BIOINFORMATICS

Vol. 20 no. 16 2004, pages 2534–2544
doi:10.1093/bioinformatics/bth280

Interactively optimizing signal-to-noise ratios in expression profiling: project-specific algorithm selection and detection p-value weighting in Affymetrix microarrays

Jinwook Seo[1,2], Marina Bakay[1], Yi-Wen Chen[1], Sara Hilmer[1], Ben Shneiderman[2] and Eric P Hoffman[1,]*

[1]*Research Center for Genetic Medicine, Children's National Medical Center and* [2]*Human-Computer Interaction Lab and Department of Computer Science, University of Maryland, College Park, MD 20742 USA*

Received on March 23, 2004; revised and accepted on April 16, 2004
Advance Access publication April 29, 2004

ABSTRACT

Motivation: The most commonly utilized microarrays for mRNA profiling (Affymetrix) include 'probe sets' of a series of perfect match and mismatch probes (typically 22 oligonucleotides per probe set). There are an increasing number of reported 'probe set algorithms' that differ in their interpretation of a probe set to derive a single normalized 'signal' representative of expression of each mRNA. These algorithms are known to differ in accuracy and sensitivity, and optimization has been done using a small set of standardized control microarray data. We hypothesized that different mRNA profiling projects have varying sources and degrees of confounding noise, and that these should alter the choice of a specific probe set algorithm. Also, we hypothesized that use of the Microarray Suite (MAS) 5.0 probe set detection *p*-value as a weighting function would improve the performance of all probe set algorithms.

Results: We built an interactive visual analysis software tool (HCE2W) to test and define parameters in Affymetrix analyses that optimize the ratio of signal (desired biological variable) versus noise (confounding uncontrolled variables). Five probe set algorithms were studied with and without statistical weighting of probe sets using the MAS 5.0 probe set detection *p*-values. The signal-to-noise ratio optimization method was tested in two large novel microarray datasets with different levels of confounding noise, a 105 sample U133A human muscle biopsy dataset (11 groups: mutation-defined, extensive noise), and a 40 sample U74A inbred mouse lung dataset (8 groups: little noise). Performance was measured by the ability of the specific probe set algorithm, with and without detection *p*-value weighting, to cluster samples into the appropriate biological groups (unsupervised agglomerative clustering with *F*-measure values). Of the total random sampling analyses, 50% showed a highly statistically significant difference between probe set algorithms by ANOVA [$F(4,10) > 14$, $p < 0.0001$], with weighting by MAS 5.0 detection *p*-value showing significance in the mouse data by ANOVA [$F(1,10) > 9$, $p < 0.013$] and paired *t*-test [$t(9) = -3.675$, $p = 0.005$]. Probe set detection *p*-value weighting had the greatest positive effect on performance of dChip difference model, ProbeProfiler and RMA algorithms. Importantly, probe set algorithms did indeed perform differently depending on the specific project, most probably due to the degree of confounding noise. Our data indicate that significantly improved data analysis of mRNA profile projects can be achieved by optimizing the choice of probe set algorithm with the noise levels intrinsic to a project, with dChip difference model with MAS 5.0 detection *p*-value continuous weighting showing the best overall performance in both projects. Furthermore, both existing and newly developed probe set algorithms should incorporate a detection *p*-value weighting to improve performance.

Availability: The Hierarchical Clustering Explorer 2.0 is available at http://www.cs.umd.edu/hcil/hce/. Murine arrays (40 samples) are publicly available at the PEPR resource (http://microarray.cnmcresearch.org/pgadatatable.asp; http://pepr.cnmcresearch.org; Chen *et al*., 2004).

Contact: ehoffman@cnmcresearch.org

INTRODUCTION

Simultaneous analysis of many thousands of genes on the microarray leads to an 'expression profile' of the original cell or tissue. This profile represents the subset of the 40 000 genes that are being employed by that cell or tissue, at that particular point of time. High density oligonucleotide arrays containing up to 500 000 features are used widely for many projects in biological and medical research. The most popular

*To whom correspondence should be addressed.

Affymetrix GeneChip uses about 1 million oligonucleotide probes to query most (~40 000) human mRNAs in two small (1.28 cm^2) glass arrays. Importantly, Affymetrix arrays have intrinsic redundancy of measurements for each gene, with 11–16 'perfect match' probes for different regions of each gene sequence, with each perfect match paired with a similar 'mismatch' probe with a single destabilizing nucleotide change in the center of the 25 nucleotide sequence (Liu *et al*., 2002; Hubbell *et al*., 2002). The complete set of 16 probe pairs is called the 'probe set' for any single gene. The mismatch is meant to serve as a 'noise filter'; labeled mRNA binding to the 'mismatch' is considered to represent a measure of non-specific binding, and thus a measure of 'noise' for the corresponding perfect match (see The Tumor Analysis Best Practices Working Group, 2004).

There are many confounding uncontrolled variables intrinsic to most microarray projects. For example in human patient samples, the outbred nature of humans leads to extensive genetic heterogeneity between individuals, even if sharing the same pathological condition or exposed to the same environmental or drug challenge. It is often difficult to precisely match age, sex and ethnic background of human subjects in microarray projects, leading to considerable inter-individual variability in the analyses. Furthermore, human tissue samples typically show extensive tissue heterogeneity, with small size leading to sampling error, and variability in histological severity and cell content (e.g. variable amounts of fibrosis, fatty infiltration, inflammation, regeneration). Many of these variables are not a concern in studies of inbred mouse strains. Inbred mice show very little inter-individual variability, and the experimental manipulation of groups of mice leads to homogeneous treatment groups often with relatively high numbers of replicates. Moreover, the use of whole lungs or other tissues leads to a normalization of tissue heterogeneity.

There are also technical variables that could confound interpretation, quality and preservation of the biopsy material, quality of RNA, cDNA and cRNA, hybridization and chip image variation, probe set signal algorithms and statistical analysis methods. Quality Control (QC) and Standard Operating Procedure (SOP) can mitigate many confounding technical variables with factory-produced Affymetrix arrays, and these have been found to be a relatively minor source of confounding variation if QC parameters are employed (Bakay *et al*., 2002a; DiGiovanni *et al*., 2003).

'Probe set algorithms' refer to the method of interpreting the 11–16 probe pairs (22–32 oligonucleotide probes) in a probe set on an Affymetrix microarray that query a particular mRNA transcript. Key variables in different probe set algorithms include the penalty weight given to the mismatch probe of each probe pair, the weighting of specific probes in a probe set based on empirical 'performance', the manner by which a single 'signal value' is derived from the interpretation of the probe set, and how this is normalized relative to other probe sets on the microarray or in the entire project. Most reports of new probe set algorithms, and comparison of existing algorithms, have been performance using one or a few 'test datasets' in the public domain; specifically 'spike in' control datasets from Affymetrix (http://www.affymetrix.com/analysis/download_center2.affx) and GeneLogic (http://qolotus02.genelogic.com/datasets.nsf/) (Li and Wong, 2001b; Irizarry *et al*., 2003b; Bolstad *et al*., 2003). These data have shown that using only the perfect match probe, and ignoring the mismatch probe of each probe pair can considerably increase the sensitivity of the study, particularly at low signal levels (Irizarry *et al*., 2003a). The performance of different probe set algorithms and normalization methods is typically done using receiver operating characteristic (ROC) curves, providing an assessment of signal-to-noise ratio for the spike-in control mRNAs.

As discussed above, different projects are known to have different levels of confounding noise. We hypothesized that the increased sensitivity of probe set algorithms that ignore the mismatch signal, such as robust multi-array average (RMA) (Irizarry *et al*., 2003b), would be expected to come at an increased cost of noise, where the quality of low level signals defined by RMA in 'noisy' projects would lead to data interpretations of poor integrity. Specifically, detection of spike-in controls would be expected to be independent of confounding noise within arrays and projects. However, the increased sensitivity of some probe set algorithms would be expected to lead to a high proportion of false positives in projects where there was relatively high level of unwanted noise. We hypothesized that different probe set algorithms would show a 'project-specific' performance, based upon the extent of confounding noise in a particular project.

The optimization of signal-to-noise ratio is a critical issue in microarray experiments, where tens of thousands of transcripts are analyzed simultaneously. If a highly sensitive probe set algorithm is used in a noisy project, then the resulting data will have very poor quality and specificity, with many thousands of 'false positives'. This would lead to both misclassification of samples, and very noisy results that could absorb large amounts of experimental time to parse through. Even though such noises and noise filtering methods strongly influence data analysis, signal-to-noise ratios are rarely optimized, or even considered in microarray data analyses. This is partly because of the lack of analysis tools that allow researchers to interactively test and verify various combinations of parameters for noise analysis.

Another aspect of microarray data interpretation that could alter results is the 'weighting' of specific probe sets. Typically, once a particular probe set algorithm is employed on a microarray project, each probe set signal is considered as equal weight with any other probe set signal. However, probe sets that detect transcripts expressed at a very high level would be expected to show a 'more robust' signal with greater quality, compared to probe sets that are performing poorly or detecting very low level transcripts (near background). A measure

of the confidence of the performance of the probe set is a continuous 'detection p-value' assignment, which is a function of the signal difference between the perfect match (PM) and mismatch (MM) probes in a probe set and the signal intensity. In Affymetrix MAS 5.0, the Discrimination score, $R = (\mathrm{PM} - \mathrm{MM})/(\mathrm{PM} + \mathrm{MM})$, is calculated for each probe pair, and the one-sided Wilcoxon's signed rank test against a small positive number (default = 0.015) is performed to generate the detection p-value (Affymetrix, 2001a,b,c http://www.affymetrix.com/products/software/specific/mas.affx, https://www.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf). Two threshold values and α_2 are assigned where poor detection p-values (less than α_1) are assigned an 'absent' call, while more robust detection p-values (greater than α_2) are assigned a 'present' call (default $\alpha_1 = 0.04$ and $\alpha_2 = 0.06$). It is now standard practice in many publications using Affymetrix arrays to use the 'present/absent' calls as a form of noise filters. For example, a '10% present call' noise filter requires any specific probe set to show a 'present' call in at least 1 in 10 microarrays in that project, otherwise it is excluded from all further analyses (DiGiovanni *et al*., 2003, 2004; Zhao *et al*., 2002, 2003). Use of a threshold is not as statistically valid as a continuous weighting method, and here we tested the effect of weighting of all probe set algorithms by MAS 5.0 detection p-values.

We hypothesized that it would be possible to identify the most appropriate probe set analysis and noise filtering methods by conducting permutational analysis of the probe set 'signal' algorithm, and noise filters using continuous MAS 5.0 probe set detection p-values. The goal was to use unsupervised hierarchical clustering to find the signal algorithm that maximized the separation of the 'known' biological variable, while minimizing confounding 'noise'. We enhanced our interactive visual analysis tool, the Hierarchical Clustering Explorer to enable researchers to perform the permutational study and to help them interactively evaluate the result. We report the analysis results of such permutational studies with very noisy human muscle biopsies samples and much cleaner inbred mouse lung biopsies samples.

In our previous work (Seo *et al*., 2003, http://ieeexplore.ieee.org/ie15/8655/27434/01221348.pdf), we performed a pilot permutational study with a small subset (25 samples of 3 groups) of our 105 human muscle biopsies. We varied probe set signal algorithms (MAS 5.0, RMA), 'present call' filter thresholds, and clustering linkage methods, and 'visually' investigated the results in HCE2 (the Hierarchical Clustering Explorer 2.0). For the dataset, the strength of the biological variable was maximized, and noise minimized, using MAS 5.0, 10% present call filter, and average linkage (or average group linkage). In this paper, we extend the pilot study to the extent that (1) we test not only the human muscle data with extensive noise but also the inbred mouse lung data expected to show considerably less biological noise [varying genetic background (polymorphisms), tissue heterogeneity], (2) compare three more signal algorithms (dChip, dChip difference model, Probe Profiler), (3) use a novel continuous noise filtering method instead of the binary 10% 'present call' filtering used previously and (4) evaluate the unsupervised clustering results not only using visual inspection but also using a general external evaluation measure (F-measure).

We first explain our permutation study design and datasets in detail. Then, our novel noise filtering method incorporated into the unsupervised hierarchical clustering algorithm is presented. An external clustering evaluation measure—F-measure is explained and application of the measure to a hierarchical clustering result is explained in the following section. Then, we talk about how those two methods are implemented in *HCE2W* (the improved version of the Hierarchical Clustering Explorer 2.0 with p-value weighting and F-measure). After presenting results with discussions, we conclude our paper.

SYSTEMS AND METHODS

We selected two large Affymetrix datasets that were expected to differ in amount of mitigating, uncontrolled biological noise. Data generation for both datasets was subjected to standardized quality control and standard operating procedure. The first dataset was a mouse experimental asthma project, of 40 individual mouse lungs studied in 8 biological groups (5 mice as independent replicates within each group) (see http://microarray.cnmcresearch.org/pgadatatable.asp; U74A microarrays utilized). The studied biological variables were exposed to dust mite allergen and time points after exposure. This dataset was expected to be relatively low in confounding biological noise; entire lungs were used that effectively removed tissue heterogeneity as an uncontrolled variable, and the inbred nature of the mouse lines used effectively removed uncontrolled genetic heterogeneity between individuals.

The second dataset was a human muscle biopsy project, with 105 muscle biopsies used individually on U133A microarrays, in 11 biological (diagnostic) groups. Clinical heterogeneity, different human patients may or may not share the same exact underlying initiating biochemical problem, is a major confounding variable in most human mRNA profiling studies. It is important to point out that clinical heterogeneity was not a confounding variable in the human samples studied here, as patients within a diagnostic group were mutation-positive for the same gene (e.g. shared the same 'ground truth' in primary biochemical disorder) (Duchenne muscular dystrophy, Becker muscular dystrophy, spastic paraplegia, dysferlin deficiency, Fukutin related protein deficiency, Calpain III deficiency, Fascioscapulohumeral muscular dystrophy, Emery Dreifuss muscular dystrophy). In two groups, there is no known single gene causative of the disorder, but all patients in these groups were clearly affected by the condition as diagnosed by an acknowledged leader in that specific disorder

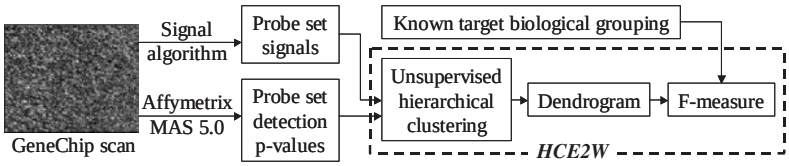


Fig. 1. Permutation study framework using unsupervised clustering in *HCE2W* (the improved version of the Hierarchical Clustering Explorer 2.0 with *p*-value weighting and *F*-measure). Inputs to the Hierarchical Clustering Explorer are two files, signal data file and *p*-value file. Each column of the two input files has values for a sample (or a chip), and the known target biological group index is assigned to each column of the signal data file. Success is measured using *F*-measure of a dendrogram and the known biological grouping.

(Acute Quadriplegic, Juvenile dermatomyositis). The 11 diagnostic groups were normal skeletal muscle from volunteers in exercise studies ($n = 19$) (Chen *et al*., 2003), Duchenne muscular dystrophy ($n = 9$) (dystrophin mutations; Chen *et al*., 2000; Bakay *et al*., 2002a,b), Acute Quadriplegic Myopathy ($n = 5$) (TGFbeta/MAPK activation; DiGiovanni *et al*., 2004), spastic paraplegia ($n = 4$) (spastin mutations; Molon *et al*., 2004), dysferlin deficiency ($n = 9$) (dysferlin gene mutations; unpublished data), Juvenile Dermatomyositis ($n = 18$) (autoimmune disease; Tezak *et al*., 2002), Fukutin related protein hypomorph ($n = 7$) (homozygous missense for glycosylation enzyme, M. Bakay, K. Gorni and E. P. Hoffman, unpublished data), Becker muscular dystrophy ($n = 5$) (hypomorph for dystrophin; see Hoffman *et al*., 1988, 1989; M. Bakay, Y.-W. Chen and E. P. Hoffman, unpublished microarray data), Calpain III deficiency ($n = 11$) (Calpain III gene mutations; see Chou *et al*., 1999; M. Bakay and E. P. Hoffman, unpublished microarray data), Fascioscapulohumeral muscular dystrophy ($n = 13$) (deletion of chromosome 4q; Winokur *et al*., 2004) and Emery Dreifuss muscular dystrophy ($n = 4$) (lamin A/C missense mutations; M. Bakay, G. Melcon and E. P. Hoffman, unpublished microarray data). This dataset was expected to have considerably greater confounding biological noise. The age and sex of subjects varied, tissue heterogeneity is known to be significant, and genetic heterogeneity between subjects is substantial. Moreover, the differences between groups were expected to be relatively minor for some groups. For example, Duchenne muscular dystrophy and Becker muscular dystrophy are both caused by mutations of the same dystrophin gene; however, Duchenne affects children and is caused by nonsense mutations, while Becker muscular dystrophy affects adults and is caused by partial-loss-of-function mutations. Thus, any attempt to distinguish some groups using unsupervised methods is expected to be considerably more challenging than for the murine lung dataset. Note that all data were subjected to the same QC/SOP protocols, as described on our website (http://microarray.cnmcresearch.org), and was generated in the same laboratory (Center for Genetic Medicine, Children's National Medical Center, Washington DC).

For the two datasets, we processed CEL files using five different probe set signal algorithms; MAS 5.0, dChip perfect match only, dChip difference, Probe Profiler and RMA. MAS 5.0 results were obtained using Affymetrix Laboratory Information Management Systems (LIMS) software, dChip results were generated using the official software release (Li and Wong, 2001a), Probe Profiler results were obtained using the Probe Profiler software by Corimbia Inc. (www.corimbia.com) and the RMA results were obtained using the affycomp package of the Bioconductor Project (http://www.bioconductor.org).

Previous comparison studies using well-known benchmark datasets such as spike-in and dilution experiments have evaluated probe set signal algorithms in terms of the known expected features (Baugh *et al*., 2001; Hill *et al*., 2001). Cope *et al*. (2003) have developed a graphical tool to evaluate probe set signal algorithms using statistical plots and summaries. They also utilized the benchmark datasets to identify the statistical features of the data for which the expected outcome is known in advance. These studies can provide a general guideline of which method is suitable for a specific investigation. While one method is better than others in general, according to the studies using the benchmark data, the 'ideal' method of probe set analysis could be different for different projects. What we suggest in this paper is a permutation study framework (Fig. 1) to help researchers choose a probe set signal algorithm that optimizes the signal-to-noise balance for their projects.

Samples (or columns in the input file) were clustered using the unsupervised hierarchical agglomerative clustering algorithm in *HCE2W* (the improved version of the Hierarchical Clustering Explorer 2.0), and the 'unsupervised' clustering results are compared with the grouping by our target biological variable. In this manner, we can evaluate the probe set signal algorithms by comparing the groupings naturally derived from the input dataset to the groupings formed by our target biological variable.

Hierarchical agglomerative clustering has been the most commonly used method for microarray data clustering (Moreau *et al*., 2002) since Eisen *et al*. (1998) first applied it to microarray data analysis. In hierarchical agglomerative clustering, when we want to cluster m data items, each data item initially occupies a cluster by itself. The most similar two clusters are then merged to construct a new cluster. The similarity/distance values between the new cluster and the remaining clusters are then updated using a specific linkage algorithm. We ran HCE2W using the average linkage method

(aka UPGMA: Unweighted Pair Group Method with Arithmetic Mean) in this study. We have previously studied the effect of different linkage algorithms in agglomerative clustering, and found that the UPGMA linkage method provided the best sample distinction by visual output (Seo *et al*., 2003). Typically for microarray data, the average linkage method gives acceptable results (Quackenbush, 2001). It is at least included or used as default measures in many standard microarray analysis tools (GeneSpring, Spotfire DecistionSite, S-plus ArrayAnalyzer, R and so on).

Average linkage is summarized as follows. Let C_n be a new cluster, a merger of C_i and C_j at a stage of the hierarchical agglomerative clustering process. Let C_k be one of the remaining clusters. Then the distance between C_n and C_k can be calculated using the following equation where Dist(C_a, C_b) is the distance (or dissimilarity) between C_a and C_b, $|C_a|$ is the number of items in a cluster C_a:

$$\mathrm{Dist}(C_n, C_k) = \mathrm{Dist}(C_i, C_k)^*|C_i|/(|C_i| + |C_j|) + \mathrm{Dist}(C_j, C_k)^*|C_j|/(|C_i| + |C_j|).$$

The merge and update are repeated until there remains only one cluster of size m.

We also developed a novel probe set weighting scheme for data analysis. Newer Affymetrix MAS 5.0 software generates a probe set detection p-value; this provides an assessment of the assuredness of the distinction between perfect match and mismatch probes across the entire 22 feature probe set, and thus a measure of the 'performance' of the probe set. It would be expected of probe sets that performed well (e.g. highly significant detection p-value) would provide 'better' data than poorly performing probe sets. A corollary to this hypothesis is that weighting of probe sets so that clustering is driven more strongly by well-performing probe sets would provide a novel noise filter that would improve clustering results. Towards this end, we used each probe set algorithm tested with and without a continuous weighting of all probe sets based upon MAS 5.0 probe set detection p-value. For each input signal dataset, we ran HCE twice to obtain 20 comparison results in total (2 experiments × 5 signal algorithms × 2). First, we ran HCE without weighting using the Affymetrix MAS 5.0 detection p-values. Second, we ran HCE with weighting each signal value in the input dataset with the detection p-values as explained in the following section. By comparing the two results, the effect of noise filtering methods can be verified across the five probe set signal algorithms and two datasets of different noise-level.

Incorporating probe set detection p-value to similarity calculation

Affymetrix noise calculations give us two outputs: one is the continuous detection p-value assignment, and the other is a simple detection call ('present/absent'). Each signal intensity value has a confidence factor—detection p-value, which contributes to determining the detection call for the corresponding probe set. When the probe set detection p-value reaches a certain level of significance, then the probe set is assigned a 'present' call, while all these probe sets with less robust signal-to-noise ratios are assigned an 'absent' call. This enables the use of a 'present call' threshold noise filter that has been used in many published studies (Chen *et al*., 2000, 2002; DiGiovanni *et al*., 2003, 2004; Hittel *et al*., 2003). In our previous study (Seo *et al*., 2003), we reported that a '10% present call' noise filter did improve the performance of probe set signal algorithms. While such 'present call' based filtering improves performance, it is clearly an arbitrary threshold method, and thus it is highly possible that potentially important signals that might be conveyed by the probe sets are filtered out.

Affymetrix MAS 5.0 uses a two-step procedure to determine the detection p-value for a probe set. It calculates the discrimination score, $R = (\mathrm{PM} - \mathrm{MM})/(\mathrm{PM} + \mathrm{MM})$ for each probe pair, and then tests R against a small positive threshold value. It assigns a rank to each probe pair according to the distance from R and the given threshold, and then the one-sided Wilcoxon's signed rank test is used to generate the detection p-value for the probe set. The discrimination score R describes the ability of a probe pair to detect its intended target, so the detection p-values are a reliable continuous indicator of how well the measured transcript is detected. Even though these detection p-values are given by Affymetrix MAS 5.0, they can be used with other signal algorithms since (1) all signal algorithms used the CEL files as their inputs and detection p-values are directly calculated from CEL files and (2) the signal algorithm and detection algorithm are independent of each other in MAS 5.0. We used the detection p-values from MAS 5.0 as a continuous weighting for probe sets for all five tested signal algorithms in this study. By involving this confidence factor in the clustering process, we believe it would give greater potential sensitivity by considering all probe sets in an analysis without the cost of poor signal-to-noise ratio.

There are many possible similarity measures for unsupervised clustering methods, and it is also possible to develop weighted versions of most similarity measures. For example, we can derive a weighted Pearson correlation coefficient as follows from the Pearson correlation coefficient that has been widely used in the microarray analysis. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be the vectors representing two arrays to be compared, and let $p(y) = [p(y_1), \ldots, p(y_n)]$ and $p(x) = [p(x_1), \ldots, p(x_n)]$ be the vectors representing p-values for x and y, respectively. Then the weighted Pearson correlation coefficient is given by

$$r_{xy} = \frac{\sum w_i (x_i - \overline{x_w})(y_i - \overline{y_w})}{\sqrt{\sum w_i (x_i - \overline{x_w})^2 \sum w_i (y_i - \overline{y_w})^2}}, \qquad (1)$$

where $w_i = [(1 - p(x_i) + (1 - p(y_i))]/2$, $\overline{y}_w = \sum w_i y_i / \sum w_i$, $\overline{x}_w = \sum w_i x_i / \sum w_i$.

```
Overall_F-measure=0
FOR EACH class i
BEGIN
    F(i)=0 // the current maximum f-measure F(i) for class i
    FOR EACH subtree j
    BEGIN
        cacluate F(i,j) using [equation 2]
        IF F(i,j) is greater than F(i) THEN F(i)=F(i,j)
    END
    Overall_F-measure = Overall_F-measure
                + (the number of samples of class i)*F(i)/(the total number of samples)
END
```

Fig. 2. The pseudo code for the overall F-measure calculation.

The weighted Pearson correlation coefficient has been used in many microarray data analysis tools, e.g. in Eisen's Cluster software (http://rana.lbl.gov/manuals/ClusterTreeView.pdf). Our extension is that we use the complement of detection p-value to calculate the weight for each term since the smaller the p-value is, the more significant the signal is. Other similarity measures such as Euclidean distance, Manhattan distance and cosine coefficient can be extended to their weighted version in a similar way to the weighted Pearson correlation coefficient.

Using external measure for evaluation of unsupervised clustering results

In our previous pilot study (Seo *et al.*, 2003), we visually inspected the unsupervised clustering results to see how well the clustering result fit to the known biological variable. Visual inspection was the right choice for the study since we only have 25 arrays of 3 different groups of samples. But since we now have 105 arrays of 11 different groups of samples, visual inspection is not realistic. Therefore, we decided to use reasonable clustering evaluation measures in addition to visual inspection in this study.

There are two kinds of clustering result evaluation measures, internal and external. The former is for the case where one is not certain what the correct clustering is. It compares the clusters using internal measures, such as distance matrix without any external knowledge. The latter is for the case where we already know the correct classes of our samples. In this study, we already know the correct class labels of samples, and thus use external measures. Possible external measures include purity, entropy and F-measures. Among them, F-measures (Van Rijsbergen, 1979, http://www.dcs.gla.ac.uk/Keith/Preface.html) have been used as an external clustering result evaluation measure in many studies across many fields including information retrieval and text-mining (Lewis and Gale, 1994; Bjornar and Aone, 1999; Cohen and Richman, 2002). Furthermore, F-measure has been successfully applied to hierarchical clustering results (Bjornar and Aone, 1999).

We applied F-measure to the entire hierarchical structure of clustering results and also to the set of clusters determined by the minimum similarity threshold in *HCE2W*. Let $C_1, \ldots, C_i, \ldots, C_n$ be the right clusters according to the target biological variable. Let $HC_1, \ldots, HC_j, \ldots, HC_m$ be the clusters from the hierarchical clustering results. In F-measure, each cluster is considered as a query and each class (or each correct cluster) is considered the correct answer of the query. The F-measure of a correct cluster (or a class) C_i and an actual cluster HC_j is defined as follows:

$$F(i,j) = \frac{2P(i,j) \cdot R(i,j)}{P(i,j) + R(i,j)},$$

where

$$P(i,j) = \frac{|C_i \cap HC_j|}{|HC_j|}, \quad R(i,j) = \frac{|C_i \cap HC_j|}{|C_i|}. \qquad (2)$$

The precision values $P(i,j)$and recall values $R(i,j)$ are defined by the information retrieval concepts. The F-measure of a class C_i is given by

$$F(i) = \max_{j=1}^{m} F(i,j). \qquad (3)$$

Finally, the F-measure of the entire clustering result is given by

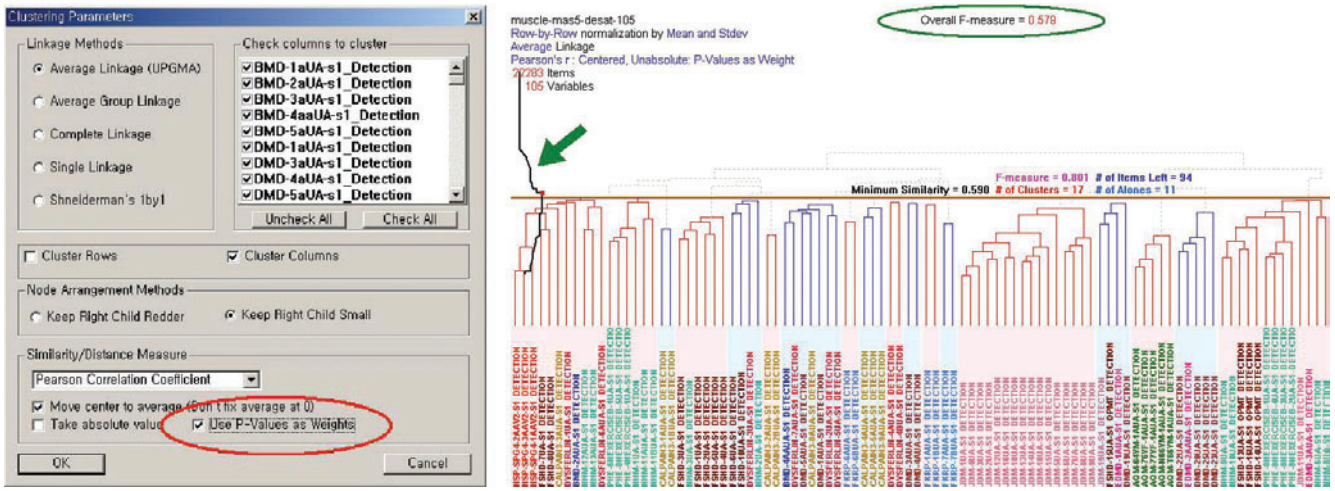
$$\sum_{i=1}^{n} \frac{|C_i|}{N} \cdot F(i), \qquad (4)$$

where N is the total number of arrays in the experiment.

The F-measure score is between 0 and 1. The higher the F-measure score is, the better the clustering result is. When we calculate the F-measure for the entire cluster hierarchy, for each external class we traverse the hierarchy recursively and consider each subtree as a cluster. Then the F-measure for an external class is the maximum of F-measures for all subtrees. The pseudo code for the overall F-measure calculation is shown in Figure 2.

Interactive visual analysis of hierarchical clustering results

HCE2 (the Hierarchical Clustering Explorer 2.0) is an interactive visualization tool for hierarchical clustering results



(a) Clustering DialogBox (b) Visualization of a clustering result of human muscle samples

Fig. 3. Software development of HCE2W for probe set selection and detection p-value weighting. (**a**) Researchers can check the option checkbox (highlighted with a red oval) to use the MAS 5.0 detection p-values as weights for distance/similarity measures. (**b**) Each sample name is color-coded by its biological class. Overall F-measure is highlighted with a green oval. The F-measure distribution is shown, as the distance from the left-hand side, over the dendrogram display as indicated by an arrow mark.

(Seo and Shneiderman, 2002, http://www.cs.umd.edu/hcil/hce/). HCE2 users load a microarray experiment dataset from a tab-delimited file, and apply their desired hierarchical clustering methods to generate a dendrogram and a color mosaic. Users can immediately observe the entire clustering result in a single screen that enables identification of high-level patterns, major clusters and distinct outliers. They can adjust the color mapping to highlight the separation of groups in the dataset. Then they start their exploration of the groupings. Instead of using fingers and pencils on a static clustering results, HCE2 users can use a dynamic query device called 'minimum similarity bar' to find meaningful groups. The Y-coordinate of the bar determines the minimum similarity threshold. A cluster (a subtree of the dendrogram) will be shown only if any two items in the cluster are more similar than the minimum similarity threshold specified by the minimum similarity bar. Thus, users see tighter clusters as they pull the bar lower to increase the minimum similarity threshold. HCE2 is provided as a public domain software tool.

A troublesome problem related to clustering analysis is that there is no perfect clustering algorithm. Clustering results highly depend on the distance calculation method and linkage method used in the clustering process. Therefore, molecular biologists and other researchers need some mechanism to examine and compare two clustering results. HCE2 users can select two different clustering methods and compare the two clustering results in a single screen. When users double click on a cluster in one clustering result, HCE2 shows the mapping to the other clustering result by connecting the same items with a line (for detail see http://www.cs.umd.edu/hcil/hce/). Through this comparison, users can determine clustering parameters that most faithfully assemble items into the appropriate biological groups according to their known biological function.

Since sample clustering is the main task of this study, we implemented an improved version of HCE2, *HCE2W*, to enable users to better understand sample (or chip) clustering results. With *HCE2W*, users can focus on either sample clustering or gene clustering by switching the main dendrogram view between sample and gene. When the sample clustering result is on the main dendrogram view, each sample name is color-coded according to its biological class so that users can assess the quality of clustering from the visual representation. To facilitate signal-to-noise ratio analyses for microarray experiments, we incorporated a weighting method for distance/similarity function and an external clustering evaluation method into *HCE2W* as described in the previous sections. *HCE2W* users can choose the option of using p-values as weights in the clustering dialogue box (Fig. 3a) and get an instantaneous graphical feedback of F-measure for each minimum similarity threshold value (Fig. 3b).

As users drag the minimum similarity bar, a line graph of F-measure score is overlaid on the main dendrogram view so that they can easily see the overall distribution of F-measure values right on the clustering result. Since the maximum F-measure value is highlighted with red dot on the F-measure distribution curve, users can easily know when to stop dragging the minimum similarity bar to get the best clustering results in terms of F-measure. This F-measure is calculated based on the current grouping determined by the current value

of minimum similarity threshold. While this *F*-measure helps users find natural groupings in the dataset, we need another measure that evaluates the clustering structure as a whole to compare many clustering results reasonably. We used the overall *F*-measure described in the previous section for this purpose. The overall *F*-measure evaluates the entire cluster hierarchy instead of considering only the groups by the current minimum similarity threshold. *HCE2W* shows the overall *F*-measure value at the top center of the main dendrogram view that is calculated by the pseudo code in the previous section (Fig. 2).

RESULTS AND DISCUSSION

We felt that the 'ideal' method of probe set analysis was likely different for different projects. Application of any noise filter can be appropriate in one context, and inappropriate in another, depending on the sensitivity desired, and the relative cost of noise that generally accompanies increased sensitivity. For example, the RMA method performs very well with known 'spike-in' RNAs, providing greater sensitivity and more stable 'signals' from probe sets. However, the greater sensitivity of the RMA method would be expected to come at a cost to specificity; the less weight given to the mismatch 'noise' filter by RMA would be expected to lead to greater signal-to-noise ratio problems in complex solutions. The testing of two cell samples that vary only due to a single highly controlled variable would be best analyzed by RMA. On the other hand, comparison of human muscle biopsy profiles (as below) are complicated by many uncontrolled variables, such as inter-individual variation, and the biopsy content of different constituent cell types (myofiber, connective tissue, vasculature). In the latter experiment, the greater sensitivity of RMA would be offset by the high cost of specificity and noise resulting from non-specific hybridization and uncontrolled variables.

We investigated the systematic alteration of signal-to-noise ratios by iteratively altering the probe set analysis algorithm (five methods), and weighting of genes using MAS 5.0 probe set detection *p*-value. The latter is, to our knowledge, a novel method of continuous weighting based upon the observed performance of each probe set, with better performing probes given greater weight in the resulting clustering. We also developed a new implementation, *HCE2W*, of our public domain HCE2 software, to effectively interrogate optimal signal-to-noise ratios by visualizing *F*-measures in unsupervised clustering analyses. To test the effectiveness of these methods, we utilized two large datasets that were expected to differ considerably in the amount of confounding and uncontrolled biological noise intrinsic to the projects; a 'noisy' 105 human muscle biopsy U133A dataset, and a 'less noisy' 40 microarray U74A inbred mouse lung dataset (see Systems and Methods section for description of the datasets). All microarrays were processed in the same laboratory, following the same quality control and standard operating procedures, thus minimizing non-biological technical noise in both projects.

All arrays were analyzed using five different signal algorithms, including Affymetrix MAS 5.0, dChip perfect match only model, dChip difference model, Probe Profiler and RMA method. We used the continuous probe set detection *p*-value as a 'weighting' function. Spreadsheets corresponding to each profile were then loaded into *HCE2W*. Unsupervised clustering of the profiles was done using permutations of signal algorithms, with and without a noise filter (continuous probe set detection *p*-value weighting). For each signal algorithm, we prepared two data files; a signal value file and a detection *p*-value file where each column is a sample and each row is a probe set. Our *HCE2W* program supports five different linkage methods: average, average group, complete, single and one-by-one linkage (Seo and Shneiderman, 2002). In this study, we choose average linkage since it is the most widely used linkage method and it was one of the most desirable linkage methods in our previous study (Seo *et al*., 2003).

For each signal algorithm, we first ran *HCE2W* without applying any noise filter. Then, *HCE2W* was run again applying the noise filter (using the detection *p*-values as a continuous weighting function) to the dataset. We visualized the unsupervised clustering of the dataset to determine the method that provided the best clustering according to our 'known' biological variable (specific biochemical defect, patient diagnosis), and thus was most effective in reducing undesirable noise. In the following bar graphs (Fig. 4), we have determined the 'performance' for each probe set algorithm using *F*-measure, either weighted by Affymetrix MAS 5.0 probe set detection *p*-value (the 'wt' bars) or unweighted (the 'no-wt' bars).

As expected, the two projects showed different results, with the inbred mouse lung data (low noise) showing greater success of unsupervised clustering into appropriate biological variables by all probe set algorithms and weighting methods. This reflects the much more highly controlled nature of the mouse data, with less confounding biological noise, as described above.

Using probe set *p*-value as a weight improved the performance of dChip difference model, Probe Profiler and RMA probe set algorithms in both datasets (Fig. 4). There was no detectable change in the performance of MAS 5.0 and dChip PM only algorithms using unsupervised clustering and *F*-measure (Fig. 4). This suggests that utilizing a continuous weighting with MAS 5.0 detection *p*-value would improve data analysis with three of the most commonly used probe set algorithms and clustering methods.

We then compared the relative performance of the different probe set algorithms. Most obvious was the differences in performance of RMA in the two datasets. RMA, a probe set algorithm that is thought to be among the most sensitive with low signal intensities, performed very well in the mouse

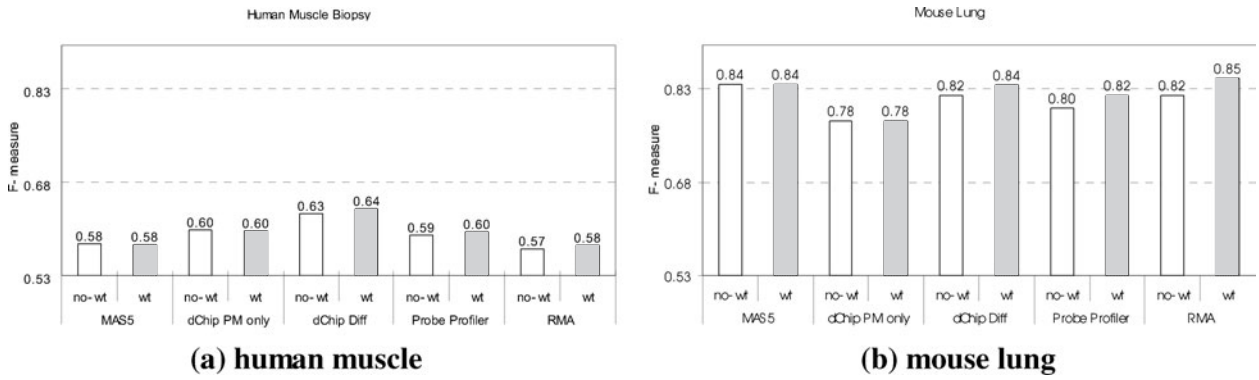


Fig. 4. External evaluation results using F-measure of unsupervised clustering for the human muscular dystrophy data and the mouse lung biopsy data. 'no-wt' bar represents the result without MAS 5.0 detection p-value weighting, and 'wt' bar represents the result with p-value weighting.

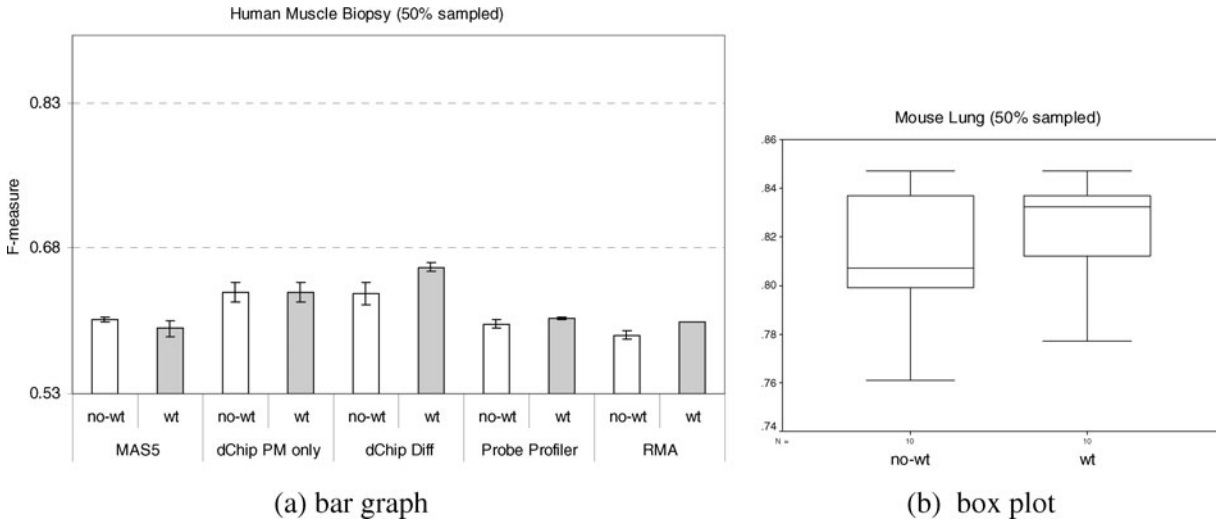


Fig. 5. Experiment results with 50% random sampled datasets. (**a**) dChip difference model with detection p-value weighting outperformed other probe set signal algorithms in human muscle data [$F(4, 10) > 14$, $p < 0.0001$]. (**b**) The result with p-value weighting ('wt') was statistically significantly better than that without weighting ('no-wt') for mouse lung data [$t(9) = -3.675$, $p = 0.005$].

dataset, if used with MAS 5.0 detection p-value weighting (Fig. 4b). However, this same RMA algorithm showed the poorest performance of all algorithms in the human data (Fig. 4a). We can conclude that the greater sensitivity of the RMA algorithm with low signal strengths is an advantage in projects with low confounding noise (e.g. inbred mouse data), but this same advantage becomes a liability driving poor performance in the human data with high levels of confounding noise. It is important to point out that the large majority of human subjects studied had a 'known' primary genetic defect (e.g. gene mutation positive), as described in the Systems and Methods section. Thus, underlying clinical heterogeneity could be ruled out in this specific project.

We performed paired t-tests with the two results to see if there is a statistically significant difference between the results with or without continuous detection p-value weighting. There was no statistically significant difference in the human muscle data. This is because the performance of MAS 5.0 and dChip PM only model was unchanged or slightly worse with the p-value weighting while those of others get better. Excluding the two cases, the difference was statistically significant. There was a statistically significant difference in the mouse lung data [$t(4) = -3.687$, $p = 0.021$]. We conclude that use of MAS 5.0 detection p-value weighting is recommended for dChip difference model, ProbeProfiler and RMA.

We then used a random-sampling permutation study to determine the statistical significance of differences in performance found between the different probe set algorithms and to verify the previous t-test result on the effect of continuous detection p-value weighting with more samples (Fig. 5). We random-sampled 50% of probe sets to partition our original datasets into two small datasets with only half the

number of probe sets. For each randomly sampled partition of input data, we repeated the previously mentioned permutation study to get two-times larger external evaluation result.

We then conducted 5×2 two-way ANOVA on the effect of five probe set signal algorithms and our novel detection p-value weighting. The analysis showed that the probe set signal algorithms did have a statistically significant effect on the external evaluation measure for both the mouse and human experiments [$F(4, 10) > 14, p < 0.0001$]. The effect of the detection p-value weighting was statistically significant only for the inbred mouse lung data [$F(1, 10) > 9, p < 0.013$]. We also re-ran paired t-tests to verify the significance of the detection p-value weighting with more samples. The t-test results showed that the continuous detection p-value weighting made a more statistically significant difference for the inbred mouse lung data [$t(9) = -3.675, p = 0.005$] than the previous result, but this again did not reach significance for the human muscle data.

Our data provide guidance of how researchers might optimize probe set algorithms and signal weights for individual projects. Our permutation study of noise level (two datasets), probe set analysis (five methods) and noise filtering (two methods—with or without detection p-value weighting) with *HCE2W* found that:

- Performances of all probe set signal methods were better with a less-noisy dataset (inbred mouse lung dataset) than with noisy dataset (human muscle biopsy).
- Noise filter using continuous probe set detection p-value improved the performances for dChip difference model, Probe Profiler and RMA.
- dChip difference model with MAS 5.0 probe set detection p-values as weights was the most consistent in maximizing the effect of the target biological variables on data interpretation of the two datasets.

While our current implementation only uses hierarchical agglomerative clustering in our permutation study framework, it is also possible to employ other unsupervised clustering algorithms, such as K-means clustering or SOM clustering. The novel F-measure and p-value weighting described here can also be used for these algorithms with minor modifications.

There are additional microarray experimental platforms available for mRNA profiling, including mechanically spotted cDNA and oligonucleotide arrays (for as review see The Tumor Analysis Best Practices Workshop, 2004). Spotted arrays typically have a single ratio per gene, or, in some cases, replicate spots per array. The single measurement possible with spotted arrays does not permit the development of algorithms to determine 'signal' across a larger 'probe set' as with Affymetrix arrays. Thus, the methods described here are not easily applicable to spotted microarrays.

CONCLUSION

Our data suggest that large microarray projects should undergo a systematic 'signal-to-noise ratio' analysis, as we have presented here. By using permutations of probe set signal algorithms, and noise reduction filters (continuous variable probe set detection p-values), with unsupervised clustering, the analysis method that most faithfully assembles profiles into the appropriate biological groups should maximize the signal from the biological variable, while minimizing the confounding noise intrinsic to the project. This results in a balanced signal-to-noise ratio assay that should provide the best balance between sensitivity and specificity. Our future plans are to implement a more extensive and automated project analysis, where these and other variables are systemically varied to achieve the best clustering into the desired biological variable groupings.

ACKNOWLEDGEMENTS

This work was supported by N01 NS-1-2339 from the NIH.

REFERENCES

Affymetrix (2001a) Microarray Suite User Guide, Version 5.0. Affymetrix, Santa Clara, CA.

Affymetrix (2001b) Data Analysis Fundamentals.

Affymetrix (2001c) Statistical Algorithm Reference Guide. Affymetrix, Santa Clara, CA, version 5 edition.

Bakay,M., Chen,Y.W., Borup,R., Zhao,P., Nagaraju,K. and Hoffman,E.P. (2002a) Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformatics*, **3**, 4–15.

Bakay,M., Zhao,P., Chen,J. and Hoffman,E.P. (2002b) A web-accessible complete transcriptome of normal human and DMD muscle. *Neuromuscul. Disord.*, **12**, S125–S141.

Baugh,L.R., Hill,A.A., Brown,E.L. and Hunter,C.P. (2001) Quantitative analysis of mRNA amplification by *in vitro* transcription. *Nucleic Acids Res.*, **29**, E29.

Bjornar,L. and Aone,C. (1999) Fast and effective text mining using linear-time document clustering. *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, NY, pp. 16–22.

Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Chen,Y.W., Hubal,M.J., Hoffman,E.P., Thompson,P.D. and Clarkson,P.M. (2003) Molecular responses of human muscle to eccentric exercise. *J. Appl. Physiol.*, **95**, 2485–2494.

Chen,Y.W., Nader,G., Baar,K.R., Hoffman,E.P. and Esser,K.A. (2002) Response of rat muscle to acute resistance exercise defined by transcriptional and translational profiling. *J. Physiol.*, **545**, 27–41.

Chen,Y.W., Zhao,P., Borup,R. and Hoffman,E.P. (2000) Expression profiling in the muscular dystrophies: identification of novel aspects of molecular pathophysiology. *J. Cell Biol.*, **151**, 1321–1336.

Chen,J., Zhao,P., Massaro,D., Clerch,L.B., Almon,R.R., DuBois,D.C., Jusko,W.J. and Hoffman,E.P. (2004) The PEPR GeneChip data warehouse and implementation of a dynamic time series query tool (SGQT) with graphical interface. *Nucleic Acids Res.*, **32**, D578–581.

Chou,F.L., Angelini,C., Daentl,D., Garcia,C., Greco,C. Hausmanowa-Petrusewicz,I., Fidzianska,A., Wessel,H. and Hoffman,E.P. (1999) Calpain III mutation analysis of a heterogeneous limb-girdle muscular dystrophy population. *Neurology*, **52**, 1015–1020.

Cohen,W.W. and Richman,J. (2002) Learning to match and cluster large high-dimensional datasets for data integration. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, NY, pp. 475–480.

Cope,L.M., Irizarry,R.A., Jaffee,H., Wu,Z. and Speed,T.P. (2003) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **1**, 1–10.

DiGiovanni,S., Knoblach,S.M., Brandoli,C., Aden,S.A., Hoffman,E.P. and Faden,A.I. (2003) Gene profiling in spinal cord injury shows role of cell cycle in neuronal death. *Ann. Neurol.*, **53**, 454–468.

DiGiovanni,S., Molon,A., Broccolini,A., Melcon,G., Mirabella,M., Hoffman,E.P. and Servidei,S. (2004) Myogenic atrophy in acute quadriplegic myopathy is specifically associated with activation of pro-apoptotic TGF beta-MAPK cascade. *Ann. Neurol.*, **55**, 195–206.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 14863–14868.

Hill,A.A., Brown,E.L., Whitley,M.Z., Tucker-Kellogg,G., Hunter,C.P. and Slonim,D.K. (2001) Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Proc. Natl Acad. Sci., USA*, **98**, 31–36.

Hittel,D.S., Kraus,W.E. and Hoffman,E.P. (2003) Skeletal muscle dictates the fibrinolytic state after exercise training in overweight men with characteristics of metabolic syndrome. *J. Physiol.*, **548**, 401–410.

Hoffman,E.P., Fischbeck,K.H., Brown,R.H., Johnson,M., Medori,R., Loike,J.D., Harris,J.B., Waterston,R., Brooke,M., Specht,L. *et al*. (1988) Dystrophin characterization in muscle biopsies from Duchenne and Becker muscular dystrophy patients. *N Eng. J. Med.*, **318**, 1363–1368.

Hoffman,E.P., Kunkel,L.M., Angelini,C., Clarke,A., Johnson,M. and Harris,J.B. (1989) Improved diagnosis of Becker muscular dystrophy by dystrophin testing. *Neurology*, **39**, 1011–1017.

Hubbell,E., Liu,W.M. and Mei,R. (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585–1592.

Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.

Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Lewis,D.D. and Gale,W.A. (1994) A sequential algorithm for training text classifiers. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, NY, pp. 3–12.

Li,C. and Wong,W. (2001a) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci., USA.*, **98**, 31–36.

Li,C. and Wong,W. (2001b) Model-based analysis of oligonucleotide arrays:model validation, design issues and standard error application. *Genome Biol.*, **2**, RESEARCH0032.1–RESEARCH0032.11.

Liu,W.M., Mei,R., Di,X., Ryder,T.B., Hubbell,E., Dee,S., Webster,T.A., Harrington,C.A., Ho,M.H., Baid,J., and Smeekens,S.P. (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, **18**, 1593–1599.

Molon,A., DiGiovanni,S., Chen,Y.W., Clarkson,P.M., Angelini,C., Pegoraro,E. and Hoffman,E.P. (2004) Large-scale disruption of microtubule pathways in morphologically normal human spastin-haploinsufficient muscle. *Neurology*, **62**, 1097–1104.

Moreau,Y., De Smet,F., Thijs,G., Marchal,K. and De Moor,B. (2002) Functional bioinformatics of microarray data: from expression to regulation. *Proc. IEEE*, **90**, 1722–1743.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.*, **2**, 418–427.

Seo,J. and Shneiderman,B. (2002) Interactively exploring hierarchical clustering results. *IEEE Comput.*, **35**, 80–86.

Seo,J., Bakay,M., Zhao,P., Chen,Y., Clarkson,P., Shneiderman,B. and Hoffman,E.P. (2003) Interactive color mosaic and dendrogram displays for signal/noise optimization in microarray data analysis. *Proc. IEEE Int. Conf. on Multimedia and Expo.*, pp. III-461–III-464.

Tezak,Z., Hoffman,E.P., Lutz,J., Fedczyna,T., Stephan,D., Bremer,E.G., Krasnoselska-Riz,I., Kumar,A. and Pachman,L.M. (2002) Gene expression profiling in DQA1*0501^{+} children with untreated dermatomyositis: a novel model of pathogenesis. *J. Virol.*, **168**, 4154–4163.

The Tumor Analysis Best Practices Working Group (2004) Guidelines: expression profiling—best practices for data generation and interpretation in clinical trials. *Nat. Rev. Genet.*, **5**, 229–237.

Van Rijsbergen,C.J. (1979) *Information Retrieval*, 2nd edn. Butterworth, London.

Winokur,S.T., Chen,Y.W., Masny,P.S., Martin,J.H., Ehmsen,J.T., Tapscott,S.J., Van Der Maarel,S.M., Hayashi,Y. and Flanigan,K.M. (2004) Expression profiling of FHSD muscle supports a defect in specific stages of myogenic differentiation. *Hum. Mol. Genet.*, **12**, 2895–2907.

Zhao,P., Iezzi,S., Sartorelli,V., Dressman,D. and Hoffman,E.P. (2002) Slug is downstream of myoD: identification of novel pathway members via temporal expression profiling. *J. Biol. Chem.*, **277**, 30091–30101.

Zhao,P., Seo,J., Wang,Z., Wang,Y., Shneiderman,B. and Hoffman,E.P. (2003) *In vivo* filtering of *in vitro* MyoD target data: an approach for identification of biologically relevant novel downstream targets of transcription factors. *Comptes Rendus Biologies*, **326**, 1049–1065.