STATISTICAL ISSUES IN THE DESIGN AND ANALYSIS OF GENE EXPRESSION MICROARRAY STUDIES OF ANIMAL MODELS

Lisa M. McShane

Joanna H. Shih

Aleksandra M. Michalowska

Biometric Research Branch, National Cancer Institute

Bethesda, MD 20892

October 15, 2003

Corresponding author:

Lisa M. McShane, Ph.D.
Biometric Research Branch, DCDT, NCI
Room 8126, Executive Plaza North, MSC 7434
6130 Executive Boulevard
Bethesda, MD 20892-7434
(301) 402-0636 (voice)
(301) 402-0560 (fax)
email: lm5h@nih.gov

Abstract: 150 words; Main text: 7920 words; References: 49 (1199 words); Tables: 1; Figures: 5 (430 words in legends);

Running head: Statistical issues in gene expression microarray studies of animal models

Abbreviations: MM, mismatch; PM, perfect match; SAM, Significance Analysis of Microarrays; SOMs, self-organizing maps

ABSTRACT

Appropriate statistical design and analysis of gene expression microarray studies is critical in order to draw valid and useful conclusions from expression profiling studies of animal models. In this article, several aspects of study design are discussed, including the number of animals that need to be studied to ensure sufficiently powered studies, usefulness of replication and pooling, and allocation of samples to arrays. Data preprocessing methods for both cDNA dual-label spotted arrays and Affymetrix-style oligonucleotide arrays are reviewed. High-level analysis strategies are briefly discussed for each of the types of study aims, namely class comparison, class discovery, and class prediction. For class comparison, methods are discussed for identifying genes differentially expressed between classes while guarding against unacceptably high numbers of false positive findings. Various clustering methods are discussed for class discovery aims. Class prediction methods are briefly reviewed, and reference is made to the importance of proper validation of predictors.

Key words: gene expression profiling; animal models of cancer; statistics; study design; analysis of microarray data

INTRODUCTION

Gene expression microarray analysis of animal models of mammary cancer holds the potential for a better understanding of mammary cancer development and the mechanisms of action of agents for prevention and treatment of mammary cancers. Careful statistical design and analysis of these microarray studies will enhance the insights gained and ensure the validity of conclusions drawn from these studies. Microarray studies of human breast tumors are complicated by the heterogeneous clinical presentation of the tumors, the genetic diversity in the normal human breast tissue backgrounds from which the cancers arose, and the varied environmental factors to which the breast tissues have been exposed. Further, it is difficult to obtain human specimens from early or sub-clinical stages of breast cancer progression. Carefully designed animal model experiments can control for factors such as genetic strain, tumor induction mechanism, and timing of observations throughout the stages of cancer progression. Better understanding of the genetic alterations that occur in the controlled animal model settings should translate to an improved understanding in the more complex human tumor setting.

Appropriate study design and statistical analysis recommendations cannot be made for a particular study until there is a clear statement of the scientific aims of the study. Microarray analysis permits the measurement of expression levels of thousands of genes simultaneously on each specimen, generating an "expression profile" for each specimen. However, the generation of large amounts of data does not remove the need for a clear scientific focus and sound study design. As described by others (1-3), the aims of most microarray studies fall into three general categories: class comparison, class discovery, and class prediction. For the class comparison aim, the goal is to determine whether the average expression pattern in one group (class) of specimens differs from that in another group, and, if they differ, what genes appear to be responsible for the differences. In class discovery, there are no pre-specified classes, and the goal is to discover natural groupings of specimens or genes with the property that there is some homogeneity within groups but differences between groups. The third aim of class prediction involves the development of a multivariate mathematical model that takes as input an expression profile for a specimen or gene and gives as output a prediction of the class to which the specimen or gene belongs.

In the next section, we discuss design issues to be considered for microarray animal experiments. The design recommendations may differ depending on the scientific aims and array platform. In this paper we refer to two widely used array platforms – dual-label spotted cDNA arrays and single-label Affymetrix-style oligonucleotide arrays. We feel that sound statistical design is critical to the validity of conclusions drawn from these studies, and animal models provide the opportunity for carefully designed and well-controlled studies. Due to its great importance, the reader will notice that we have devoted a large part of our discussion in this paper to study design. Following the design discussion, analysis strategies appropriate for each type of study aim are discussed. Embedded in those discussions is advice regarding proper interpretation of the analysis results. We conclude with summary remarks presented in the last section.

STUDY DESIGN

Animal model experiments are particularly well suited to address class comparison-type questions such as whether there are differences in gene expression between different tissue types, between stages of tumor progression, or between tumors induced by different mechanisms. They also provide the opportunity to assess activity of genes induced by an experimental intervention or to study the coordinated expression of multiple genes through time series experiments. Factors one must consider in the design of such experiments include the number of animals or specimens needed per group (or per timepoint in a time series experiment), the number and type of technical replicates, the optimal sample pooling strategies (if pooling is needed at all), and, for dual-label arrays, how to pair specimens for co-hybridization to the arrays. Detailed discussion of many of these elements is given in (3-6). We highlight the main points of those discussions in this paper. Since class prediction is not a frequent goal of animal model experiments, we will not provide a separate discussion of design issues for class prediction studies, although a number of the issues are the same as for class comparison studies (3). Studies with class discovery as their goal are by their very nature more exploratory, and therefore, tend to have fewer controlled design factors. However, we will make brief mention of a few design issues that remain relevant, such as technical replication and allocation of specimens to dual-label arrays.

Types of Designs

There are a variety of designs that can be considered for class comparison studies. The simplest type of design is a single-factor design in which one compares among two or more groups corresponding to levels of the factor. For example, one might wish to compare the gene expression profiles of tumors induced by different carcinogenic agents. A time series experiment can also be viewed as a single-factor design with each timepoint corresponding to a group. For example, a drug may be delivered to a large collection of animals with tumors, and different subsets of animals are sacrificed after different lengths of observation. One could be interested in testing for expression differences between any two timepoints or in examining for trends in the expression trajectories across timepoints. For the design described, the arrays at different timepoints correspond to independent groups of animals. There are situations in which arrays at the different timepoints would not be strictly independent; for example, in cell culture experiments, it is possible that RNA could be harvested on repeated occasions from each of several cultures. A detailed discussion of the ramifications of such designs is beyond the scope of this paper. However, a key point is that it would be important when analyzing data generated by these types of designs to account for culture effects that might result in greater similarity across timepoints of samples originating from the same culture compared to samples originating from independent cultures.

More complicated multi-factorial designs can be used when there are several factors of interest, with each factor having multiple levels. For example, an investigator might be interested in studying the anti-tumor effects (as reflected in expression profiles) of three experimental chemotherapy drugs (D1, D2, D3) on two strains of mice (S1, S2) whose tumors were induced by two different tumor-inducing agents (A1, A2). This experiment has three factors: strain (2 levels), tumor-inducing agent (2 levels), and drug (3 levels). Multi-factorial designs allow not only examination of the effect of each factor individually, but also they allow examination of interactions between the factors.

Loosely speaking, an interaction is present when two or more factors together have a greater or lesser effect than would be predicted by the sum of their individual effects. For example, suppose that strains S1 and S2 develop tumors reflective of their different normal background expression profiles, and that tumors induced by each agent show characteristic alterations in expression in a particular subset of genes, with the subset depending on the agent but not on the strain. However, suppose that the alterations in expression resulting from treatment with the drugs are different depending on the strain. In the situation described, we would say that there is a "main effect" of tumor-inducing agent, but an interaction between strain and drug. See figure 1 for a schematic representation of an example of a strain by drug interaction.

In multi-factorial designs one must also take special care to avoid confounding effects of different factors. For example, it would be a faulty design to allocate only drug D1 to S1 mice and drugs D2 or D3 to S2 mice because one could then not separate the effects due to strain and drug. Similar considerations apply to technical factors. If the specimens studied in an array experiment were analyzed using arrays produced in two different print batches or different manufacturing lots, it would be unwise to assay all specimens from S1 mice using the first batch of arrays and to assay all specimens from S2 mice using the second batch of arrays.

As stated earlier, class discovery studies have no pre-specified classes, so the above discussion of single and multi-factor designs mostly does not apply. However, care should be taken to avoid introducing artifactual "class" structure resulting from technical variations such as mid-stream changes in array print batches or manufacturing lots, fluorescent dye lots, common reference pool (see discussion of common reference design for dual-label arrays below), or even laboratory technician.

Allocation of Samples to Dual-label Arrays

Optimal allocation of samples to dual-label arrays has been a subject of considerable debate and has generated much confusion. Here we briefly discuss allocation schemes in the context of class comparison studies, but we mention which of the designs discussed are suitable when class discovery is also an aim. The reader is referred to Dobbin and Simon (7) for a more comprehensive discussion. Yang and Speed (4) discuss some allocation schemes for multi-factorial or time-series experiments.

The purpose of the dual-labeling for printed glass slide arrays is to provide a means of standardization to account for variation in size and shape of spots across arrays and for possibly unequal distribution of sample across an individual array. There are many ways in which samples can be paired for analysis on the arrays. Below we describe three major types of sample allocation designs for dual-label arrays for studies involving group comparisons: the common reference design, the balanced block design, and the loop design (7-8). We compare them on the basis of both practical considerations and statistical efficiency. We use the word “efficiency” here in the following sense: if the goal in a study is to estimate the mean difference in expression between different classes, then the more efficient design is the one that allows estimation of the difference with greatest precision (smallest variance).

For the traditional common reference design, each study “test” sample is paired with an aliquot of RNA from a large reference pool, so the number of arrays required is equal to the number of test samples. See Figure 2a. The reference pool serves as an internal standard. The test samples are always labeled with one dye color, and the reference pool samples are always labeled with the other color dye. Because the common reference pool is measured on every array in this design, this design has been criticized as wasteful of arrays, particularly given that the reference sample is often of no biological interest in its own right. However, the advantages of the reference design are that it allows for easy and efficient comparisons of any number of groups and cluster analyses. Also, results from different studies using a common reference design with the same reference pool can theoretically be compared, assuming no other major technical differences exist between the studies.

An alternative allocation strategy is a balanced block design. In the simplest case of a two group comparison, test samples from the two groups being compared are randomly paired, one from each group, for co-hybridization to each array. Labeling of test samples with red versus green dye is alternated from one array to the next. See figure 2b. Only one aliquot of each test sample is required, and the number of arrays required is half of the number required for the common reference design. Disadvantages are that if the number of test samples in the two groups is not the same or if there is more than one two-group comparison of interest, complicated modifications of the design are needed. Also, if the investigator wishes to conduct some class discovery analyses such as cluster analysis (see Class Discovery section) in addition to the comparison of two pre-specified classes, this design is problematic because the pairing of different test samples on arrays introduces artificial correlations between the samples in each pair that could influence results of cluster analyses. A variation on the balanced block design is a situation in which tumor and normal tissue from the same animal is paired on an array, with each array corresponding to a different animal. This design is a good choice if one views the normal tissue background as a nuisance and is interested in comparing expression in tumor relative to normal tissue. This approach would avoid detection of expression differences between tumors that were solely due to differences in expression in the normal tissues from which the tumors arose.

The loop design requires two aliquots of each test sample that are labeled alternately with the red and green dyes and co-hybridized on arrays with two other test samples. See figure 2c. It requires the same number of arrays as the common reference design (equal to the number of distinct test samples). For a fixed number of arrays, the loop design will be less efficient than the balanced block design for a two group class comparison study aim, but more efficient than the common reference design for comparison of only two classes. However, as the number of classes being compared increases, the efficiency advantage of the loop design compared to the reference design is lost. The loop design is very inefficient for cluster analyses (efficiency for cluster analyses refers to the accuracy with which clusters are defined). Furthermore, if there are technical difficulties with some arrays rendering them unusable, the "loop" is broken and appropriate statistical analysis of the data becomes extremely difficult. The loop design also has the important practical limitation that it can only be used if duplicate aliquots of each test sample are available – a major limitation in animal studies in which the specimens may be very small. Our general recommendation is to avoid the loop design.

Model-based methods of analysis (9-11) must be used to analyze data derived from balanced block or loop designs. Simpler analyses involving t-tests or F-test or their non-parametric counterparts are possible when the common reference design is used (see Class Comparison section).

Replication

A commonly asked question is whether it is necessary to perform replicate array assays. This discussion has to begin with an understanding of the different sources of variation in microarray experiments. Except possibly in very preliminary pilot studies, the goal of a microarray experiment is to make inferences about biological populations. The only way to make those inferences is to sample multiple individuals from each of the biological populations under study. We will refer to these as biological replicates. One might ask whether it is also advisable to obtain technical replicates. Technical replicates include multiple aliquots of RNA from a single batch of RNA extracted from each test specimen, multiple RNA extractions from each test specimen, or reverse fluor ("dye swap") replicates. Dye swap replicates involve the co-hybridization of the same two RNA samples to two arrays, but with the dye labeling of the specimens reversed between the first array and the second.

Technical replication can be useful for quality control purposes to assure that the microarray assays are working properly, and reverse fluor replicates for some arrays will be necessary if the investigator wishes to interpret individual ratios from a single array experiment. Expression measurements from technical replicates, such as multiple aliquots from a single RNA batch or multiple RNA extractions from a specimen, can be averaged to increase precision of the expression measures or can be used as checks on clustering results (see Class Discovery section), but it is important to understand that no amount of technical replication is a replacement for biological replication. For example, even if one performs 100 technical replicates on a single biological sample from one population and 100 technical replicates on a single sample from another population, conclusions can only be drawn about the two biological samples that were studied, and not about the two populations from which they came. Replication efforts are most profitably directed at biological replication.

Unless one wishes to interpret individual ratios, it is generally not necessary to perform dye-swap experiments for every array in class comparison studies. Usually dye swaps of 5-10 arrays will suffice to estimate gene-specific dye biases for the entire experiment (12). If there is no plan to interpret individual ratios and only analyses to compare classes will be conducted, no dye swap experiments are needed, because gene-specific dye biases cancel out when expression levels are compared between groups. If the goal of the study is class discovery, then the value of dye swap experiments will depend on the distance metric used in the cluster analyses (see Class Discovery section) and the size of the dye biases relative to the degree of separation expected between the classes one is trying to discover.

Number of biological replicates

For two-group class comparison studies using single-channel arrays or dual-label arrays with the common reference design, we describe a simple method (6, 13a) for determining the number of independent biological samples needed to detect a specified level of differential expression with a desired statistical power. For multi-group or multi-factor designs or for dual-label array experiments using sample allocation schemes other than the common reference design, the calculations are more complicated and beyond the scope of this paper (see references 12 and 14). In the calculations that follow, it is assumed for dual-label arrays that the expression measurement for a gene is a log-ratio; for single-channel arrays, the expression measurement is assumed to be the log signal.

The sample size calculations presented here focus on detection of differential expression for a single gene between two independent classes. The calculations will require specification of parameters that describe the anticipated variability of measurements for the gene within each class and a target minimum detectable average expression difference. The variability is specified through σ, the standard deviation of expression level for the gene within each class; δ denotes the difference in average expression level between the two classes. For example, with base 2 logarithms, $\delta = 1$ corresponds to a 2-fold average difference between the two classes. In our experience with several animal model studies utilizing inbred strains of mice or rats, we have observed that within-class standard deviation σ is usually small, typically ranging from 0.1 to 0.5. For example, in the study conducted by Desai et al. (15), the median (across genes) within-class standard deviation was in the range 0.2-0.25 for their 8.7K arrays and in the range 0.3-0.5 for their 2.7K arrays.

It should be noted that σ can vary according to a number of factors, including the specific gene examined, the particular animal model under study, and the laboratory performing the microarray assays. If the samples size estimate is based on standard deviations anticipated for genes with average levels of within-class variability, that sample size may not provide adequate power to detect differential expression for genes that tend to have very large within-class variability. If an investigator does not have readily available estimates of variability from related experiments conducted in his or her laboratory, a choice must be made to either accept estimates from other studies or to conduct a pilot study. Our recommendation would be to consider a combination of the two options. A limitation in estimating variability from a pilot study is that it generally requires fairly large sample sizes to obtain precise estimates of standard deviations. Given the imprecision expected in estimates obtained from a small pilot study, one would be well-advised to take the additional step of comparing such preliminary estimates with estimates available from other labs. This also provides the added safeguard that if the estimates obtained in the pilot appear substantially larger than estimates reported for comparable situations in other labs, it might prompt one to investigate for technical problems that could be introducing large amounts of measurement error into the array measurements and to check for specimen quality problems.

Two types of error may occur in conducting a statistical hypothesis test. One is false discovery (false positive), and the other is false negative. Declaring a gene that is truly not differentially expressed between classes as significantly differentially expressed is referred to as a false discovery. Conversely, mistakenly declaring a gene that is truly differentially expressed between classes as not significantly differentially expressed is referred to as a false negative. False negatives are likely to occur when the number of independent specimens in each class is too small or variability within each class is large. We need to control the two error rates in sample size planning through the specification of significance level and power. The significance level, denoted by α, is the probability of declaring that the expression level of a gene is different between classes when in fact there is no difference ($\delta = 0$). The false negative rate, denoted by β, is the probability of declaring that the expression level between the two classes is not different when in fact the mean difference is not zero (δ ≠ 0). The statistical power is equal to 1-β, the probability of obtaining statistical significance when the true difference in mean expression level between the two classes is δ. The specification of α and β determines the average number of false discoveries and false negatives. Since gene expression levels for thousands of genes will be examined, the value set for α should be small. If *N* genes are tested, then the expected number of false discoveries is *N* × α or less, where the maximum value *N* × α applies to the most extreme situation in which all *N* genes tested are truly non-differentially expressed. If, on the other hand, *m* of the *N* genes are truly differentially expressed with mean difference at least δ, then the expected number of false negatives is *m* × β or less. Common choices for α and β in the analysis of microarray gene expression data are 0.001 and 0.05, respectively. Thus if 10,000 non-differentially expressed genes are examined, the average number of false discoveries will be 10 or less. If 1% of the 10,000 examined genes are truly differentially expressed, then the average number of false negatives will be 5 or less.

The approximate formula for the total sample size required, based on the assumption that the mean expression measurement is approximately normally distributed among specimens of the same class, is

$$n = (T_{n-2,\alpha/2} + T_{n-2,\beta})^2 \sigma^2 / (\delta^2 q(1-q))$$

where $T_{n-2,p}$ denotes the (1-*p*) percentile of the *t* distribution with *n-2* degrees of freedom with *p* equal to α/2 or β, and q is the proportion of biological specimens allocated to the first class. (The value *n* is the total sample size and should not be confused with the number of genes, *N*.) An iterative computational procedure is used to solve the above equation for the total sample size *n*. For example, using α = 0.001, β = 0.05, σ= 0.25 and δ= 1 in the above formula gives approximately 12 as a total sample size with 6 in each class. Often the number of biological specimens is limited and one is interested in knowing the power that the statistical test would have for detecting a certain mean fold difference in gene expression between two classes. In this case, the formula for calculating power is given by

$$Power = P_{n-2}[(\delta/\sigma)\sqrt{nq(1-q)} - T_{n-2,\alpha/2}] + P_{n-2}[-(\delta/\sigma)\sqrt{nq(1-q)} - T_{n-2,\alpha/2}]$$

where P_{n-2} denotes the cumulative probability distribution for a t distribution with n-2 degrees of freedom. Table I lists number of independent specimens per class and power for detecting the difference in mean expression level between two classes at 0.001 significance level for a range of σ and δ applicable to animal microarray experiments. It shows that a larger standard deviation requires a larger sample size to achieve the same power as achieved in the presence of a smaller standard deviation, and a larger sample size allows one to detect a smaller fold-difference. In addition, the bottom rows of the table show that when the number of specimens is fixed, power decreases with increasing within-class variability.

Pooling of samples

Pooling involves mixing together RNA from several biological specimens before labeling and hybridization. Investigators may wish to pool samples because there is not enough RNA available from each individual sample to hybridize on an array or because they want to reduce the number of arrays for the purpose of saving cost. However, a single pool for each class does not allow estimation of biological or technical variability necessary for statistical inference. Taking multiple aliquots from each pool and applying each aliquot to an array does not alleviate the problem, because variability from multiple aliquots only allows one to make inference to that pool which represents a finite subset of all specimens from that class. In order to make a valid statistical inference, one needs to generate multiple independent pools from each class (6, 16). Each pool is constructed from a different set of biological specimens and represents a true replication in the population. However there are still some disadvantages to this approach. First, it does not allow one to understand the contribution of individual RNA samples to the observed expression measurements, thus complicating the identification of outlier or low quality samples. Second, there may be great power loss in comparing gene expression profiles between classes.

Comparisons of power between designs without pooling and designs with different pooling strategies depend on the ratio of biological variation to technical variation (16), total number of specimens and number of specimens per pool. Biological variation represents variability among different specimens, and technical variation represents measurement error arising from array-to-array variability. The overall variance in measured expression levels for a particular gene within each class is the sum of biological variance and technical variance; we will refer to this total variance as σ^2. Kendziorski et. al (16) observed that the ratio of biological to technical variance in an RT-PCR experiment was in the range approximately 4 to 10 fold and suggested that microarray-based assays might have somewhat smaller variance ratios. We estimated a variance ratio of approximately 3.7 when we examined one unpublished data set on human cell lines for which technical replicates had been performed. Pooling is more beneficial when the ratio is large, because the impact of pooling is to reduce the biological variability, and this results in increased precision of the expression measures (16). Figure 3 plots power as a function of effect size defined by δ/σ with 2 independent specimens per pool, significance level set to 0.001, and with the ratio of biological variance to technical variance set to 4. For example, under the design with the total number of samples n = 12 and effect size = 4, without pooling (number of arrays = 12) there is approximately 95% power to declare gene expression differs between the two classes, whereas with pooling of specimens two at a time (number of arrays = 6) power drops to 30%. Even if the total number of original specimens is increased to 16 (number of arrays = 8), the power is decreased to 79%. In order to achieve at least 95% when pooling specimens by pairs, 20 specimens (number of arrays = 10) are required. Power could also be calculated for designs with pool sizes other than 2, but we will not present those additional calculations here. There is a moderate effect of the ratio of biological to technical variance on these power comparisons. Consider the above selected design (effect size = 4, significance level = .001, 2 specimens per pool). If the ratio is reduced to 1 (i.e. technical variance is equal to biological variance), then pooling results in power 22%, 67% and 93% when the original number of specimens is equal to 12 (6 arrays), 16 (8 arrays), and 20 (10 arrays), respectively. In conclusion, pooling can result in substantial loss of power for detecting genes differentially expressed between groups and generally should be avoided unless necessary due to insufficient amount of RNA available from individual samples. If pooling is necessary for reasons of small RNA samples or preferred due to the high cost of arrays, then the number of individual samples utilized in the experiment should be increased appropriately to compensate for the decrease in power caused by pooling samples.

STATISTICAL ANALYSIS

The analysis of microarray data is a multi-step process. First, there must be an analysis of the data quality, followed by calculation of gene-level summary statistics that have been suitably corrected and normalized to remove noise due to experimental artifacts and to correct for systematic array and dye effects. Following this, the main analyses can be conducted to address the study aims.

Quality Screening

For either single-sample or dual-label arrays, there must be either visual or automated screening of arrays to detect experimental artifacts. For single-sample Affymetrix-style arrays, these artifacts may result, for example, from manufacturing defects in the chip or debris in the hybridization chamber. Dual-label glass slide arrays are prone to problems such as dust specs, scratches, fibers, bubbles, or excessive fluorescence of a slide itself, due to problems with the coating or washing of the slide. Array elements flagged as poor quality during image analysis or associated with low signals and large variability should be excluded from analysis. Entire arrays or regions of arrays should be excluded if too many spots have been flagged, if the range of signals is narrow, or if the signals are uniformly low.

Calculation of Gene-level Expression Measures and Normalization for Dual-label cDNA Arrays

For dual-label cDNA arrays using reference design, preliminary gene expression measures are calculated for each gene as the signal in the test sample divided by the signal in the reference channel at the feature location corresponding to that gene. Signal is usually defined as foreground intensity for the gene's feature minus background intensity (many methods are available for defining background signal). The distribution of these signal ratios will frequently be centered away from 1.0 due to different physical properties of the dyes, different uptake of the two dyes, error in measuring sample amounts, and adjustments of the photomultiplier tube when scanning the slides. The reader is referred to Yang et al. (17) for a discussion of various normalization procedures to correct for these biases. The normalization procedures range from application of a single global correction factor to all ratios to correction factors that are dependent on signal intensity, slide region, or print tip. The final normalized log ratios are the measures of gene expression that are used in subsequent analyses.

Normalization and Probe Set Summary Calculation for Affymetrix Arrays

In order to control for array-to-array technical variability arising from variation in amount of sample, chip manufacture, and hybridization and scanning conditions, normalization is also required for Affymetrix arrays. In this array platform, each gene is interrogated with a probe set consisting of 11-20 oligonucleotide pairs. Every pair is comprised of a perfect match (PM) 25-mer oligonucleotide and a mismatch (MM) oligonucleotide.

Each PM is typically a 25-mer oligonucleotide with sequence exactly complementary to some nucleotide sequence of length 25 from an exon in the gene of interest. The MM member of each pair is a 25-mer oligonucleotide with sequence that differs from the PM sequence by a single nucleotide in the center position. Normalization is performed at either gene (probe set) level or at probe level, and probe set summaries are calculated either before or after normalization. Specifically, Affymetrix Microarray Suite, Version 5.0 (18) calculates probe set summaries based on robust weighted averages of PM and MM pair differences using unnormalized probe-level data, and then the resulting gene-level summaries are normalized. Others (19-22) have proposed model-based methods for calculating probe-set summaries and normalizing data. A major difference between model-based methods and the Affymetrix approach is that the model-based approaches account for probe effects which are estimated using data across multiple arrays, whereas the Affymetrix algorithm does not account for probe effects. In addition, model-based methods normalize probe-level data before calculating model-based probe set summaries, or they incorporate probe-level normalization into their model fitting procedures. Generally, summary measures produced by model-based methods have a substantial advantage in that they tend to stabilize variability at low signals. Similar to normalization for dual-label cDNA arrays, normalization can be done globally, or the magnitude of correction may vary depending upon the signal range. The reader is referred to Bolstad et al (23) for a detailed account of various normalization methods for Affymetrix array data. Software implementing the above preprocessing methods for Affymetrix data is available as part of the R package Affy of the Bioconductor software project http://www.bioconductor.org. These final normalized gene summary measures (or their log transforms which we prefer) are then used in subsequent analyses to address the study aims.

Analysis Strategies for Addressing Study Aims

Appropriate analysis methods for the class discovery aim are often termed unsupervised methods because they do not require pre-specified class information, whereas, the best analysis strategies for the class comparison and class prediction aims are supervised methods that make use of available class information (1, 24). In the discussions that follow, it is assumed that the expression measures being analyzed are normalized log ratios (from a common reference design) in the case of dual-label cDNA arrays or normalized gene summary measures for Affymetrix arrays. We do not discuss analysis strategies based on model-based methods that are applied to individual channel data.

Class Comparison

The goal in comparing specimens from different pre-specified classes is to establish whether gene expression profiles differ between classes, and if they do, identify genes that are differentially expressed between classes. One usually identifies genes that are differentially expressed between known classes of specimens using univariate analyses on each gene, such as t-tests for two classes or F-tests for more than two classes. The F-test is used to determine whether there is some difference among the classes. When an F-test is rejected, one may be interested in assessing which classes have different expression levels with various post-hoc comparisons (25a). For example, classes may represent different transgenic mouse models, and one could assess pairwise differences.

Both the t-test and F-test assume that the mean of the expression measurement in each class is normally distributed, which may not hold, particularly when the number of specimens or animals per group is small. In this case, one could use the non-parametric counterparts of the t-test and F-test, namely the Wilcoxcon test (26a) and Kruskal-Wallis test (26b) or permutation tests. Another example is the study of gene expression in a mammary cancer progression model in mice. Here, breast tissues might be taken from mice at different time points following tumor initiation. In this case, one might be interested not only in whether there is a difference in gene expression between stages of mammary cancer but also in the trajectory of gene expression change over time. One could examine for trends in these trajectories using regression methods (25a-b).

As described in the section discussing the number of biological replicates needed, it is important to take into account the problem of multiple testing in order to control for false discoveries. Several multiple testing procedures are available to identify differentially expressed genes between classes, and they differ by controlling for different aspects of false discoveries. A simple procedure is to control the average number of false discoveries to be no more than some number *u* by conducting each of the *N* (number of genes) univariate tests at significance level *u/N*. The procedure of Tusher et al. (27), called SAM, aims to control the expected proportion of false discoveries. Efron et al. (28) describes an empirical Bayes procedure that controls the expected proportion of false discoveries. A potential drawback to these approaches is that although they control the expected number or expected proportion of false discoveries, the actual number of false discoveries may be quite different from the expected number. Alternatively, a multivariate permutation approach such as the one proposed by Korn et al. (29) can be used to control for the number or proportion of false positives. Their procedure, generalizing the step-down permutation procedures of Westfall and Young (30) to control the familywise error rate, controls number or proportion of false discoveries with stated confidence. The procedure of Korn et al. (29) is available in BRB-ArrayTools software (http://linus.nci.nih.gov/BRB-ArrayTools).

Analysis of class comparison studies is particularly challenging when the number of independent specimens available in each class is small (between 2 and 5). In this case, the variance estimate for each gene is not very accurate and the power of statistical tests is low. Several statistical approaches based on borrowing strength from the data of other genes have been developed. One approach is to assume that the variances are the same for all genes and estimate the common variance by pooling within-class variances over all genes. However, the assumption of a common variance is often not realistic. Other statistical approaches take an intermediate approach with respect to the variances of the genes. Basically, they assume that different genes have different variances, but that these variances represent random quantities generated from a common distribution. It is this additional distributional assumption on the variances that allows one to borrow strength from information of other genes. The statistical approaches vary depending on the additional distributional assumptions on other parameters, including the empirical Bayes approach of Efron et al. (28), Bayesian approaches of Baldi and Long (31) and Broet et al. (32) and the Frequentist approach of Wright and Simon (33). In general, these statistical approaches to detect differentially expressed genes will be most useful when the within-class variability of gene expression tends to be small and the spread of the gene-specific variances is not too large; these criteria are often satisfied for specimens of inbred stains of mice or a cell line within the same class. The approach developed by Wright and Simon is available in BRB-ArrayTools (http://linus.nci.nih.gov/BRB-ArrayTools).

If one is interested in establishing that there is an overall difference in average expression profile between classes without specifically identifying genes responsible for those differences, global permutation tests can be used. To conduct a permutation test, one first chooses a summary measure of difference between classes. Examples of summary measures include number of genes univariately significant at 0.001 level and sum of squared univariate t-statistics. Then one calculates the summary measure statistic many times, each time randomly permuting the class labels. The statistical evidence is assessed by the permutation p-value which is the proportion of permutations with the same or larger value of the summary measure as the one observed in the original data. Obtaining a p-value of less than 0.01 or 0.05 is sufficient to establish that the expression profiles between classes are different. The global test described above is available in BRB-ArrayTools software (http://linus.nci.nih.gov/BRB-ArrayTools).

Class Discovery

When class discovery is the goal of the microarray investigation, unsupervised analysis strategies such as clustering methods can be used. Methods to find cluster structure in data include hierarchical clustering, self-organizing maps, *k*-means clustering, likelihood-based methods, and Bayesian methods. In this paper, we will briefly discuss hierarchical clustering, *k*-means clustering, and self-organizing maps along with some graphical display techniques. We refer the reader to the review article of Jain et al. (34)

for discussion of other clustering algorithms. Sometimes expression data are pre-processed prior to clustering to remove genes that vary little across the arrays or to center the expression data within each gene. For more discussion of these pre-processing steps, see Simon et al. (13b).

A description of hierarchical clustering methods applied to microarray data was given by Eisen et al. (35). We describe it here in the context of clustering specimens, but it can also be used to cluster genes. Hierarchical clustering of specimens starts by defining a distance between the profiles of each pair of specimens. The algorithm proceeds by merging the two closest (most similar) specimens first, and then successively merging specimens or groups of specimens in order of greatest similarity. Two distance metrics commonly used in clustering gene expression profiles are Euclidean distance and one minus the Pearson correlation coefficient. Euclidean distance measures, in absolute terms, the closeness of two profiles, whereas correlation measures the similarity of patterns in the sense of how closely the values in one profile can be approximated by a linear function (scalar multiple or shift) of the values in the other profile. Euclidean distance has the property that when clustering specimens, a constant shift in all expression measurements for any given gene (for example due to gene-specific dye bias) will subtract out in the distance calculation. While the distance metric defines distance between two individual specimens, additional specifications are required to define distance between two clusters. This specification refers to the linkage method. Average linkage specifies merging clusters whose average distance between all pairs of items (one item from each cluster) is minimized. Complete linkage specifies merging clusters to minimize the maximum distance within any resulting cluster. Single linkage merges clusters at minimum distance from one another.

Hierarchical clustering results can be depicted by a tree structure called a dendrogram. Figure 4a shows a dendrogram representing a clustering of cDNA microarray data obtained in an experiment comparing mouse models of cancer (15). The distance metric used was one minus Pearson correlation, and average linkage was used. These data were generated using a 2.7K mouse oncochip spotted cDNA array. Mammary tumor RNA sample from each of 36 mice was analyzed on a different array, co-hybridizing with samples obtained from a common reference pool of normal mouse mammary tissue RNA. Normalized $\log_2$ ratios were calculated as described in Desai et al. (15). Genes not having at least two measurements in each of the mouse model subgroups were excluded from the analyses. Of the remaining genes, only those for which the ratio of the 95th percentile of ratios to the 5th percentile of the ratios exceeded 3 were used. This left 1137 genes for the analyses. At the bottom of the tree, each specimen occupies its own cluster. At the top of the tree, all specimens have been merged into a single cluster. Mergers between two specimens, or between two clusters of specimens, are represented by horizontal lines connecting them in the dendrogram. The height of each horizontal line represents the distance between the two groups it merges. A number of differences in the smaller clusters are observed when different linkage methods are used (see figures 4b and 4c). Overall, the conclusion is that there is some evidence for clustering according to animal model class because there were many within-class merges at distance of 0.2 or less (correlation 0.8 or more), and between-class merges tended to occur at higher distances. However, the gene expression-based clusters do not completely duplicate the known mouse model classes.

A color image plot is another popular display for hierarchical clustering results. It is a rectangular array of boxes, with the color of each box representing the expression level of one gene on one array. Shades of red are usually used to represent degrees of increasing expression, and shades of green are used to represent decreasing expression. Each column of boxes represents an array, and each row corresponds to a gene. The columns are ordered according to the array ordering in the dendrogram of clustered specimens. The rows are ordered according to the clustering of genes. The end result is a color image with patches of red and green color indicating combinations of genes and specimens that exhibit high or low expression. See figure 5 at (link to website) for an image plot of the mouse model data. Further examples are given in Eisen et al. (35).

The classical *K*-means algorithm and self-organizing maps (SOMs) are both partitional clustering methods that have been applied to microarray data (36,37). Partitional methods try to find a single partition of the items being clustered, whereas hierarchical methods look for a nested series of partitions. The *K*-means algorithm, as described by MacQueen (38), begins with either an initial partition of the objects into *K* subgroups or an initial specification of *K* cluster centroids, and the algorithm iteratively reallocates objects and updates centroids. SOMs can be viewed as a generalization of *K*-means in which the algorithm favors the choice of clusters whose centers can be arranged on a two-dimensional grid of nodes with distances roughly representing distances between centroids. SOMs can be particularly useful for clustering expression profiles of genes obtained in time course experiments (37). However, both *k*-means and SOMs are known to be sensitive to choices of parameters such as initial cluster centroids, grid configurations, and learning rates.

It is important to understand when evaluating cluster analysis results that clustering algorithms will always produce clusters even if the clusters formed only represent noise in the data. For this reason, some assessment of cluster reproducibility is advisable. McShane et al. (39) and references therein, describe methods for global tests of clustering and for assessment of reproducibility of particular clusters. A simple means of partially assessing clustering results is to include a few technical replicate arrays for some specimens and see if the replicates cluster together. This would provide evidence that the cluster analyses are finding at least some real clusters.

Class Prediction

Class prediction has been an infrequent aim of animal model microarray studies. While for human tumor studies there has been interest in developing diagnostic and prognostic gene expression-based predictors, most animal experiments are not geared toward developing clinical tools, but rather toward understanding tumor biology. Application of class prediction methods in animal models might more likely be directed at predicting functional classes of genes than predicting classes of specimens. Our treatment of class prediction methods in this paper will be very brief.

Class predictors are mathematical functions that take as their input a vector of measurements on an object, e.g. expression profiles of specimens or genes, and then output a predicted class membership. Numerous methods for building class predictors have been applied to microarray data, including but not limited to Fisher linear discriminant analysis (40) and its variants weighted voting method (1) and compound covariate prediction (41-42), regression trees (43), neural networks (44), support vector machines (45-46), and nearest centroid and relatives (47). We refer the interested reader to Dudoit et al. (48), and references therein for description and comparison of several of these methods. Interestingly, Dudoit et al.'s (48) findings were that simpler class prediction methods, such a diagonal linear discriminant analysis and nearest neighbor methods, performed better than more complicated methods on the microarray data sets they considered. Radmacher et al. (42) and Simon et al. (24) discuss considerations in the proper building and validation of class predictors to avoid overfitting (fitting to noise in the data) and to obtain unbiased assessments of a predictor's accuracy.

DISCUSSION

In this review we have highlighted what we feel are the most important issues in the design and analysis of animal model microarray studies. Our intent was not to provide an encyclopedic account of all published design and analysis methods that have been proposed for microarray studies. We decided what to include on the basis of our combined many years of experience in consulting and collaborating with investigators at our institution on the design and analysis of both human and animal studies. From those experiences we have gained a sense for what design and analysis strategies for microarray studies are practical and useful without necessarily being complex. For readers interested in more comprehensive discussions of many of the topics covered in this paper at a level accessible to scientists with limited statistical background, we recommend the books by Simon et al. (13) and Knudsen (49).

Acknowledgments

We thank Dr. Jeffrey E. Green and Dr. Kartiki V. Desai for allowing us use of the mouse model microarray data. We are grateful to Dr. Ming-Chung Li for assistance with figures.

References

1. Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Dowing J, Caligiuri M, Bloomfield C, Lander E. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531-537.

2. Miller LD, Long PM, Wong L, Mukherjee S, McShane LM, Liu ET. Optimal gene expression analysis by microarrays. Cancer Cell 2002;2:353-361.

3. Simon R, Radmacher MD, and Dobbin K. Design of studies using DNA microarrays. Genet Epidemiol 2002;23:21-36.

4. Yang YH, and Speed T. Design issues for cDNA microarray experiments. Nat Rev Genet 2002;3:579-588.

5. Simon R and Dobbin K. Experimental design of DNA microarray experiments. Biotechniques 2003;34:S16-S21.

6. Dobbin K, Shih J, and Simon R. Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. J Natl Cancer Inst 2003; 95(18):1362-1369.

7. Dobbin K and Simon R. Comparison of microarray designs for class comparison and class discovery. Bioinformatics 2002;18:1438-1445.

8. Kerr MK and Churchill GA. Statistical design and the analysis of gene expression microarray data. Genet Res 2001;77:123-8.

9. Kerr MK and Churchill GA. Experimental design for gene expression microarrays. Biostatistics 2001;2:183-201.

10. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, and Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 2001;8:625-638.

11. Lee M-L, Kuo FC, Whitmore GA, and Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA 2000;97:983-9839.

12. Dobbin K, Shih J, and Simon R. Statistical design of reverse dye microarrays. Bioinformatics 2003;19(7):803-810.

13. Simon R, Korn E, McShane LM, Radmacher MD, Wright GW, Zhao Y. Design and analysis of DNA microarray investigations. Springer Verlag (a: chapter 3; b: chapter 9; in press, publication anticipated December 2003).

14. Neter J, Wasserman W, Kutner MH. Applied linear statistical models, 2nd edition. Homewood (Illinois): Richard D. Irwin, Inc; 1985, pp. 547-549, 700-702, 818, 919-920.

15. Desai KV, Xiao N, Wang W, Gangi L, Greene J, Powell JI, Dickson R, Furth P, Hunter K, Kucherlapati R, Simon R, Liu ET, Green JE. Initiating oncogenic event determines gene-expression patterns of human breast cancer models. Proc Natl Acad Sci USA 2002;99:6967-6972.

16. Kendziorski CM, Zhang Y, Lan H, and Attie AD. The efficiency of pooling mRNA in microarray experiments. Biostatistics 2003; 4:465-477.

17. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30(4):e15.

18. Affymetrix. Affymetrix Microarray Suite User Guide. 5th ed. Santa Clara (CA): Affymetrix; 2001.

19. Li C and Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA 2001;98:31-36.

20. Li C and Wong WH. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol 2001;2:research0032.1-0032.11.

21. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B and Speed TP. Summaries of Affymetrix genechip probe level data. Nucleic Acids Res 2003;31(4):**e15**.

22. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4(2):249-264.

23. Bolstad BM, Irizarry RA, Astrand M, and Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics 2003;19(2):185-193.

24. Simon R, Radmacher MD, Dobbin K, and McShane LM. Pitfalls in the analysis of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14-18.

25. Snedecor GW and Cochran WG. Statistical methods. 8th edition. Ames (IA): Iowa State University Press; 1989 (a: pp. 234-236; b: chapter 9).

26. Hollander M and Wolfe DA. Nonparametric Statistical Methods, 2nd edition. New York: John Wiley & Sons, Inc; 1999 (a: pp. 106-124; b: 190-201).

27. Tusher V, Tibshirani R, and Chu G. Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. Proc Natl Acad Sci USA 2001;98:5116-5121.

28. Efron B, Tibshirani R, Storey JD, and Tusher V. Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 2001;96:1151-1160.

29. Korn EL, Troendle JF, McShane LM and Simon R. Controlling the number of false discoveries: Application to high-dimensional genomic data. J Stat Plan Infer 2003 (in press).

30. Westfall, PH and Young, SS. Resampling-based multiple testing. New York: John Willey & Sons, Inc; 1993, pp. 72-74.

31. Baldi P and Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 2001;17(6):509-519.

32. Broet P, Richardson S and Radvanyi F. Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. J Comput Biol 2002;9(4):671-683.

33. Wright G and Simon R. The random variance model for differential gene detection in small sample microarray experiments. Bioinformatics 2003 (in press).

34. Jain AK, Murty MN, and Flynn PJ. Data clustering: A Review. ACM Comput Surv 1999;31(3):264-323.

35. Eisen MB, Spellman PT, Brown PO, and Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863-14868.

36. Tibshirani R, Hastie T, Eisen M, Ross D, Botstein D, and Brown P. Clustering methods for the analysis of DNA microarray data. (Stanford, CA: Stanford University Department of Statistics Technical Report).

37. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999;96:2907-2912.

38. MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability 1967;1:281-97.

39. McShane LM, Radmacher MD, Freidlin B, Yu R., Li M., Simon R. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics 2002;18:1462-1469.

40. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics 1936;7:179-188.

41. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J. Gene expression profiles of hereditary breast cancer. N Engl J Med 2001;344:549-548.

42. Radmacher MD, McShane LM, and Simon R. A paradigm for class prediction using gene expression profiles. J Comput Biol 2002;9:505-511.

43. Breiman L, Friedman J, Stone C, and Olshen R. Classification and regression trees. Belmont (CA): Wadsworth; 1984.

44. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7:673-679.

45. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16:906-914.

46. Brown MPS, Grundy WN, Lin D, Cristiani N, Sunet CW, Furey TS, Ares M, Haussler D. Knowedge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 2000;97:26

47. Tibshirani R, Hastie T, Narasimhan B, and Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002;99:6567-6572.

48. Dudoit S, Fridlyand J, and Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77-87.

49. Knudsen S. A biologist's guide to analysis of DNA microarray data. New York: John Wiley&Sons; 2002.

Table I. Power and Sample Size for Planning Animal Model Microarray Studies

σ (SD of log expression level for the gene within each class)	δ (true difference in mean log expression level between the two classes)	Fold-difference (2^{δ})	Number of independent specimens (or animals) per class	Power (%)
0.1	1	2	4	95
0.2	1	2	5	95
0.25	1	2	6	95
0.3	1	2	8	95
0.4	1	2	11	95
0.5	1	2	15	95
0.1	1.32	2.5	3	95
0.2	1.32	2.5	4	95
0.25	1.32	2.5	5	95
0.3	1.32	2.5	6	95
0.4	1.32	2.5	8	95
0.5	1.32	2.5	10	95
0.1	1	2	5	> 99
0.2	1	2	5	97
0.25	1	2	5	82
0.3	1	2	5	60
0.4	1	2	5	28
0.5	1	2	5	14
0.1	1.32	2.5	5	> 99
0.2	1.32	2.5	5	> 99
0.25	1.32	2.5	5	98
0.3	1.32	2.5	5	90
0.4	1.32	2.5	5	59
0.5	1.32	2.5	5	34

FIGURE LEGENDS

Figure 1. Schematic representation of an interaction between two factors of a three-way factorial experiment. The effect of tumor-inducing agent for a given gene is represented by the vertical distance between each pair of parallel lines. The parallel lines on the right are shifted upwards relative to the parallel lines on the left, representing strain effects. The interaction between strain and drug is represented by the fact that all three drugs result in equivalent gene expression for strain 1, but for strain 2 gene expression is the same for drugs 1 and 2 and then increases for drug 3.

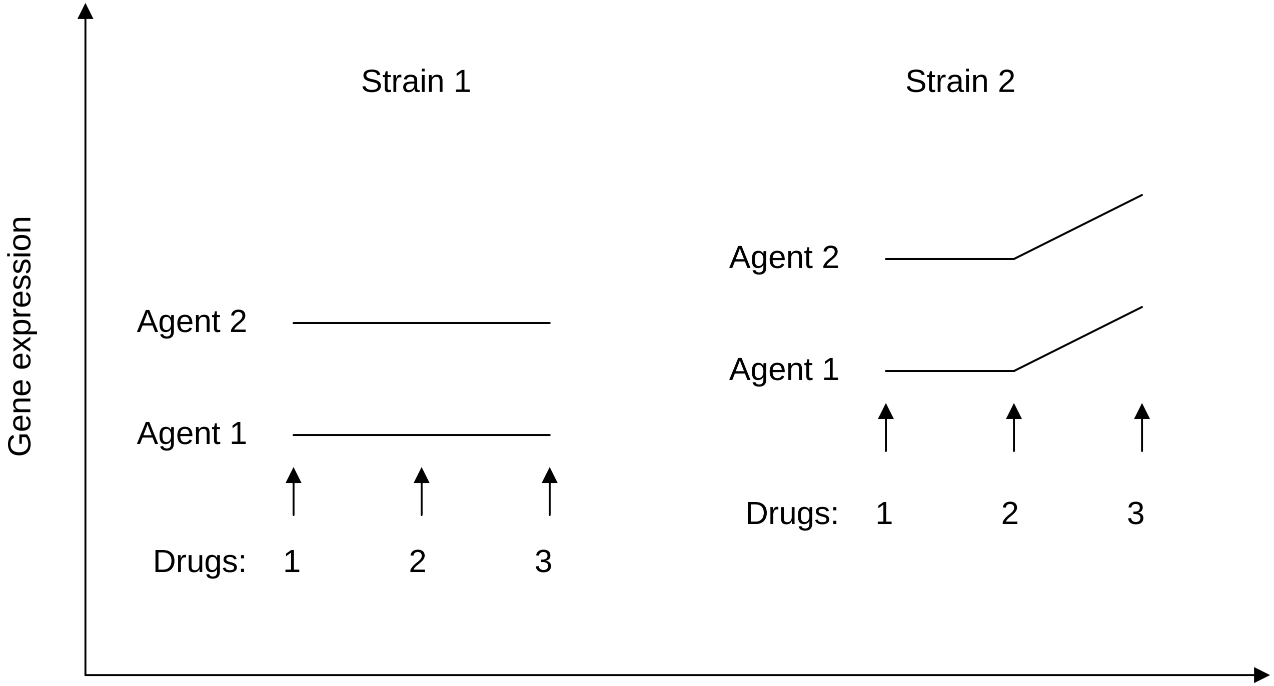
Figure 2. Schematic of allocation schemes for pairing samples on dual-label cDNA microarrays for studies involving a two-group comparison: a) Common Reference Design, b) Balanced Block Design, c) Loop Design

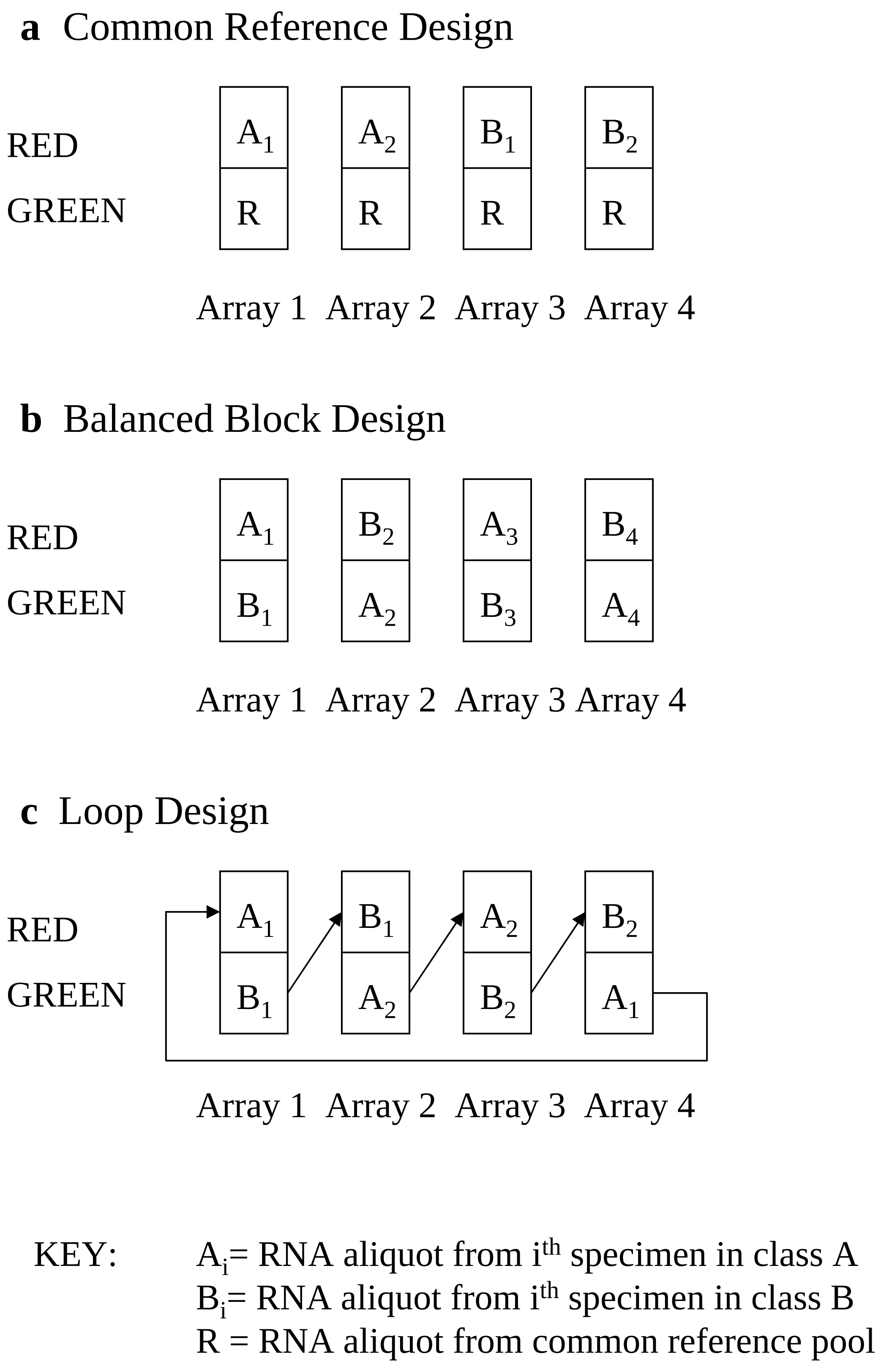
Figure 3. Power to detect a range of effect sizes (δ/σ) when combining 2 independent specimens per pool, where σ = SD of $\log_2$ expression measurements for the gene within each group, and δ = true difference in mean $\log_2$ expression level between the two groups. For all calculated values represented on the curves, significance level is set to 0.001, and the ratio of biological variance to technical variance is set to 4. The number of independent biological samples per group is *n*/2.

Figure 4. Dendrogram representing hierarchical cluster analysis of normalized $\log_2$ expression ratios for 1137 genes showing high variability across mammary tumor RNA

specimens from 36 mice representative of 6 different mouse models of cancer. The type of array used was a 2.7K mouse oncochip spotted cDNA microarray. The distance metric used was one minus the Pearson correlation. Linkage methods used were a) average linkage, b) complete linkage, and c) single linkage. Data are from Desai et al. (15).

Figure 5. Image plot representing hierarchical cluster analysis of normalized $\log_2$ expression ratios for 1137 genes showing high variability across mammary tumor RNA specimens from 36 mice representative of 6 different mouse models of cancer. The type of array used was a 2.7K mouse oncochip spotted cDNA microarray. The distance metric used was one minus the Pearson correlation, and average linkage was used. Each row in the image plot represents a single gene and each column represents a tumor specimen. Red and green bars indicate over-expressed and under-expressed genes in breast tumor compared to the pool of normal mouse mammary RNA, respectively. Black bars indicate genes with approximately equivalent expression levels, and gray bars indicate missing or filter-excluded data. The dendrogram displayed at the top is identical to the one in figure 4a. Data are from Desai et al. (15).


Gene expression
Strain 1
Agent 2
Agent 1
Drugs: 1 2 3
Strain 2
Agent 2
Agent 1
Drugs: 1 2 3

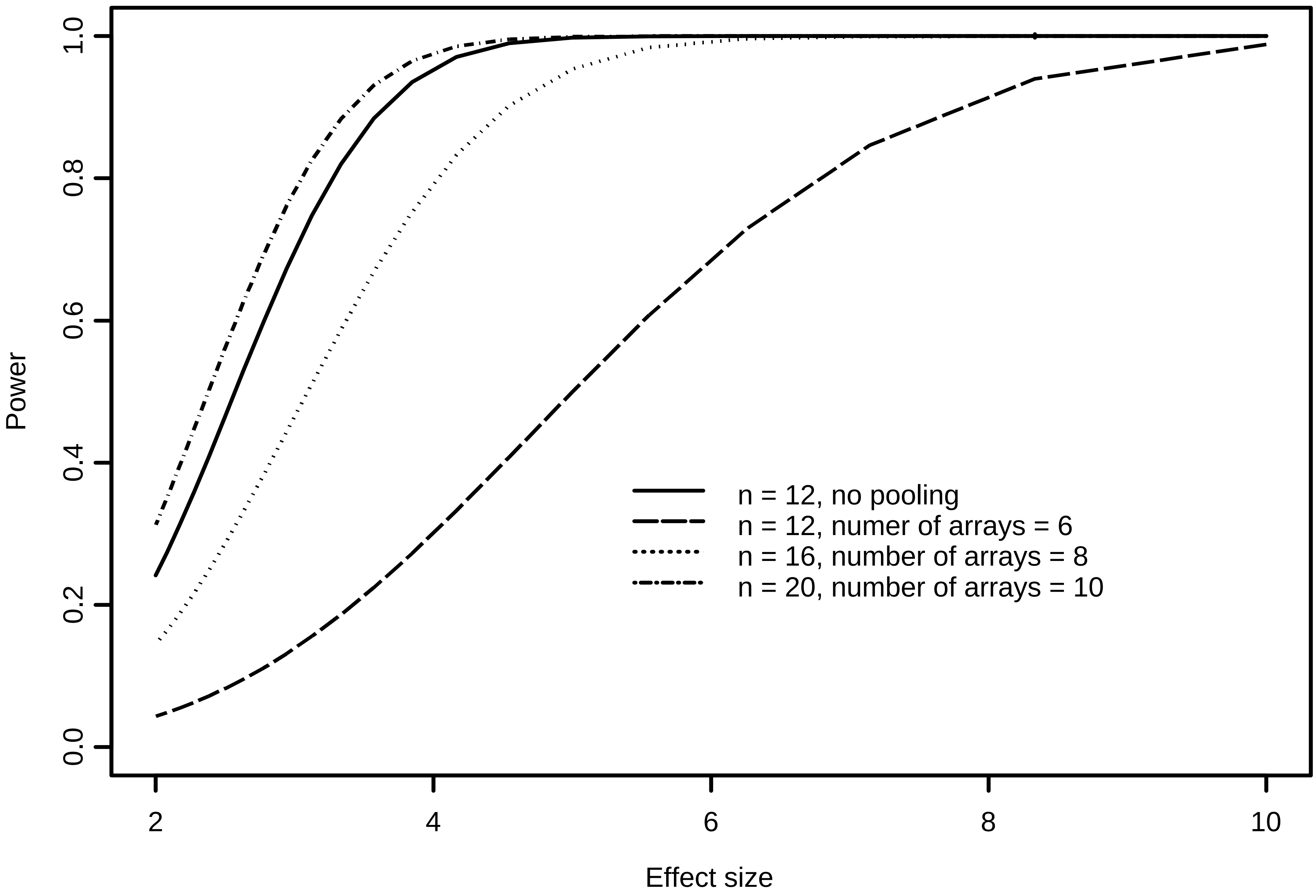


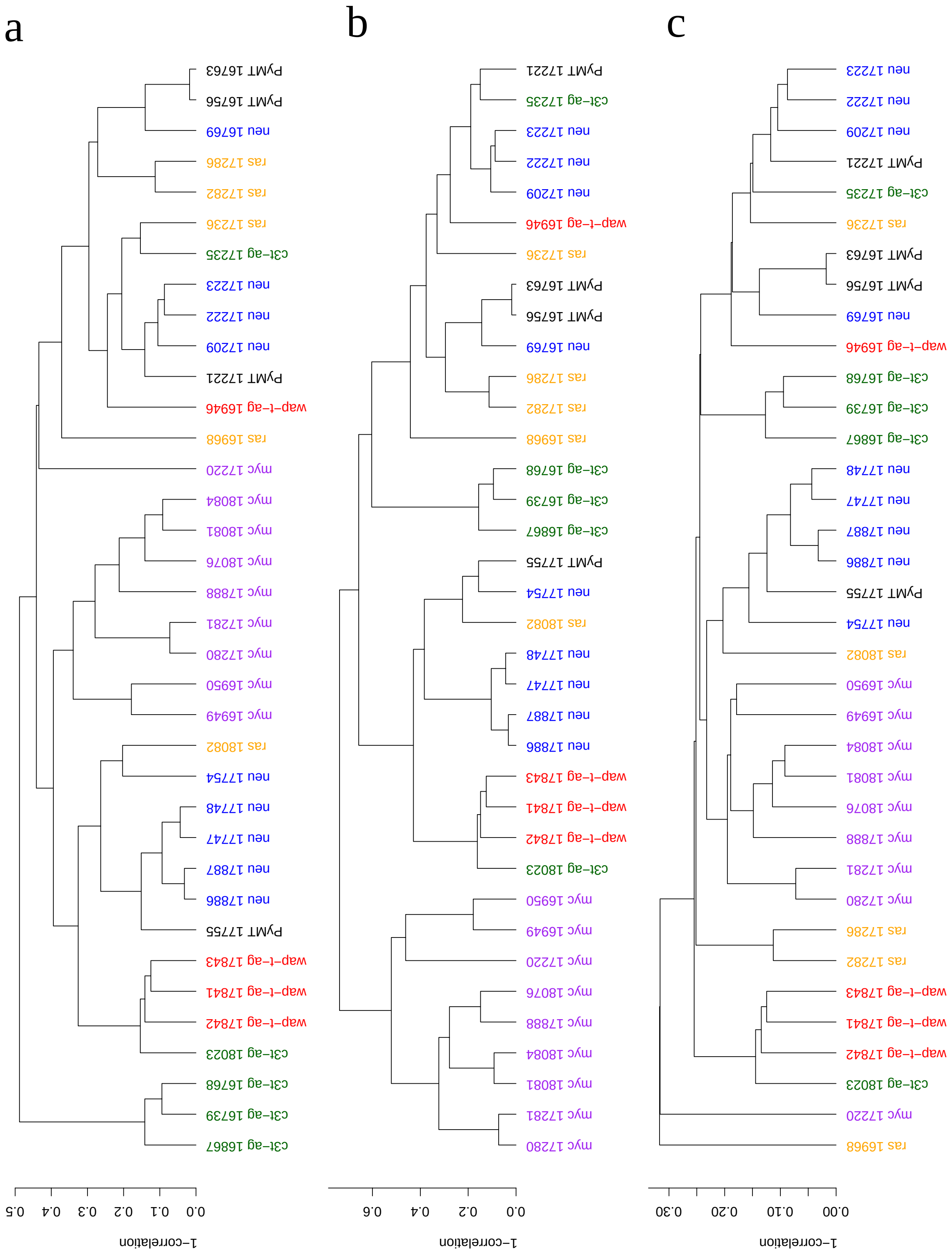
a Common Reference Design

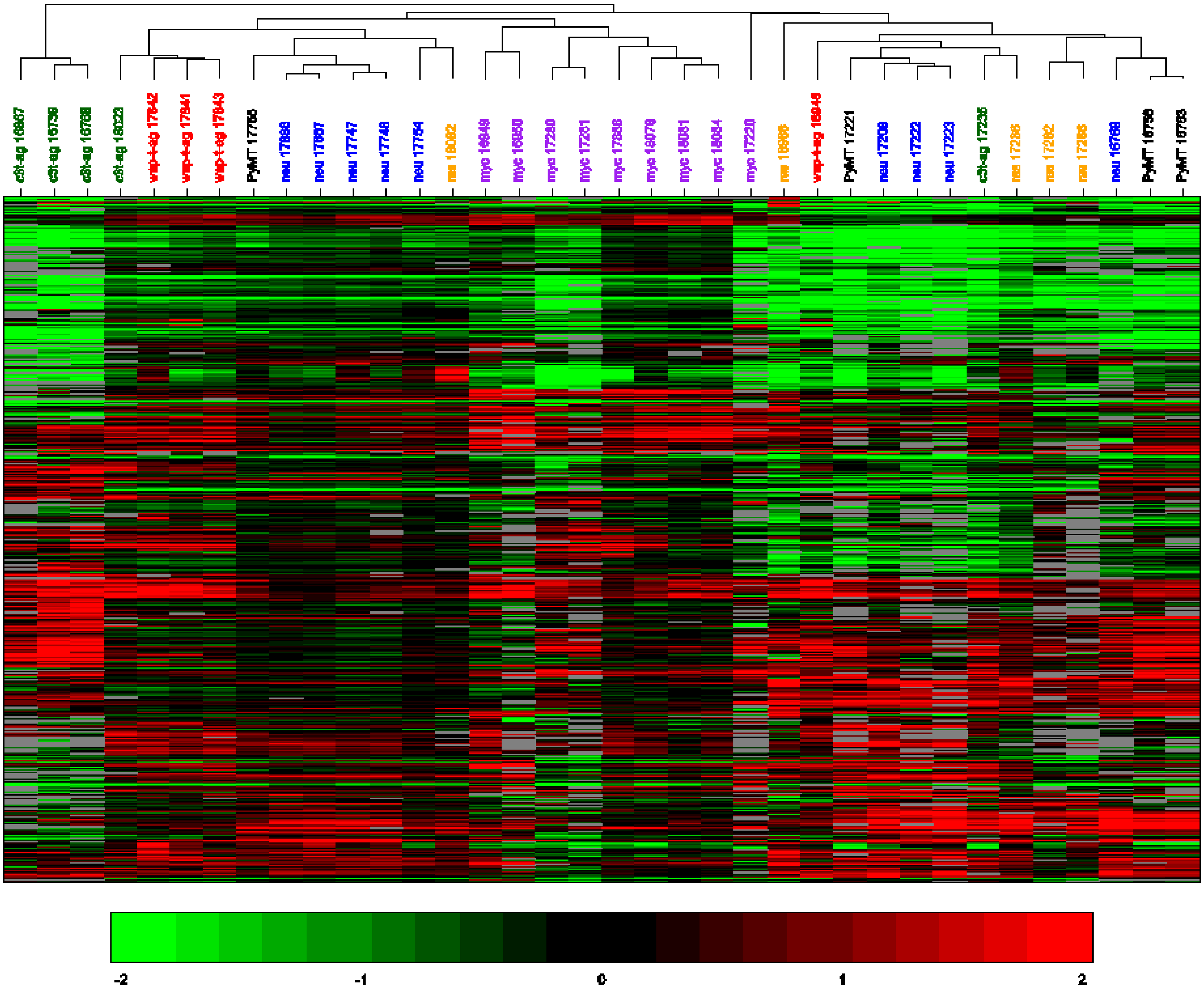
b Balanced Block Design

c Loop Design

KEY: A_i= RNA aliquot from i^{th} specimen in class A
B_i= RNA aliquot from i^{th} specimen in class B
R = RNA aliquot from common reference pool


Power
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
Effect size
n = 12, no pooling
n = 12, numer of arrays = 6
n = 16, number of arrays = 8
n = 20, number of arrays = 10


a
PyMT 16763
PyMT 16756
neu 16769
ras 17286
ras 17282
ras 17236
c3t−ag 17235
neu 17223
neu 17222
neu 17209
PyMT 17221
wap−t−ag 16946
ras 16968
myc 17220
myc 18084
myc 18081
myc 18076
myc 17888
myc 17281
myc 17280
myc 16950
myc 16949
ras 18082
neu 17754
neu 17748
neu 17747
neu 17887
neu 17886
PyMT 17755
wap−t−ag 17843
wap−t−ag 17841
wap−t−ag 17842
c3t−ag 18023
c3t−ag 16768
c3t−ag 16739
c3t−ag 16887
0.5
0.4
0.3
0.2
0.1
0.0
1−correlation
b
PyMT 17221
c3t−ag 17235
neu 17223
neu 17222
neu 17209
wap−t−ag 16946
ras 17236
PyMT 16763
PyMT 16756
neu 16769
ras 17286
ras 17282
ras 16968
c3t−ag 16768
c3t−ag 16739
c3t−ag 16887
PyMT 17755
neu 17754
ras 18082
neu 17748
neu 17747
neu 17887
neu 17886
wap−t−ag 17843
wap−t−ag 17841
wap−t−ag 17842
c3t−ag 18023
myc 16950
myc 16949
myc 17220
myc 18076
myc 17888
myc 18084
myc 18081
myc 17281
myc 17280
0.6
0.4
0.2
0.0
1−correlation
c
neu 17223
neu 17222
neu 17209
PyMT 17221
c3t−ag 17235
ras 17236
PyMT 16763
PyMT 16756
neu 16769
wap−t−ag 16946
c3t−ag 16768
c3t−ag 16739
c3t−ag 16887
neu 17748
neu 17747
neu 17887
neu 17886
PyMT 17755
neu 17754
ras 18082
myc 16950
myc 16949
myc 18084
myc 18081
myc 18076
myc 17888
myc 17281
myc 17280
ras 17286
ras 17282
wap−t−ag 17843
wap−t−ag 17841
wap−t−ag 17842
c3t−ag 18023
myc 17220
ras 16968
0.30
0.20
0.10
0.00
1−correlation



c3t-ag 16867
c3t-ag 16736
c3t-ag 16768
c3t-ag 18023
wap-t-ag 17842
wap-t-ag 17841
wap-t-ag 17843
PyMT 17765
neu 17888
neu 17697
neu 17747
neu 17748
neu 17754
ras 18082
myc 16849
myc 16850
myc 17280
myc 17281
myc 17936
myc 18078
myc 18081
myc 18084
myc 17220
ras 16806
wap-t-ag 16849
PyMT 17221
neu 17208
neu 17222
neu 17223
c3t-ag 17235
ras 17286
ras 17282
ras 17285
neu 16768
PyMT 16736
PyMT 16783
-2
-1
0
1
2