425	0.400	0.425
0.0005	0.350	0.325	0.325	0.350	0.375	0.375	0.425	0.450	0.450	0.450
0.0001	0.400	0.375	0.375	0.400	0.425	0.425	0.475	0.525	0.525	0.500
0.00005	0.425	0.400	0.375	0.400	0.450	0.450	0.500	0.525	0.550	0.525
0.00001	0.475	0.425	0.400	0.450	0.500	0.500	0.550	0.600	0.600	0.600
Mu 11K A array Cortex $N=3$ *P*-value	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
0.05	0.150	0.150	0.150	0.150	0.150	0.167	0.167	0.167	0.167	0.133
0.01	0.200	0.200	0.183	0.200	0.200	0.217	0.233	0.233	0.217	0.217
0.005	0.233	0.217	0.200	0.217	0.233	0.233	0.250	0.267	0.250	0.250
0.001	0.267	0.250	0.250	0.250	0.283	0.283	0.300	0.317	0.317	0.300
0.0005	0.283	0.267	0.250	0.267	0.300	0.300	0.333	0.350	0.333	0.333
0.0001	0.333	0.300	0.283	0.317	0.333	0.333	0.367	0.383	0.400	0.383
0.00005	0.350	0.317	0.300	0.317	0.350	0.350	0.400	0.417	0.417	0.400
0.00001	0.367	0.333	0.333	0.350	0.400	0.400	0.433	0.450	0.467	0.467

From each array and tissue type, sampling distributions of the Increase Ratio − Decrease Ratio for each decile were assembled for $N=2,\ldots,6$. The threshold values in each bin corresponding to a frequency P were then tabulated. The results for $N=2$ and $N=3$ are shown. Complete tables can be found in the Supplemental Data (www.neumetrix.info).

Table 1. continued

Mu 11K B array Striatum $N=2$ *P*-value	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
0.05	0.175	0.200	0.200	0.200	0.200	0.200	0.225	0.225	0.250	0.275
0.01	0.250	0.250	0.250	0.275	0.275	0.300	0.325	0.350	0.375	0.425
0.005	0.275	0.275	0.300	0.300	0.300	0.325	0.350	0.400	0.425	0.475
0.001	0.350	0.350	0.350	0.375	0.375	0.425	0.450	0.500	0.525	0.575
0.0005	0.375	0.375	0.400	0.400	0.400	0.450	0.475	0.525	0.550	0.625
0.0001	0.450	0.475	0.475	0.475	0.475	0.500	0.550	0.600	0.625	0.725
0.00005	0.475	0.500	0.500	0.500	0.500	0.525	0.600	0.650	0.675	0.800
0.00001	0.525	0.550	0.550	0.575	0.550	0.600	0.675	0.750	0.775	0.900
Mu 11K B array Striatum $N=3$ *P*-value	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
0.05	0.150	0.150	0.150	0.150	0.167	0.167	0.183	0.183	0.200	0.217
0.01	0.200	0.200	0.200	0.217	0.217	0.233	0.250	0.267	0.283	0.317
0.005	0.217	0.217	0.233	0.233	0.250	0.267	0.283	0.300	0.317	0.367
0.001	0.267	0.267	0.283	0.283	0.300	0.317	0.350	0.383	0.400	0.433
0.0005	0.300	0.300	0.300	0.317	0.317	0.333	0.367	0.400	0.417	0.483
0.0001	0.350	0.350	0.350	0.367	0.367	0.400	0.433	0.467	0.500	0.567
0.00005	0.367	0.383	0.367	0.383	0.383	0.417	0.450	0.500	0.517	0.600
0.00001	0.400	0.417	0.417	0.433	0.433	0.467	0.517	0.567	0.600	0.667
Mu 11K B array Cerebellum $N=2$ *P*-value	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
0.05	0.200	0.200	0.200	0.200	0.200	0.200	0.225	0.225	0.225	0.225
0.01	0.250	0.275	0.250	0.275	0.275	0.275	0.300	0.325	0.350	0.375
0.005	0.300	0.300	0.300	0.300	0.300	0.325	0.325	0.375	0.400	0.425
0.001	0.350	0.350	0.350	0.350	0.375	0.400	0.400	0.475	0.500	0.525
0.0005	0.375	0.375	0.375	0.375	0.375	0.425	0.450	0.500	0.550	0.575
0.0001	0.425	0.425	0.425	0.425	0.425	0.475	0.500	0.575	0.625	0.650
0.00005	0.450	0.450	0.450	0.450	0.450	0.500	0.525	0.600	0.650	0.700
0.00001	0.500	0.500	0.500	0.500	0.500	0.575	0.600	0.675	0.750	0.825
Mu 11K B array Cerebellum $N=3$ *P*-value	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
0.05	0.150	0.167	0.150	0.167	0.167	0.167	0.167	0.183	0.183	0.183
0.01	0.200	0.217	0.217	0.217	0.217	0.217	0.233	0.267	0.267	0.283
0.005	0.233	0.233	0.233	0.233	0.250	0.250	0.267	0.283	0.300	0.333
0.001	0.267	0.283	0.267	0.267	0.283	0.300	0.317	0.350	0.383	0.400
0.0005	0.300	0.300	0.300	0.300	0.300	0.317	0.333	0.383	0.400	0.433
0.0001	0.333	0.333	0.333	0.333	0.350	0.367	0.383	0.433	0.467	0.500
0.00005	0.350	0.350	0.350	0.350	0.367	0.383	0.417	0.467	0.500	0.550
0.00001	0.383	0.400	0.400	0.383	0.400	0.433	0.450	0.517	0.567	0.617
Mu 11K B array Cortex $N=2$ *P*-value	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
0.05	0.200	0.200	0.200	0.200	0.200	0.200	0.225	0.200	0.200	0.175
0.01	0.300	0.300	0.300	0.275	0.275	0.300	0.300	0.275	0.300	0.300
0.005	0.350	0.350	0.325	0.325	0.325	0.325	0.350	0.300	0.325	0.375
0.001	0.450	0.425	0.425	0.400	0.400	0.425	0.425	0.375	0.400	0.475
0.0005	0.500	0.475	0.475	0.425	0.425	0.450	0.450	0.400	0.425	0.500
0.0001	0.575	0.525	0.550	0.475	0.475	0.525	0.525	0.450	0.500	0.575
0.00005	0.600	0.575	0.575	0.500	0.500	0.550	0.550	0.475	0.525	0.600
0.00001	0.675	0.650	0.650	0.575	0.575	0.625	0.625	0.500	0.575	0.725
Mu 11K B array Cortex $N=3$ *P*-value	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
0.05	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.150
0.01	0.233	0.233	0.233	0.217	0.217	0.233	0.233	0.217	0.233	0.233
0.005	0.267	0.267	0.267	0.250	0.250	0.250	0.267	0.233	0.250	0.283
0.001	0.350	0.333	0.333	0.300	0.300	0.317	0.333	0.283	0.317	0.350
0.0005	0.367	0.350	0.367	0.317	0.333	0.350	0.350	0.300	0.333	0.367
0.0001	0.433	0.417	0.417	0.367	0.367	0.400	0.400	0.350	0.383	0.433
0.00005	0.467	0.433	0.450	0.400	0.400	0.417	0.433	0.367	0.400	0.467
0.00001	0.517	0.500	0.500	0.433	0.433	0.467	0.483	0.400	0.450	0.550

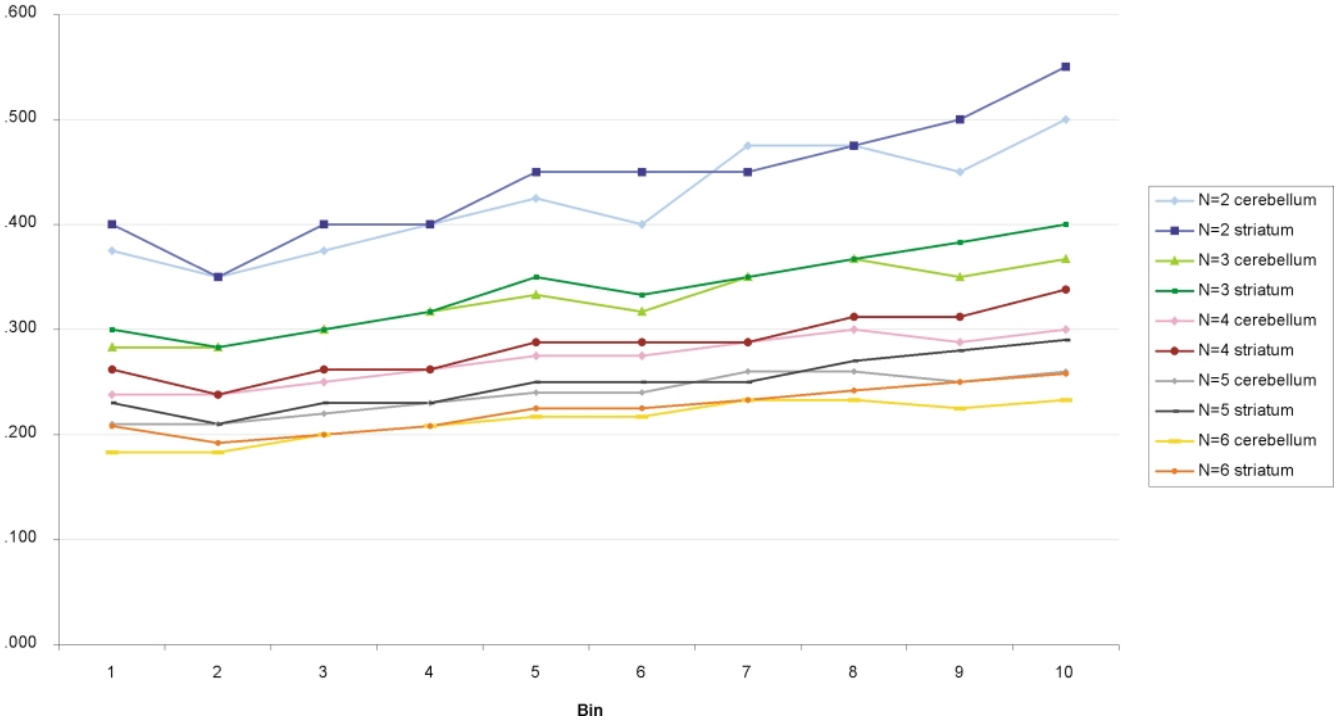

Figure 6. R_g cutoffs associated with $P < 0.001$ for the striatum and cerebellum Mu11K A array models and different N. The Mu11K A array R_g thresholds associated with $P < 0.001$ for each of the 10 bins are plotted for the cerebellum and striatum error models. The lines for each model are quite close together for $N = 2, \ldots, 6$. The increased power conferred by replication is seen in the diminishing thresholds as N increases.

change to guide the choice of genes to be examined in confirmatory studies.

Our method may not be the best means of assessing significance or identifying interesting genes when the experimental question of interest is not a simple comparison of two populations. A time series is one example of a more complex experimental design. Larger sample sizes also make this method unwieldy, but as sample size increases, standard *t*-statistics can be applied. The advantages of probe pair level analysis like ours are greatest in the event that there are few replicates. But even in other instances, information about the individual probe pairs might allow the development and application of more powerful statistical tests than tests based on single-value averages such as the Average Difference.

Because genes were binned in the error models to get stable *P*-value estimates, each array essentially turned into a generic array of 10 'genes' independently of biological identity and experimental detail. Thresholds derived for the striatum, for instance, can reasonably be used on the cortex data as a more conservative estimator of *P*-values. This leads to the question of how general are the thresholds we determined. The issue here is relating the technical variability of one experiment to that of another. If the arrays and samples are more variable than the data used to construct the tables, then too many false positives would be labeled significant; alternatively, if the samples were less variable, then the table values would be conservative. Fine-tuning thresholds for individual studies or extrapolation of our numbers to other studies might be possible if one could relate the variability of one experiment to another. As a practical matter, merely repeating the *P*-value assignments using tables for larger or smaller N approximates fine-tuning, since this raises or lowers the *P*-value cutoffs.

METHODS

Calculation of error model based *P*-values

To simplify the calculations, we tabulate frequencies for probe pair counts, compute the expected frequencies for various N, where N is the number of replicates, and then normalize the counts by dividing by N times the predominant number of probes per gene. Let $f_1(x)$ be the fraction of times that we observe an (Increase − Decrease) score of x in a bin in the replicate comparisons. On the Mu11K mouse arrays, 20 probe pairs typically represent genes, and for convenience we ignore the few probe sets with other numbers when performing the calculations. Thus x ranges from −20 to 20.

To calculate frequencies for $N \geq 2$, set

$$f_i(x) = \sum_{y=-20}^{20} f_1(y) f_{i-1}(x-y) \quad \text{for } i = 2, \ldots, N \text{ and}$$

$$x = -(i \times 20), \ldots, (i \times 20),$$

Table 2. The number of genes with an approximate *P*-value < 0.001 in the different brain tissues and models and the results of confirmation studies

Model		Striatum				Cerebellum			Cortex		
R6/2	Number of replicates	$N=2$	$N=2$	$N=2$	$N=2$	$N=2$			$N=3$	$N=3$	$N=3$
	Age (weeks)	2	4	6	12	12			2	6	12
	$P<0.001$	7	1	29	147	183			25	62	182
	1 − FDR	0.00	0.00	0.55	0.91	0.93			0.48	0.81	0.93
	Confirmation	0 of 2	Not done	3 of 3	9 of 10	13 of 16			0 of 2	3 of 3	12 of 12
DRPLA	Number of replicates	$N=2$	$N=2$	$N=2$		$N=4$	$N=4$	$N=4$			
	Comparison	Q26 vs WT	Q65 vs WT	Q65 vs Q26		Q26 vs WT	Q65 vs WT	Q65 vs Q26			
	$P<0.001$	24	36	29		100	469	448			
	1 − FDR	0.46	0.64	0.55		0.87	0.97	0.97			
	Confirmation	Not done	Not done	Not done		Not done	9 of 9	Not done			
N171	Number of replicates	$N=2$	$N=2$	$N=2$		$N=4$	$N=4$	$N=4$			
	Comparison	18Q vs WT	82Q vs WT	82Q vs 18Q		18Q vs WT	82Q vs WT	82Q vs 18Q			
	$P<0.001$	15	13	42		73	165	224			
	1 − FDR	0.13	0.00	0.69		0.82	0.92	0.94			
	Confirmation	Not done	Not done	Not done		Not done	7 of 7	Not done			
YAC	Number of replicates	$N=2$				$N=5$					
	Expression level and age	Low, 12 months				Low, 12 months					
	$P<0.001$	16				84					
	1 − FDR	0.19				0.85					
	Confirmation	2 of 3				1 of 3					
Aronin	Number of replicates	$N=6$	$N=6$								
	Comparison	Mild vs WT	Severe vs WT								
	$P<0.001$	174	77								
	1 − FDR	0.93	0.83								
	Confirmation	0 of 3	0 of 4								

The number of genes assigned an approximate *P*-value <0.001 from each mouse model is shown. We score the confirmation of the genes labeled at this significance level. More complete descriptions of confirmation of microarray data are given in (1–4). Random chance would have ~ 13 genes at this level on the two Mu11K arrays. The False-Discovery Rate (FDR) is equal to 13 divided by the number of observed genes. The comparisons for the R6/2 and YAC mice are all with age- and sex-matched control or wild-type (WT) mice. The cortex numbers reflect the application of the striatum error model thresholds to the cortex data. This was done because the number of striatal specimens was more appropriate for generating an error model, and that model has more conservative thresholds (Fig. 6). Application of the more lenient cortex model results in 62, 85 and 238 genes being labeled at 2, 6 and 12 weeks. There is a trend in the R6/2 striatal and cortex profiles for more gene expression changes as the mice age. Another general trend is the expected finding that more genes are labeled as significant as the number of replicates increases, this is shown most clearly in the DRPLA and N171 models where the striatum experiments were performed with $N=2$ and the cerebellum experiments with $N=4$.

with $f_{i-1}(x - y) = 0$ if $|x - y| > (i \times 20)$. The quantity $f_i(x)$ is the fraction of times that the sum of N independent counts is x. Set

$$g_i(x) = \sum_{y \le x} f_i(y) \quad \text{for } i = 1, \ldots, N \text{ and}$$
$$x = -(i \times 20), \ldots, (i \times 20).$$

The quantity $g_i(x)$ is the fraction of times the sum of N independent counts is smaller than x. Computing $g_i(x)/(N \times 20)$ provides P-values.

Assigning *P*-values to experimental differences

The algorithm can be summarized as follows.

1. For each gene g, let D_g be the mean Average Difference over all arrays, and assign a bin, or decile, to each gene based on its mean signal. Our convention is that bin 1 contains the lowest-signal genes and bin 10 the highest.
2. For each gene g, let R_g be the average over all experiment-to-baseline comparisons of the Increase Ratio minus the Decrease Ratio.
3. There are separate tables for each tissue, type of array and N. To assign a P-value for R_g, select the appropriate table and bin for gene g. Cutoff values for R_g are found in this column. Approximate P-values are found by reading across to the leftmost column. To illustrate, if a striatum gene in bin A10 $N = 2$ on the Mu11K A array has $R_g = 0.51$, then a P-value range of $0.001 < P \le 0.005$ is assigned. A computer script to map the P-value estimates to the data is in the Supplemental Data (www.neumetrix.info).

If the number of experiment and baseline samples is different, then using the smaller of the two numbers as N will assign more conservative P-values. When determining what P-value to use as a cutoff for selecting interesting genes, one should keep in mind that approximately a fraction P of all genes would be indicated as significant at a particular level. For example, if 0.05 were chosen as the P-value cutoff, then 0.05×6600 or 330 genes would be expected to appear significant by chance alone on each Mu11K array. If at this significance level 500 genes are observed, then a false-discovery rate of $330/500 = 0.66$ would be expected. In general, it is recommended that one use a more conservative P-value than 0.05. To be very conservative, a multiple comparison correction, such as the Bonferoni or Westfall–Young (26) can be applied. A less conservative P-value can be used when selecting a group of genes for further analysis. The false-discovery rate (13) as described above can now guide the choice of P-value.

When applying the algorithm to real data for small N, R_g is typically calculated by averaging over all possible combinations of experiment (E)-to-baseline (B) comparisons. The rationale behind this is to minimize the effects of experimental bias due to a single outlying array or sample and not discard data in cases where the experiment and control samples have different numbers of replicates. Formally, the reference distribution is valid for the average of E1 − B1 and E2 − B2. However, the correlation between this average and the average of E1 − B2 and E2 − B1 is usually very high (∼0.9). If anything, inclusion of the additional comparisons leads to P-values that are slightly on the conservative side. For these analyses, if $N \le 4$, we performed all the possible comparisons and averaged. If $N \ge 5$, we selected and compared arbitrary pairs of samples and averaged. The exceptions were with sets of replicates that were generated at different times, since it was observed that technical variance began to contribute more than desired to the observed differences. In those situations, only comparisons between experiments and baselines generated in parallel were used in the computations.

Sample preparation and image analysis

Sample preparation and array processing were done per manufacturer's specification (9). Prior to analysis, normalization was performed by global scaling, setting the target intensity of each array to 1000 arbitrary intensity units, with all other parameters at the default levels.

ACKNOWLEDGEMENTS

We thank Mark Aronszajn for writing computer script, Cassie Neal and Jeff Delrow of the FHCRC array facility for expert technical assistance and helpful discussions, Third Millennium for database development, and the Hereditary Disease Array Group. These studies were funded by the Hereditary Disease Foundation Cure HD Initiative (J.M.O.) and the National Institutes of Health (NS42157 to J.M.O. and CA74841 to C.K.).

REFERENCES

1. Chan, E.Y.W., Luthi-Carter, R., Strand, A.D., Solano, S.M., Hanson, S.A., DeJohn, M.M., Kooperberg, C., Chase, K.O., DiFiglia, M., Young, A.B. *et al.* (2002) Increased huntington protein length reduces the severity of polyglutamine-induced gene expression changes in mouse models of Huntington's disease. *Hum. Mol. Gen.*, **11**, 1939–1951.
2. Luthi-Carter, R., Hanson, S.A., Strand, A.D., Bergstrom, D.A., Chun, W., Peters, N.L., Woods, A.M., Chan, E.Y.W., Kooperberg, C., Young, A.B. *et al.* (2002) Dysregulation of gene expression in the R6/2 model of polyglutamine disease: parallel changes in muscle and brain. *Hum. Mol. Gen.*, **11**, 1911–1926.
3. Luthi-Carter, R., Strand, A.D., Hanson, S.A., Kooperberg, C., Schilling, G., La Spada, A.R., Merry, D.E., Young, A.B., Ross, C.A., Borchelt, D.R. *et al.* (2002) Polyglutamine and transcription: gene expression changes shared by DRPLA and Huntington's disease mouse models reveal context-independent effects. *Hum. Mol. Gen.*, **11**, 1927–1937.
4. Sipione, S., Rigamonti, D., Valenza, M., Zucato, C., Pritchard, J.I., Kooperberg, C., Olson, J.M. and Cattaneo, E. (2002) Early transcriptional profiles in huntington-inducible striatal cells by microchip analysis. *Hum. Mol. Gen.*, **11**, 1953–1965.
5. DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A. and Trent, J.M., (1996) Use of a cDNA microarray to analyze gene expression patterns in human cancer. *Nat. Genet.*, **14**, 457–460.
6. Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C., Trent, J.M., Staudt, L.M., Hudson, J., Jr, Boguski, M.S. *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.
7. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Che, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
8. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. and Lockhart, D.J. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.*, **21**(Suppl.), 20–24.

9. *Affymetrix Microarray Suite User Guide Version 4.0* (2000) Affymetrix, Santa Clara, CA.
10. Luthi-Carter,R., Strand, A., Peters, N.L., Solano, S.M., Hollingsworth, Z.R., Menon, A.S., Frey, A.S., Spektor, B.S., Penney, E.B., Schilling, G. *et al.* (2000) Decreased expression of striatal signaling genes in a mouse model of Huntington's disease. *Hum. Mol. Gen.*, **9**, 1259–1271.
11. Olson, J.M., Asakura, A., Snider, L., Hawkes, R., Strand, A., Stoeck, J., Hallahan, A., Pritchard, J. and Tapscott, S.J. (2001) NeuroD2 is necessary for development and survival of central nervous system neurons. *Dev. Biol.*, **234**, 174–187.
12. Porter, J.D., Khanna, S., Kaminski, H.J., Rao, J.S., Merriam, A.P., Richmonds, C.R., Leahy, P., Li, J., Guo, W. and Andrade, F.H. (2002) A chronic inflammatory response dominates the skeletal muscle molecular signature in dystrophin-deficient mdx mice. *Hum. Mol. Genet.*, **11**, 263–272.
13. Storey, J.D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc. B*, **10**, 479–498.
14. Lönnstedt, I. and Speed, T.P. (2001) Replicated microarray data. *Statist. Sinica*, **12**, 31–46.
15. Tushner, V., Tibshirani, R.J. and Chu, C. (2001) Significance analysis of microarrays applied to ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
16. Baldi, P. and Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.
17. Kooperberg, C., Sipione, S., LeBlanc, M.L., Strand, A.D., Cattaneo, E. and Olson, J.M. (2002) Evaluating test statistics to select interesting genes in microarray experiments. *Hum. Mol. Gen.*, **11**, 2223–2232.
18. Novak, J.P., Sladek, R. and Hudson, T.J. (2002) Characterization of variability in large-scale gene expression data: implications for study design. *Genomics*, **79**, 104–113.
19. Yang, Y.H., Dudoit, S., Luu, P. and Speed, T.P. (2001) Normalization for cDNA microarray data. In Bittner, M.L., Chen, Y., Dorsel, A.N. and Dougherty, E.R. (eds), *Microarrays: Optical Technologies and Informatics.* Proceedings of SPIE, San Jose, CA, vol. 4266, p. 31.
20. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
21. Pritchard, C.C., Hsu, L., Delrow, J., Nelson, P.S. (2001) Project Normal: defining normal variance in mouse gene expression. *Proc. Natl. Acad. Sci. USA*, **98**, 13266–13271.
22. Fan, J. and Gijbels, I. (1996) *Local Polynomial Modeling and its Applications.* Chapman & Hall, London.
23. Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
24. Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol. 2001*, **2**, RESEARCH0032.
25. *Affymetrix Microarray Suite User Guide Version 5.0* (2001) Affymetrix, Santa Clara, CA.
26. Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica*, **12**, 111–139.