Building a Coherent Data Pipeline in Microarray Data Analyses: Optimization of Signal/Noise Ratios Using an Interactive Visualization Tool and a Novel Noise Filtering Method

Jinwook Seo[1,2], *Marina Bakay*[1], *Yi-Wen Chen*[1], *Sara Hilmer*[1],
Ben Shneiderman[2], *Eric P Hoffman*[1,*]
12/21/2003
[1]Research Center for Genetic Medicine, Children's National Medical Center
[2]Human-Computer Interaction Lab and Department of Computer Science, University of Maryland

Abstract

Motivation: Sources of uncontrolled noise strongly influence data analysis in microarray studies, yet signal/noise ratios are rarely considered in microarray data analyses. We hypothesized that different research projects would have different sources and levels of confounding noise, and built an interactive visual analysis tool to test and define parameters in Affymetrix analyses that optimize the ratio of signal (desired biological variable) versus noise (confounding uncontrolled variables).
Results: Five probe set algorithms were studied with and without statistical weighting of probe sets using Microarray Suite (MAS) 5.0 probe set detection p values. The signal/noise optimization method was tested in two large novel microarray datasets with different levels of confounding noise; a 105 sample U133A human muscle biopsy data set (11 groups) (extensive noise), and a 40 sample U74A inbred mouse lung data set (8 groups) (little noise). Success was measured using F-measure value of success of unsupervised clustering into appropriate biological groups (signal). We show that both probe set signal algorithm and probe set detection p-value weighting have a strong effect on signal/noise ratios, and that the different methods performed quite differently in the two data sets. Among the signal algorithms tested, dChip difference model with p-value weighting was the most consistent at maximizing the effect of the target biological variables on data interpretation of the two data sets.
Availability: The Hierarchical Clustering Explorer 2.0 is available at http://www.cs.umd.edu/hcil/hce/ , and the improved version of the Hierarchical Clustering Explorer 2.0 with p-value weighting and F-measure is available upon request to the first author. Murine arrays (40 samples) are publicly available at the PEPR resource (http://microarray.cnmcresearch.org/pgadatatable.asp) (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 in time. High density oligonucleotide arrays containing up to 500,000 features are widely used for many projects in biological and medical research. The most popular 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

[*] To whom correspondence should be addressed.

single destabilizing nucleotide change in the center of the 25 nucleotide sequence. 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.

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. QC (Quality Control) and SOP (Standard Operating Procedure) 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; Di Giovanni *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 done using one or a few "test data sets" in the public domain; specifically "spike in" control data sets 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 ROC (Receiver Operating Characteristic) curves, providing an assessment of signal/noise 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 RMA (Robust Multi-array Average) (Irizarry *et al*. 2003b), would be expected to come at an increased cost of noise, where their integrity of low level signals defined by RMA in "noisy" projects would lead to data interpretations of poor integrity (Figure 1). 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 (Figure 1). We hypothesized that different probe set algorithms would show a "project-specific" performance, based upon the extent of confounding noise in a particular project.

<< Fig. 1 will be shown around here>>

The optimization of signal/noise 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 integrity 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/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 integrity, compared to probe sets that are performing poorly or detecting very low level transcripts (near background) (Figure 1). 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 PM (perfect match) and MM (mismatch) probes in a probe set and the signal intensity. In Affymetrix MAS 5.0 (Microarray Suite 5.0), the Discrimination score, R=(PM-MM)/(PM+MM), is calculated for each probe pair, and run the One-Sided Wilcoxon's Signed Rank test against a small positive number (default=0.015) to generate the detection p-value (Affymetrix, 2001a, 2001b, 2001c). Two threshold values α_1 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 α_1=0.04 and α_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; DiGiovanni *et al*. *in* press; 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 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), 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 Group Linkage. In this paper, we extend the pilot study to the extent that (1) we test not only the human muscle data of extensive noise but also the inbred mouse lung data expected to show considerably less biological (SNP, tissue) noise, (2) compare 3 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 data sets in detail. Then, our novel noise filtering methods 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 things 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.

Methods and Systems

We selected two large Affymetrix data sets that were expected to differ in amount of mitigating, uncontrolled biological noise (Figure 1). Data generating for both data sets was subjected to standardized quality control and standard operating procedure. The first data set 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 exposure to dust mite allergen, and time points after exposure. This data set is 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 data set was a human muscle biopsy project, with 105 muscle biopsies used individually on U133A microarrays, in 11 biological (diagnostic) groups. 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; Bakay *et al*. 2002b), Acute Quadriplegic Myopathy (n=5) (TGFbeta/MAPK activation; DiGiovanni *et al*. *in press*), spastic paraplegia (n=4) (spastin mutations; Molon *et al*. *in press*), dysferlin deficiency (n=9) (unpublished), Juvenile Dermatomyositis (n=18) (autoimmune disease; Tezak *et al*. 2002), Fukutin related protein hypomorph (n=7) (homozygous missense for glycosylation enzyme; unpublished), Becker muscular dystrophy (n=5) (hypomorph for dystrophin; see Hoffman *et al*. 1988, Hoffman *et al*. 1989; microarray data unpublished), Calpain III deficiency (n=11) (see Chou *et al*. 1999; microarray data unpublished), Fascioscapulohumeral muscular dystrophy (n=13) (deletion of chromosome 4q; Winokur *et al*. 2003), and Emery Dreifuss muscular dystrophy (n=4) (lamin A/C missense mutations; microarray data unpublished). This data set was expected to have considerably greater confounding biological noise (Figure 1). The age and sex of subjects varied, tissue heterogeneity is known to be significant, and genetic heterogeneity between subjects 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 data set. Note that all data was subjected to the same QC/SOP protocols, as described on our web site (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 data sets, we processed CEL files using five different probe set signal algorithms; MAS5, dChip perfect match only, dChip difference, Probe Profiler and RMA. MAS5.0 results were obtained using Affymetrix LIMS (Laboratory Information Management Systems) 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 affycomp package of the Bioconductor Project (http://www.bioconductor.org).

Previous comparison studies using well-known benchmark data sets 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*. have developed a graphical tool to evaluate probe set signal algorithms using statistical plots and summaries. They also utilized the benchmark data sets to identify the statistical features of the data for which the expected outcome is known in advance (Cope *et al*., 2003). 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 (see Figure 2) to help researchers choose a probe set signal algorithm that optimizes the signal/noise balance for their projects.

<< Fig. 2 will be shown around here>>

Samples (or columns in the input file) were clustered using unsupervised hierarchical agglomerative clustering algorithm in *HCE2W* (the improved version of the Hierarchical Clustering Explorer 2.0), and the "unsupervised" clustering results compared to the grouping by our target biological variable. In this way, we can evaluate the probe set signal algorithms by comparing the groupings naturally derived from the input data set to the groups by our target biological variable. Hierarchical clustering algorithms have been widely used to analyze expression profile data sets. Among many kinds of hierarchical clustering algorithms, the agglomerative algorithm is a de facto standard for microarray experiment data analysis. When we want to cluster *m* data items, initially, each data item occupies a cluster by itself. Among the current clusters, the most similar two clusters are merged together to make a new cluster. Then, the similarity/distance values between the new cluster and the remaining clusters are updated using a linkage method. We ran HCE using UPGMA (Unweighted Pair Group Method with Arithmetic Mean) linkage method in this study; we found this algorithm to provide better sample distinction by visual output in our pilot report (Seo *et al*. 2003). UPGMA linkage is summarized as follows. Let C_n be a new cluster, a merge of C_i and C_j at a stage of hierarchical agglomerative clustering process. Let C_k be a remaining cluster. Then the distance between C_n and C_k is defined as :

$$Dist(C_n,C_k)=Dist(C_i,C_k)*|C_i|/(|C_i|+|C_j|) + Dist(C_j,C_k) *|C_j|/(|C_i|+|C_j|)$$

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 that 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 data set, we ran HCE twice to obtain20 comparison results in total (2 experiments x 5 signal algorithms x 2). First, we ran HCE without weighting with the Affymetrix MAS5.0 detection p-values. Second, we ran HCE with weighting each signal value in the input data set 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 data sets 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 determine 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 those probe sets with less robust signal/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; DiGiovanni*et al*. *in press*; Hittel, 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 MAS5 uses a two-step procedure to determine the detection p-value for a probe set. It calculates the discrimination score, R=(PM-MM)/(PM+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-side Wilcoxon signed rank test to generate the detection p-value for the probe set. The discrimination score R describes the ability for a probe pair to detect its intended target, so the detection p-values are a reliable continuous indicator 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 firstly all signal algorithms used the CEL files as their inputs and detection p-values are directly calculated from CEL files, secondly the signal algorithm and detection algorithm are independent of each other in MAS5. We used the detection p-values from MAS5 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 a cost of poor signal-noise ratio.

There are many possible similarity measures for unsupervised clustering methods, and it is also possible to develop a weight measure for most similarity measures. For example, we can derive a weighted Pearson correlation coefficient as follows from the Pearson correlation coefficient thathas been widely used in the microarray analysis. Let $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ be the vectors representing two arrays to be compared, and let $p(y) = (p(y_1), ..., p(y_n))$ and $p(x) = (p(x_1), ..., p(x_n))$ 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 $w_i = \frac{(1 - p(x_i)) + (1 - p(y_i))}{2}$, $\bar{y}_w = \sum w_i y_i / \sum w_i$, $\bar{x}_w = \sum w_i x_i / \sum w_i$

We should note here that we use the complement of detection p-value to calculate the weight for each term since the smaller 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 now we have 105 arrays of 11 different groups of samples, so visual inspection is not realistic. Therefore, we decided to use a 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, F-measures and etc. Among them, F-measures (Rijsbergen, 1979) have been used as an external clustering result evaluation measure in many studies across many fields including information retrieval, and text-mining (Lewis, 1994; Bjornar, 1999; Cohen, 2002). Furthermore F-measure has been successfully applied to hierarchical clustering results (Bjornar, 1999).

We applied F-measure on the entire hierarchical structure of clustering results and also on 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 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:

$$F(i,j) = \frac{2P(i,j) \cdot R(i,j)}{P(i,j) + R(i,j)}, \text{ where } P(i,j) = \frac{|C_i \cap HC_j|}{|HC_j|},\ R(i,j) = \frac{|C_i \cap HC_j|}{|C_i|}. \qquad (2)$$

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

$$\sum_{i=1}^{n} \frac{|C_i|}{N} \cdot F(i), \text{ where } N \text{ is the total arrays in the experiment.} \qquad (4)$$

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 3.

<< Fig. 3 will be shown around here>>

Interactive Visual Analysis of Hierarchical Clustering Results

HCE2 (the Hierarchical Clustering Explorer 2.0) is an interactive visualization tool for hierarchical clustering results (Seo and Shneiderman, 2002; http://www.cs.umd.edu/hcil/hce/). HCE2 users load a microarray experiment data set 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 data set. 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 through 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 (see http://www.cs.umd.edu/hcil/hce/ for detail). 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/noise 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 (Figure 4a) and get an instantaneous graphical feedback of F-measure for each minimum similarity threshold value (Figure 4b).

<< Fig. 4 will be shown around here>>

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 data set, 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 (Figure 3).

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/noise 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.

Here, we investigated the systematic alteration of signal/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/noise ratios by visualizing F-measures in unsupervised clustering analyses. To test the effectiveness of these methods, we utilized two large data sets 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 data set, and a “less noisy” 40 microarray U74A inbred mouse lung data set (see Methods and Description for description of the data sets). 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 5

different linkage methods: UPGMA, Average Group, Complete, Single, and One-by-one linkage (Seo *et al*. 2002). In this study, we choose UPGMA linkage since it is 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 noise filter (using the detection p-values as a continuous weighting function) to the data set. We visualized the unsupervised clustering of the data set 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 undesired noise. In the following bar graphs (Figure 5), 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 un-weighted (the "no-wt" bars).

<< Fig. 5 will be shown around here>>

As expected, the two projects showed different results, with the inbred mouse lung data showing greater success of unsupervised clustering into appropriate biological variables by all probe set algorithms and weighting methods. Using probe set p-value as a weight improved the clustering results in most cases except MAS5 and dChip PM only model with the muscle dystrophy data. This suggests that utilizing a continuous weighting with detection p value would improve data analysis with most probe set algorithms and clustering methods. Secondly, the more sensitive RMA algorithm appeared to perform more poorly with the noisier human muscle data (as we have previously reported; Seo *et al*. 2003), but performed considerably better with the cleaner mouse lung data. This fulfilled our predictions, as the inbred mouse data set of whole lung has less sources of confounding biological variability (inter-individual "SNP" noise, tissue heterogeneity, age, sex, stage of disease).

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 MAS5 and dChip PM only model slightly become 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). To make our analysis more statistically meaningful, we random-sampled our original data sets to generate more external evaluation results since it was difficult and expensive to generate more real data sets with 105 arrays. We random-sampled 50% of probe sets to partition our original data sets into two small data sets with only half number of probe sets. For each random sampled partition of input data, we repeated the previously mentioned permutation study again to get twice-bigger external evaluation result. Then we ran 5x2 two-way ANOVA tests and paired t-test for the result. ANOVA tests revealed that the probe set signal method did have a statistically significant effect on the success of unsupervised clustering for both experiments (F(4,10)>14, p<0.0001) (Figure 6(a)). T-test results showed that the continuous detection p-value weighting did make a statistically significant difference for the inbred mouse lung data (t(9)=-3.675, p=0.005) (Figure 6(b)), but it didn't for the human muscle data. For the inbred mouse lung data the detection p-value weighting also had a statistically significant effect (F(1,10)>9, p<0.013).

<<Fig. 6 will be shown around here>>

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

- **Performances of probe set signal methods were better with a less-noisy data set (inbred mouse lung data set) than with noisy data set (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 weight was the most consistent at maximizing the effect of the target biological variables on data interpretation of the two data sets.**

Conclusion

In conclusion, we feel that each project should undergo a “signal/noise” 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/noise 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 http://www.affymetrix.com/products/software/specific/mas.affx

Affymetrix (2001b) Data Analysis Fundamentals, https://www.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf

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

Baugh, L.R., Hill, A.A., Brown, E.L., 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*, 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, 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**, *in press.*

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.

Chou, F.L., Angelini, C., Daentl, D., Garcia, C., Greco, C. Hausmanowa-Petrusewicz, I., Fidzianska, A., Wessel, H., Hoffman, E.P. (1999) Calpain III mutation analysis of a heterogeneous limb-girdle muscular dystrophy population, *Neurology*, **52**, 1015-20.

Cohen, W. W. and Richman, J. (2002) Learning to match and cluster large high-dimensional data sets for data integration, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 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**, 454-68.

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

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. U S A.*, **98**, 31-6.

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.*, **548.2**, 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., Kupsky, W., Chamberlain, J., Caskey, C.T., Shapiro, F. and Kunkel, L.M. (1988) Dystrophin characterization in muscle biopsies from Duchenne and Becker muscular dystrophy patients, *New Eng. J. Med.*, **318**, 1363-1368.

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

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 W. A. Gale (1994) A Sequential Algorithm for Training Text Classifiers, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3-12.

Li, C. and Wong, W. (2001a) Model-based analysis of oligonucleotide arrays:Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U S A.,* **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

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

Rijsbergen, C. J. Van (1979) Information Retrieval, 2nd ed. Butterworth, London. (http://www.dcs.gla.ac.uk/Keith/Preface.html)

Seo, J. and Shneiderman, B. (2002) Interactively exploring hierarchical clustering results, *IEEE Computer*, **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, *Proceedings of the IEEE International Conference on Multimedia and Expo*, 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**, 4154-63.

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. (2003) Expression profiling of FHSD muscle supports a defect in specific stages of myogenic differentiation, *Hum. Mol. Genet.*, in press.

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.

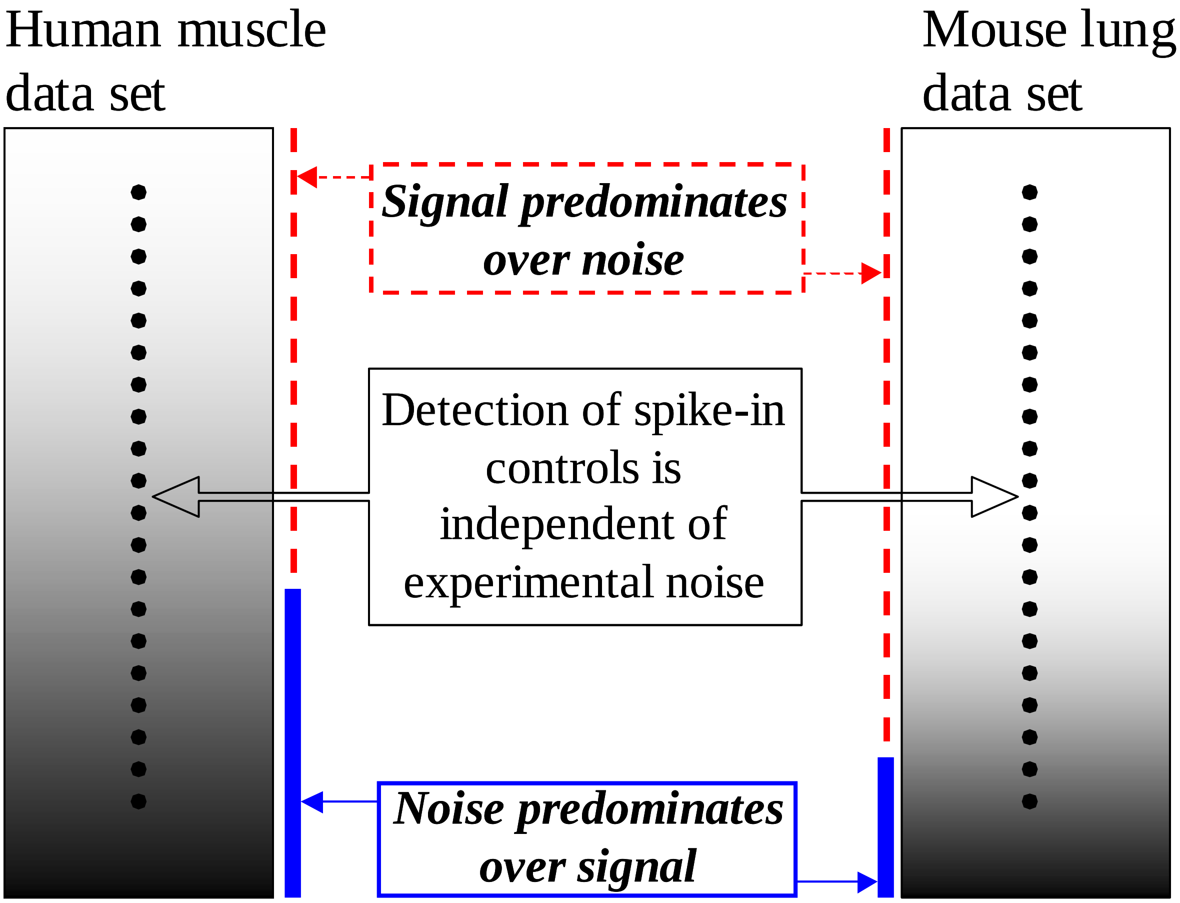

Human muscle data set
Mouse lung data set
Signal predominates over noise
Detection of spike-in controls is independent of experimental noise
Noise predominates over signal

Fig. 1. Signal/Noise ratios in the two data sets. Human muscle data set has relatively high level of confounding noise while mouse lung data set has low level of confounding noise.

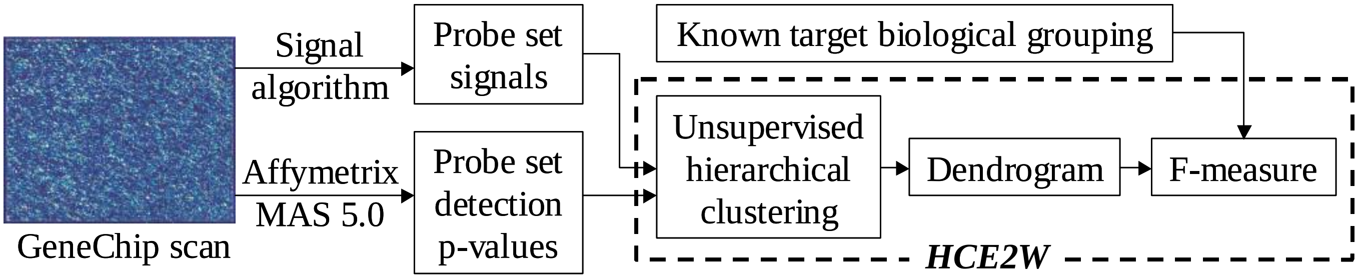
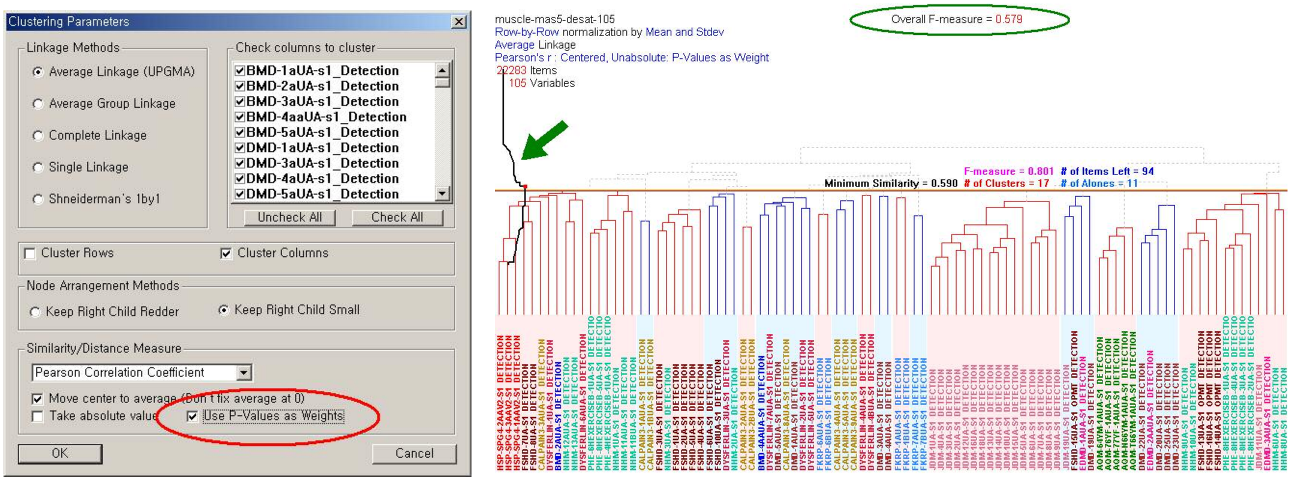

Signal algorithm
Affymetrix MAS 5.0
GeneChip scan
Probe set signals
Probe set detection p-values
Known target biological grouping
Unsupervised hierarchical clustering
Dendrogram
F-measure
HCE2W

Fig. 2. 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.

```
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. 3. The pseudo code for the overall F-measure calculation



Clustering Parameters

Linkage Methods
- ● Average Linkage (UPGMA)
- ○ Average Group Linkage
- ○ Complete Linkage
- ○ Single Linkage
- ○ Shneiderman's 1by1

Check columns to cluster
- ☑ BMD-1aUA-s1_Detection
- ☑ BMD-2aUA-s1_Detection
- ☑ BMD-3aUA-s1_Detection
- ☑ BMD-4aaUA-s1_Detection
- ☑ BMD-5aUA-s1_Detection
- ☑ DMD-1aUA-s1_Detection
- ☑ DMD-3aUA-s1_Detection
- ☑ DMD-4aUA-s1_Detection
- ☑ DMD-5aUA-s1_Detection

☐ Cluster Rows ☑ Cluster Columns

Node Arrangement Methods
- ○ Keep Right Child Redder
- ● Keep Right Child Small

Similarity/Distance Measure
- Pearson Correlation Coefficient
- ☑ Move center to average (Don't fix average at 0)
- ☐ Take absolute value
- ☑ Use P-Values as Weights

muscle-mas5-desat-105
Row-by-Row normalization by Mean and Stdev
Average Linkage
Pearson's r : Centered, Unabsolute: P-Values as Weight
22283 Items
105 Variables

Overall F-measure = 0.579

F-measure = 0.801 # of Items Left = 94
Minimum Similarity = 0.590 # of Clusters = 17 # of Alones = 11

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

Fig. 4. (a) Researchers can check the option checkbox (highlighted with a red oval) to use the MAS5 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 side, over the dendrogram display as indicated by an arrow mark.

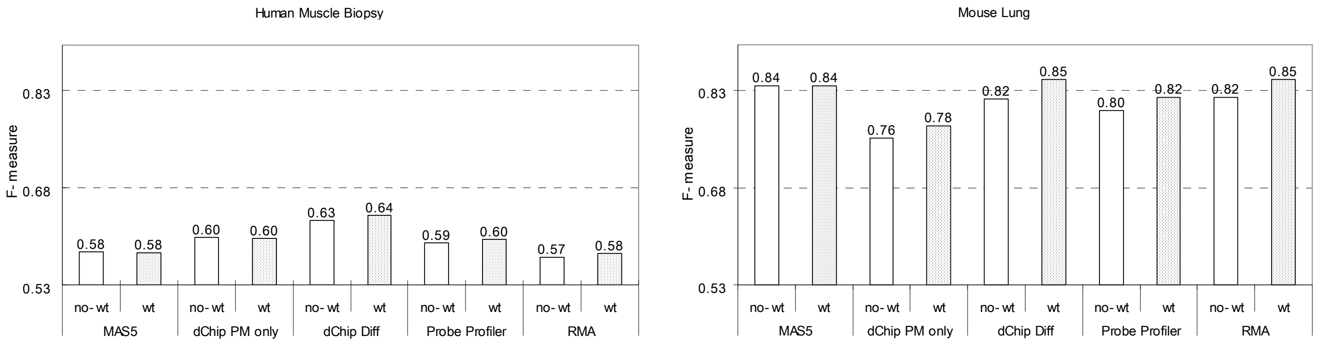
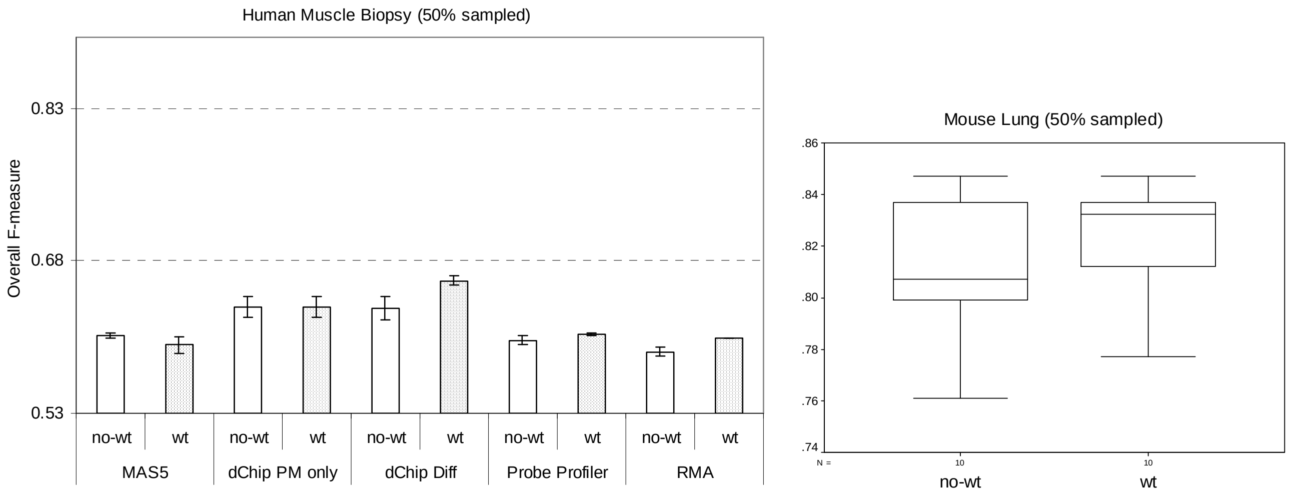


Fig. 5. Clustering results evaluation results using F-measure for the human muscular dystrophy data and the mouse lung biopsy data. "no-wt" bar represents the result without p-value weighting, and "wt" bar represents the result with p-value weighting.



(a) (b)

Fig. 6. Experiment results with 50% random sampled data sets. (a) dChip difference model with detection p-value weighting outperformed other methods. (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"). t(9)=-3.675, p=0.005)