# Design and Analysis of Microarray Experiments for Pharmacogenomics

Hsu, Jason C.

*The Ohio State University*

Rao, Youlan

*The Ohio State University*

Lee, Yoonkyung

*The Ohio State University*

Chang, Jane

*Bowling Green State University*

Bergsteinsdottir, Kristin

*Univeristy of Iceland*

Magnsson, Magnus Karl

*Landspitali-University Hospital, Iceland*

Wang, Tao

*Pfizer Inc.*

Steingrmsson, Eirikur

*University of Iceland*

Chapter 7 in *Multiple Testing Problems in Pharmaceutical Statistics* (2009), Alex Dmitrienko, Ajit Tamhane, Frank Bretz editors.

Pharmacogenomics is the co-development of a drug that targets a subgroup of patients and a device that predicts whether a patient is in the subgroup of responders to the drug. Such a development involves a training study, followed by a validation study if warranted. This chapter discusses the design of pharmacogenomic studies based on established statistical principles and describes the analysis of data collected in these studies in a way that takes the multitude of multiplicity issues into account. Both aspects are critical to the success of pharmacogenomic development. A proof of concept experiment is used to show how proper design and analysis can smooth the path from discovery to clinical use.

# 1 Potential uses of biomarkers

A biological marker (a *biomarker* for short) is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Biomarkers Definitions Working Group 2001, FDA 2005b).

Biomarkers can be based on a variety of technologies, and have different uses. Our discussion will be for biomarkers based on microarray technology.

A DNA microarray is a chip with an array of microscopic spots of DNA sequences. They are used to measure relative abundance of nucleic acid sequences in samples. This is done by hybridizing fluorophore-labeled cDNA or cRNA samples to the microarrays, and then measuring the relative intensity of fluorescence emission at each spot.

One use of biomarkers is to determine whether a patient can benefit from a drug. P450s are the major enzymes involved in drug metabolism. AmpliChip CYP450, for example, uses microarray technology to test for variations in the genes that code two of the enzymes (CYP2D6 and CYP2C19) in the P450 family, to see if a patient will have difficulty in metabolizing certain prescription drugs. This is an example of using biomarkers known to measure an important aspect of the biological process, using microarrays as a device.

Another use of biomarker is to predict disease progression. MammaPrint, for example, predicts the likelihood of breast caner recurrence of a patient using the expression levels of 70 genes. As reported in van't Veer et al (2002), this 70-gene composite biomarker was developed by measuring the expression levels of approximately 24,000 genes on 78 patients using microarray technology. The 70 genes were selected using multiple testing and machine learning techniques based on error rate control, sensitivity, and specificity considerations. Whether these genes are biologically involved in the disease process, or how they affect disease progression, was not part of the study. MammaPrint itself uses multiplex microarrays

that test eight patient samples per array, probing the 70 genes three times per sample. This is an example of developing and using a biomarker based on microarray technology to predict disease progression.

Note that MammaPrint's disease prognosis is given with no reference to any particular treatment. In the next section, we explain how pharmacogenomics can go beyond mere disease prognosis, coupling a drug with a device predicting which patients will respond to the drug.

# 2 Clinical uses of genetic profiling

Microarrays for clinical use as medical devices are subject to the regulation of the Center for Device and Radiologic Health (CDRH) of the U.S. Food and Drug Administration (FDA). In 2003, the FDA issued its guidance for disease prognostics based on multiplex testing of heritable markers (FDA, 2003). In 2007, Agendia's MammaPrint was approved by CDRH to be marketed as a medical device for breast caner prognostics. Clinical use of microarray technology is thus a reality. This technology also has potential in drug development, making more efficacious compounds available to patients, with less side effects.

It is well known that most drug development programs fail. However, even if a compound development does not succeed for the entire patient population, there is the possibility that it may still benefit a subgroup of the patient population. In terms of *efficacy*, even when a pharmaceutical trial fails to show sufficient efficacy averaged over the entire patient population, there is the potential that the compound is efficacious for a subgroup of the patient population. In terms of *safety*, even if a compound causes serious adverse events (SAEs) in some patients, there is the possibility that SAEs are confined to a subgroup of the patients. These are the rationales for *pharmacogenomics*.

Analysis of efficacy and safety for certain subgroups based on criteria such as sex and ethnicity is already routinely done. Technology such as microarrays makes it possible to form such subgroups based on the genetic profiles of patients or tissue samples, allowing more refined subgroup analysis in principle.

Pharmacogenomics thus goes beyond mere disease prognosis in that it couples a drug with a device implementing an algorithm predicting which patients will respond to the drug. It is the co-development of a *drug* that targets a subgroup of the patient population, as well as a *device* that can be used to predict whether a patient is in this subgroup of responders to the drug. Since both *drug* and *device* are involved, pharmacogenomics is subject to the joint approval by the Center for Drug Evaluation and Research (CDER) and by the Center

for Device and Radiologic Health (CDRH) of the FDA.

In 2005, the FDA issued its Voluntary Genomic Data Submission (VGDS) guidance and drug-diagnostics co-development concept paper (FDA, 2005b; FDA, 2005a). With the issuance of these documents, pharmaceutical companies have started banking blood and tissue samples from clinical trials (based on informed consent) for potential pharmacogenomic use.

After highlighting the key statistical issues in these documents in the next section, the rest of the chapter is devoted to a discussion of multiplicity issues in pharmacogenomics.

# 3 Two stages of pharmacogenomic development

Pharmacogenomic development is a two-stage process. The first stage is to identify a biomarker positive ($G^+$) subgroup of patients for which the compound is extra efficacious, compared to patients in its complement, the biomarker negative ($G^-$) subgroup.

In clinical trials for drug development, efficacy can be defined in terms of higher *average improvement* (over the control group), or in terms of higher *responder rate* (over the control group):

- In Alzheimer's disease trials, efficacy is typically established by comparing mean changes from baseline between treated and control groups.
- In schizophrenia trials, efficacy might be established by comparing mean changes from baseline of the Positive and Negative Syndrome Scale (PANSS) score between treated and control groups. Or it might be established by comparing responder rates between treated and control groups. A responder for a schizophrenia drug might be one who experiences at least a 30% reduction in total PANSS score from baseline.
- For hypertensive drugs, a responder might be one whose systolic blood pressure has been reduced to no more than 120 mm Hg.
- For diabetes drugs, a responder might be one whose HbA1c (glycosylated hemoglobin) is less than 7%.

One might look for a $G^+$ subgroup based on genotypes already suspected to affect disease outcome. In Alzheimer's disease, for example, one might compare efficacy between carriers and non-carriers of the ApoE "4" allele. Alternatively, one might attemtp to discover a $G^+$ subgroup by comparing the genetic profiles of the responder ($R^+$) patients with the profiles

of the non-responder ($R^-$) patients. Using the banked biological samples, measurements on typically a large number of biomarkers are obtained. These marker measurements may be SNP categories obtained from blood samples, or gene expression levels measured from tissue samples, for example. Biomarkers that show substantial differences between the $R^+$ and $R^-$ groups are selected, and based on a combination of these selected biomarkers, which might be called a *composite* biomarker or a gene signature, a prognostic classification algorithm is constructed to predict whether a future patient will be a responder or a non-responder to the compound.

At the end of this first stage, the *sensitivity* of the prognostic algorithm, which is the probability that a patient will be a responder given that the patient is biomarker positive ($G^+$), and the *specificity* of the algorithm which is the probability that a patient will be a non-responder given that the patient is biomarker negative ($G^-$), should be estimated. Provided that both the estimated sensitivity and specificity are sufficiently high, pharmacogenomic development proceeds. Otherwise, further pharmacogenomic development is likely to be futile.

In Section 4, we discuss multiple testing for differential expressions and for significant composite biomarkers.

If pharmacogenomic development proceeds to the second stage, a new clinical trial is conducted to independently validate the efficacy and safety of the compound for the target subgroup, and to prove that the composite biomarker has sufficient sensitivity and specificity for clinical use.

One issue that is often overlooked is the process of developing a prognostic device based on gene expressions involves a *change of platform* between the training stage and the validation (and eventual clinical use) stage. As stated in FDA (2005b):

> A new test with fewer biomarkers developed for diagnostic purposes (i.e., patient stratification) should be properly validated, ideally in clinical trials that enrolled patients with the intended indication.

Whereas the training study might use microarrays probing many biomarkers, the validation study uses the prognostic chip containing only the genes in the signature that is intended for eventual clinical use. (For example, the validation study of MammaPrint used microarrays probing 70 genes only.)

Another issue that has not been fully addressed is how to design and analyze the training study in order to properly design the validation study. As stated in FDA (2005b), it is important for a pharmacogenomics development plan to be able to compute sample sizes required to meet validation requirements:

> When validating a gene or expression pattern, instead of a set of individual biomarkers, a rigorous statistical approach should be used to determine the number of samples, and the methodology used for validation. It is recommended that the validation strategy be discussed in advance with FDA.

In this chapter, we discuss how to statistically design and analyze the training study in order to properly design the validation study. Determination of sample sizes for the validation study will be elucidated in Rao, Lee and Hsu (2009).

# 4 Multiplicity in pharmacogenomics

Two of the sources of multiplicity in pharmacogenomics are multiplicity of individual *biomarkers* and multiplicity of *subgroups*. Individual biomarkers (genes in our discussion) can be selected to form a composite biomarker. Subgroups are defined by the multitude of prediction algorithms that can be formed by the selected individual biomarkers.

To select biomarkers to form a composite biomarker, one can test for the significance of the individual biomarkers controlling a multiple error rate such as the Familywise Error Rate (FWER), generalized Familywise Error Rate (gFWER) and False Discovery Rate (FDR). These error rate are defined in Chapter 2 section 2.

Then, after proposing a prognostic algorithm based on the selected biomarkers, one can validate its discriminant power as follows. Each potential prediction algorithm divides the patients into either the biomarker positive ($G^+$) group or the biomarker negative ($G^-$) group. A patient in the $G^+$ group is predicted to be a responder, while a patient in the $G^-$ group is predicted to be a non-responder. Therefore, one can account for the multiplicity of subgroups by proving that the proposed prognostic algorithm has non-zero discriminant power, even after taking into account the multiplicity of potential prediction algorithms that can be formed by the selected biomarkers.

## 4.1 Multiplicity of genes

In the first stage, genes are tested for differential expressions, as having too many genes may hinder construction of an effective classification algorithm and inflate its variability. At this stage, expression levels are measured using microarrays that probe a large number of genes, perhaps the whole genome. For example, the first stage in the training of MammaPrint used microarrays that probe approximately 24,000 genes. With such a large number of genes, if they are tested without adjusting for multiplicity, surely some will be found to be

differentially expressed even if genetic makeup has absolutely no bearing on the response. This is the first multiplicity issue in pharmacogenomics. To confidently select genes to train a classification algorithm within the first stage, we control an appropriate multiple testing error rate, in agreement with FDA (2005b):

> Statistical considerations in deriving a small number of biomarkers from a large amount of parallel multiplexed data should be properly addressed.

## Two different definitions of Type I error rate

Suppose $k$ genes are probed in comparing expression profiles between responder and non-responder groups. Let $\mu_{Ri}, \mu_{Ni}, i = 1, \ldots, k$, denote the expected (logarithms of) expression level of the $i$th gene of a randomly sampled patient from the responder and non-responder group respectively. Let $\theta_i$ denote the difference of the expected (logarithms of) expression levels of the $i$th gene between the two groups, $\theta_i = \mu_{Ri} - \mu_{Ni}$.

In the current bioinformatics literature, the (marginal) null hypotheses being tested are

$$H_{0i} : \theta_i = 0 \tag{1}$$

for $i = 1, \ldots, k$. Multiple testing then generally involves testing $H_{0I} : \theta_i = 0$ for all $i \in I$ for $I \subseteq \{1, \ldots, k\}$.

A Type I error of testing $H_{0I}$ can be defined two different ways. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, and let $\boldsymbol{\Sigma}$ denote generically all parameters that the observed expression levels depend on (including variances, covariances, skewness, etc.). Let $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_k^0)$, and let $\boldsymbol{\Sigma}^0$ be the collection of all (unknown) true parameter values.

The traditional definition of the Type I error rate (Casella and Berger, 1990, Definition 8.3.3), is

$$\sup_{\theta_i = 0, i \in I} P_{\{}\text{Reject } H_{0I} | \boldsymbol{\theta}, \boldsymbol{\Sigma}\}, \tag{2}$$

where the supremum is taken over all possible $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ subject to $\theta_i = 0, i \in I$. In the analysis of gene expression levels, the supremum is taken over all possible joint distributions of expression levels under the null hypothesis (including all possible correlations and expression levels of the other genes with indices not in $I$). The more realistic null hypotheses (which might become more popular as bioinformatics evolves) is defined as

$$H_{0i} : \theta_i \leq \delta$$

and

$$H_{0i} : \theta_i \geq -\delta$$

with $\delta > 0$ the Type I error rate of a test for $H_{0I}$ may well depend on the values of $\theta_j, j \notin I$, in the presence of gene by gene interactions.

A different definition of the Type I error rate for testing $H_{0I}$, given in Pollard and van der Laan (2005) for example, is

$$P\{\text{Reject } H_{0I} | \boldsymbol{\theta}^0, \boldsymbol{\Sigma}^0\}. \tag{3}$$

Note that $\theta_i^0 = 0$ for $i \in I$ under $H_{0I}$, $\theta_j^0, j \notin I$, and $\boldsymbol{\Sigma}^0$ are unknown. Thus this probability is difficult to compute directly. Pollard and van der Laan (2005)'s methods, implemented as the MTP function in the multtest package in the bioconductor repository of R, estimate $\theta_j^0, j \notin I$, and $\boldsymbol{\Sigma}^0$ by resampling. Their methods control the Type I error rate *asymptotically* as the number of microarrays goes to infinity.

In practice, statistical decision procedures are applied to *different* studies over time. Therefore, error rate control is more useful if it controls the long run relative frequency of incorrect decisions across different studies. Such a claim is possible provided each test controls the Type I error rate at level $\alpha$ according to the traditional definition (2). See Berger and Wolpert (1988, pp. 71-2) and Berger (1985, p. 23, p. 44). In this sense, perhaps definition (2) is more useful than definition (3).

## Popular error rates to control in bioinformatics

Let $V$ denote the number of false rejections and let $R$ denote the total number of rejections. The two most popular quantities to control in bioinformatics are the *false discovery number* $V$ and *false discovery proportion* $V/R$. The rates of these errors are reported either as an *exceedance probability* (relative frequency of $V$ or $V/R$ exceeding a specification across many multiple tests), or an *expectation* (average of $V$ or $V/R$ across many multiple tests). Such reporting can either be *unconditional*, averaged over all studies, or *conditional* on the data, restricted to studies with an observed number of total rejections $R = r$, for example.

The FWER and gFWER are exceedance probabilities. The FDR is an expectation error rate. The step-up test of Benjamini and Hochberg (1995) is a popular FDR-controlling method.

Another popular method in bioinformatics is the Significance Analysis of Microarrays (SAM) proposed by Tusher, Tibshirani, and Chu (2001). Instead of controlling the FDR, it aims to control $E(V)/R$, under the configuration that all the null hypotheses are true.

Reporting of an error rate which is an *expectation* may be inadequate, if the quantity being controlled (or a component of it) turns out to be highly variable for a statistical

method. Gordon et al (2007) showed that, in terms of $V$, the number of false discoveries, the Benjamini and Hochberg (1995) method is less stable than the Bonferroni method.

If controlling $V$ guards against incorrect decision-making, and Type I error rate is controlled according to the traditional definition (2), then FWER or gFWER control of the exceedance probabilities of $V$ implies control of the long run relative frequency of incorrect decision across different studies, as discussed at the end of the previous chapter.

## Partitioning to control the gFWER

One approach to controlling the gFWER is to use a method that controls the familywise error rate (i.e., controlling the gFWER at $m = 0$), and then augments the rejections by automatically rejecting the null hypotheses associated with the next $m$ extreme test statistics. This is the *augmentation* approach of van der Laan, Dudoit, and Pollard (2004).

A different approach is to use the generalized partitioning principle of Xu and Hsu (2007). It partitions the parameter space into disjoint subspaces $\Theta_I^*$ just like the partitioning principle introduced in Chapter 2 Section 3.4, but adds the concept of testing *individual* hypotheses $H_{0i}$. Specifically, in each $\Theta_I^*$, it rejects all $H_{0i}$, $i \notin I$, and test $\{H_{0i} : \theta_i \in \Theta_i, i \in I\}$ at gFWER level $\alpha$. An $H_i$ is then rejected if it is rejected in $\Theta_I^*$ for all $I \ni i$.

Note that the original partitioning principle can be viewed as a special case of the generalized partitioning principle in that, when testing in $\Theta_I$, the hypotheses $H_{0i}, i \in I$, are either all accepted or rejected.

A particular application of the Generalized Partitioning Principle is to use Markov's inequality to provide a gFWER-controlling test for each $\Theta_I^*$. Suppose the level of each marginal test is $\alpha$. Then, in testing $|I|$ true null hypotheses, Markov's inequality states

$$P(V > m) \leq \frac{|I|\alpha}{m+1}.$$

Thus, one can control gFWER at level $\alpha$ when testing in $\Theta_I^*$ by testing each individual hypothesis at level $\alpha(m+1)/|I|$. The resulting multiple test is the gFWER-controlling method of Hommel and Hoffmann (1988), which was re-discovered by Lehmann and Romano (2005).

## Conditional versus unconditional inference

FDR is an unconditional error rate. It may be tempting for investigators to report conditional error rates in practice. Suppose that an investigator tested 100,000 hypotheses, using a method which controls FDR unconditionally at 1%. Then, having rejected 1,000 hypotheses,

the investigator might want to state "10 out of these 1000 discoveries are expected to be false discoveries." Such a statement, conditional on the realized rejections, is misleading since FDR is an unconditional expectation.

Step-down procedures that control the FWER and gFWER adjust for multiplicity conservatively, but only to the extent that some subset (but not all) of the null hypotheses $H_{0i}, i = 1, \ldots, k$, might be true, conditional on how many of them have been rejected. This is in contrast to a single-step procedure which typically adjusts for multiplicity under the scenario that *all* the null hypotheses are true. The critical value (threshold) used by a step-down procedure, in effect, is the one corresponding to the maximum subset hypothesis $H_{0I}$ that could be true, conditional on data. Such conditional tests keep the true error rate as close to the desired error rate as possible, while still guaranteeing conservatism.

A form of conditional FDR error rate reporting, discussed in Efron (2007) for example, is to report an estimate of $E(V)/r$ where $r$ is the realized number of rejections, $R = r$. Note that Efron's method estimates $E(V)$ unconditionally.

## Taking dependence into account

Methods based on the Bonferroni inequality for FWER control (e.g., Holm's method), or Markov's inequality for gFWER control (e.g., the method in Lehmann and Romano (2005), do not take joint distribution of the test statistics into account. They are generally conservative.

Some methods, such as Hochberg's step-up method for FWER control, and Benjamini and Hochberg's (1995) step-up method for FDR control, set critical values based on the assumption that the test statistics are independent. They are conservative under certain positive dependence structures.

For FWER control, if the test statistics have a multivariate normal or a multivariate $t$ distributions under an intersection/partitioning hypothesis $H_{0I}$, and the correlation structure has exactly or approximately a one-factor structure, then the *factor analytic* technique of Hsu (1992) is applicable. This technique amounts to modeling dependence by a latent variable. Alternatively, the variance-reduced Monte Carlo technique of Genz and Bretz (1999) can be applied. If the joint distribution of the test statistics is not multivariate normal or multivariate $t$, then resampling techniques can be used to compute the thresholds.

For gFWER control, assuming the test statistics have an exchangeable distribution under an intersection/partitioning hypothesis, Xu and Hsu (2007) constructed step-down methods that control gFWER while taking dependence among the test statistics into account. In analogy to closed/partitioning procedures based on the maximum test statistic, they pro-

posed using an order statistic to test each intersection/partitioning hypothesis $H_{0I}$. The technique for computing threshold proposed by Xu and Hsu (2007) is a special case of the factor-analytic technique, as equal correlation can clearly be generated by a single latent variable. Romano and Wolfe (2007) also proposed gFWER-controlling methods based on order statistics, using a resampling technique to compute the thresholds.

For Fdr reporting when the number of tests is large, Efron (2007) first transforms the test statistics so that those corresponding to the true null hypotheses become standard normal random variables. Then, to estimate $E(V)$, assuming pairs of transformed test statistics are bivariate normal, he extracts a one factor structure from the *multinomial* distribution of binned counts from ordered transformed test statistics.

In summary, general strategies for taking dependence into account are

- Model dependence as arising from a latent variable,
- Estimate dependence by resampling.

## 4.2 Multiplicity of subgroups

In searching for a subgroup of the patients for which a compound is especially efficacious, one must guard against the possibilities that the more subgroups are searched the more likely that one will "discover" such a subgroup by chance.

Let $\boldsymbol{X} = (X_1, \ldots, X_k)'$ represent measurements on the individual biomarkers from a typical patient. One might contemplate using a linear combination of the biomarker measurements $\boldsymbol{b}'\boldsymbol{X} = b_1X_1 + \cdots + b_kX_k$ to place patients in biomarker positive ($G^+$) and biomarker negative ($G^-$) groups. For example, $\boldsymbol{b}'\boldsymbol{X} > c$ puts the patient in the $G^+$ group, while $\boldsymbol{b}'\boldsymbol{X} \leq c$ puts the patient in the $G^-$ group.

Let $Y$ be an appropriate measure of efficacy of the compound. To see the danger of subgroup analysis without appropriate multiplicity adjustment, consider the model $Y = \beta_1X_1 + \cdots + \beta_kX_k + \epsilon$, where $\epsilon$ represents uncertainty and assume that $\beta_i \neq 0$ only if the $i$th biomarker correlates with efficacy.

Suppose none of the biomarkers correlates with efficacy (i.e., $\beta_i = 0, i = 1, \ldots, k$). Then, in testing whether individual biomarkers correlate with efficacy, $H_{0i} : \beta_i = 0$, the multiplicity of having $k$ biomarkers needs to be taken into account to avoid false positives. In testing whether any composite biomarkers in the family

$$\{H_{0\boldsymbol{b}} : b_1\beta_1 + \cdots + b_k\beta_k = 0, \boldsymbol{b} \in \Re^k\}$$

correlate with efficacy, if we assume uncertainty terms $\epsilon$ are i.i.d. with a normal distribution with a known variance for simplicity, then the appropriate multiplicity adjustment to control FWER can be made by using a threshold based on a Chi-square distribution, not a normal distribution.

Figure 1 shows how quickly the probability of incorrectly inferring at least one composite biomarker as correlating with efficacy approaches 100% as $k$ increases, if each composite biomarker is tested at an error rate of $\alpha = 0.05$.

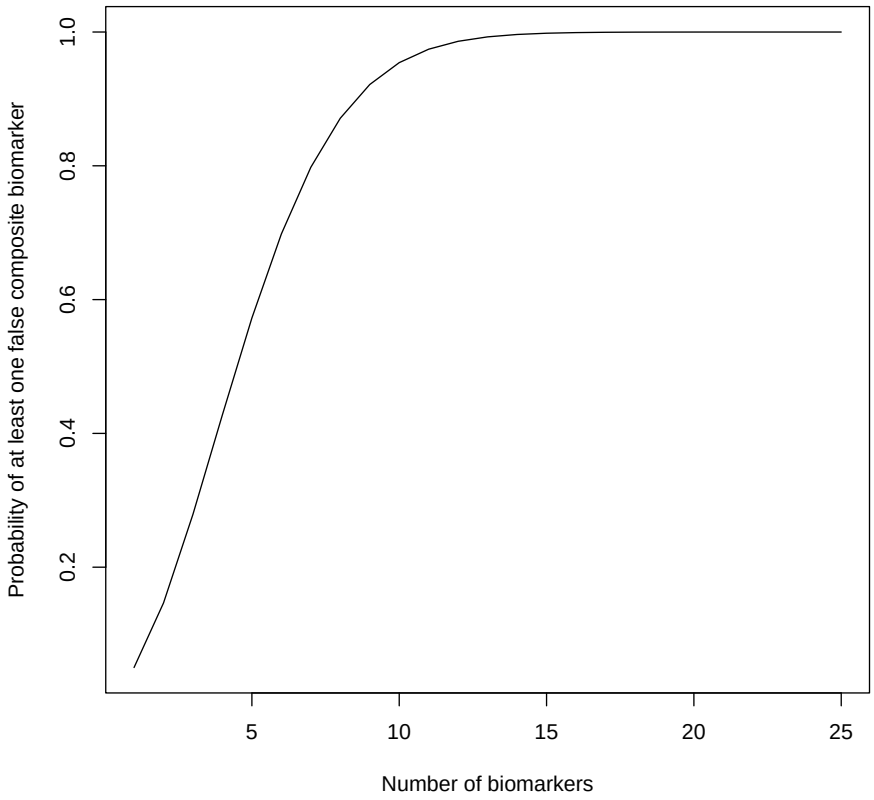


Figure 1: The probability of falsely finding a composite biomarker significant without multiplicity adjustment

Even if a (composite) biomarker is found to be correlated with efficacy, it does not mean that its discriminatory power would be high enough to be practically useful. Whether a compound is sufficiently efficacious for the $G^+$ group or not requires a test on the accuracy (sensitivity or specificity) of the prediction rule based on the composite biomarker for the

responder/non-responder status. In general, composite biomarkers correlated with efficacy that define subgroups can be nonlinear in the biomarker measurements. Thus, the multiplicity of subgroups can be far greater than what has been considered in this section, and its proper adjustment would call for a different approach to the problem.

# 5 Designing pharmacogenomic studies

Microarray experiments conducted for eventual *clinical use* are, in essence, clinical trials *in silica*. As in clinical trials, designing microarray experiments according to the statistical principles below helps to ensure that the prognostic/diagnostic algorithm derived from expressions measured on the training platform applies to expressions measured on the validation platform and in its eventual clinical use.

## 5.1 Control for platform change using external reference sample

To control for possible systematic shifts in measured expression levels changing from the training platform to the validation platform, one can normalize expression levels between platforms using the expression levels measured from samples that are available and homogeneous from the training experiment to the validation experiment, and remain so for clinical use. An example of such samples is the Universal Reference Sample from StrataGene.

In the case of MammaPrint, the training study placed a reference sample pooled from 78 patients into one of the two channels on each of the microarrays, whereas the validation study placed a reference sample pooled from 307 patients into one of the two channels on every microarray. When the training study and validation study utilize different reference samples, it is unclear to us which is the appropriate reference sample in clinical use for individual patients.

## 5.2 Design to discover group differences

Measurements on gene expression levels inherently contain variability. The five sources of variability of measured gene expression levels are as follows.

1. Group: There may be differential gene expressions between risk groups (averaged over infinitely many subjects).

2. Subject: Within each group, subjects may have the same alleles but still have natural differences in expression levels (even for inbred mice).

3. Sample: Different samples from the same subject, so-called technical replicates, nevertheless will have some difference in expression levels.

4. Probe: The probes for each gene represent different parts of a gene, and will have different amounts of RNA hybridized to them.

5. Noise: Noise could come from various non-biological sources such as experimental and technical settings which may not be identical in repeated experiments.

Having replicate measurements from each level of each factor allows one to estimate each effect, and their variabilities. One can then, in turn, not only discover group differences more readily, but also estimate sensitivity and specificity of prediction algorithms, as follows.

## Replicate to estimate and remove variabilities

Figure 2 displays observed expression levels of 99 genes from five groups of mice from our proof of concept experiment described in Section 7.7 (after background correction and normalization). Separation of the groups cannot be seen. The reason turns out to be that mouse and sample variabilities overwhelm group differences.

In order to estimate the effect of each subject, and its variability, replicate samples from each subject is needed. In order to estimate the effect of each sample, and its variability, each sample needs to be probed multiple times.

To discover group differences, if subject and sample effects can be estimated unbiasedly, then removing them may make group differences reveal themselves more readily. By treating subject and sample as fixed effects (as one would adjusting for covariate effects), one may more readily identify differentially expressed genes. Multiple tests conducted in this fashion control error rates conditionally, conditional on the subjects and the samples. Therefore, they control error rates unconditionally as well.

In analyzing expression levels to discover genes differentially expressed between groups, expression level is the response variable while group, subject, sample, and probe all are predictor variables. However, in training a classification algorithm based on differentially expressed genes, their expression levels then become predictors for treatment outcome. Sensitivity and specificity of such an algorithm depends on how variable expressions are between subjects within each group, and between samples within each subject. With replicate samples from each subject, and replicate probes for each sample, the variability of subject and sample can be estimated by considering them as *random* effects in modeling expression level data. Analysis of gene expressions should be cognizant of this distinction between the roles

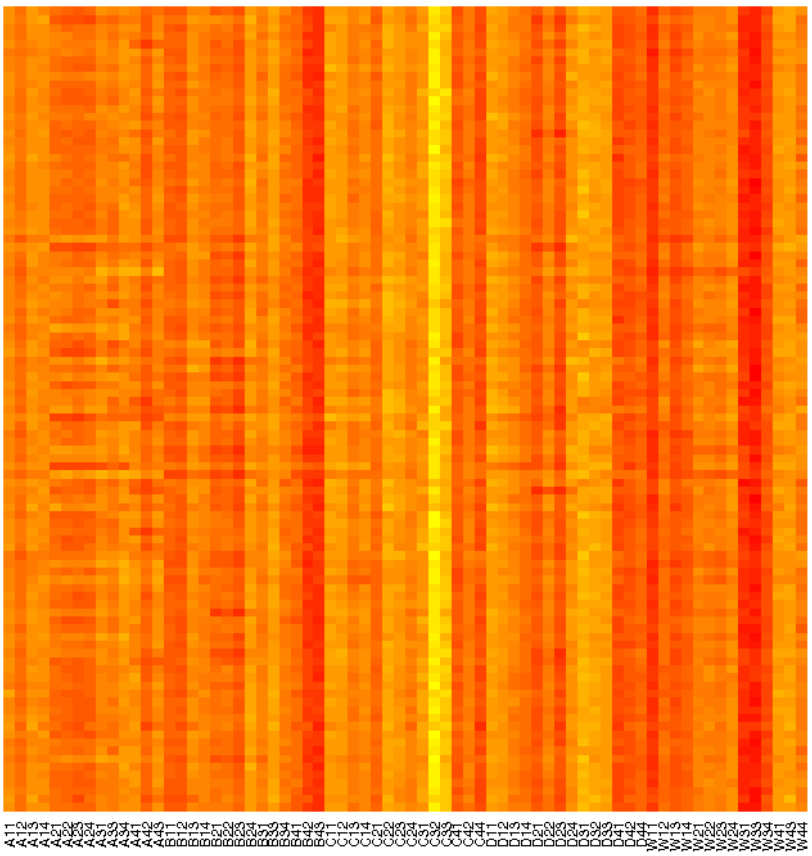


Figure 2: Observed expression levels of genes from inbred mice. Rows, from top to bottom, represent gene 1 to gene 99. Columns correspond to samples from five groups of mice, arranged from left to right: group A, B, C, D, and W (wildtype). There are four mice per group, and four samples per mouse. The 80 columns, from left to right, represent sample 1 of mouse 1 from group A, to sample 4 of mouse 4 from group W.

of expression levels. Figure 3 displays estimated expression levels of 99 genes, after estimated mouse and sample effects are removed by modeling them as fixed effects. Clustering of 80 samples with the estimated gene expression levels rediscovered the five groups of mice, arranging the columns corresponding to the 80 samples perfectly into five distinct groups.

## Block to avoid confounding

Gene expression measurements from microarrays are potentially affected by extraneous effects such as array or batch processing effects. Many microarray experiments take only one

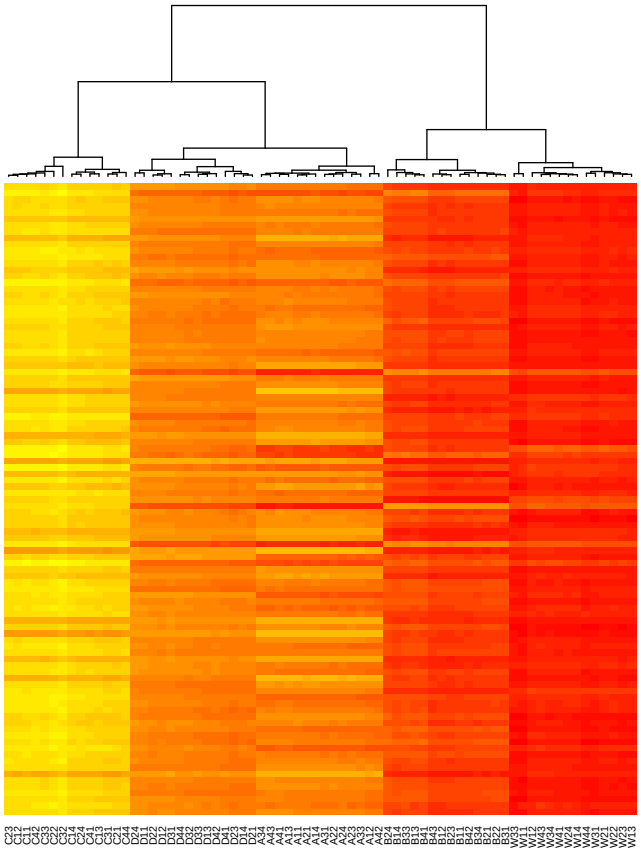

Figure 3: Estimated expression levels of genes from inbred mice, with Mouse and Sample effects removed. Rows, from top to bottom, represent gene 1 to gene 99. Arranging the columns by unsupervised machine learning (clustering) results in five distinct groups: from left to right, groups C, D, A, B, and W (wildtype).

sample from each subject, hybridizing it to one microarray. Estimates of group differences are then potentially confounded with array effects.

In microarray experiments, a statistical *block* is a condition under which measured gene expressions are likely to be equally affected by confounding factors. A block might, therefore, be a batch of arrays processed together. Some microarrays, including those made with Agilent and NimbleGen technology, can have multiple biological samples hybridized on the same array, so each array can conveniently form a block.

Statistical analysis of a block design proceeds by first comparing different risk groups within blocks, and then combining such comparisons across blocks. Such an analysis not only

avoids confounding due to extraneous effects, but also increases sensitivity and specificity because it eliminates batch/array variability in the comparisons. Basic statistical design principles also suggest keeping the proportions of samples from the groups to be compared the same across blocks. Doing so not only facilitates statistical analysis by making the effects orthogonal, but makes group comparisons more efficient as well.

### Randomize sample hybridization to avoid bias

Randomization prevents *bias*. Within each block, if placement of the samples onto microarrays or processing of the biological samples over time is not randomized, the observed differences in expression levels may contain biases due to extraneous effects. To avoid such biases, we randomize the placement of biological samples onto the microarrays and the order in which the samples are hybridized.

### Randomize probe placement to avoid change of platform issue

If placement of the probes on microarrays is not randomized, measurements from the training platform and the validation platform may have different biases. To avoid such biases in order to ensure a prediction algorithm derived from the training platform applies to expressions measured on the validation platform, we randomize the placement of probes on both platforms.

A good microarray experimental design balances the allocation of samples from the groups to be compared to wells/sub-microarrays on each microarray, and to microarrays within a batch, to avoid potential confounding array or batch effects.

## 5.3 Permutation tests may not control error rates

Permuting raw data across groups to be compared is often used as a tool to generate reference distributions under the null hypotheses to be tested. It is thought of as capable of taking dependence into account, as well as producing more exact inferences for small samples than methods based on asymptotics.

It turns out that, depending on application, there may be an assumption needed for permutation tests to control multiple testing error rates. At issue is, in comparing parameters of the marginal distributions of two sets of multivariate observations, the validity of permutation testing is affected by all the parameters in the joint distributions of the observations.

Calian, Li, and Hsu (2008) showed the surprising fact that, in the case of a linear model with i.i.d. errors, permuting raw data (instead of residuals) across groups to be compared turns out to control the FWER, if the test statistic for each partitioning hypothesis is based on ordinary least squares estimates and of the maximum test statistic form.

On the other hand, as shown in Xu and Hsu (2007), in comparing the mean expression levels of genes between two groups of subjects, permuting raw data across groups may not generate correct reference distributions under the null hypotheses, unless equalities of mean expression levels for a set of genes automatically imply equality of joint distributions of expressions levels for this set of genes (including equalities of variances, covariances, and higher cumulants).

Our view is, unless such an assumption can be made on biological ground, it is safer to take a modeling approach, an example of which is described below.

# 6 Analyzing microarray data by modeling

Suppose that a microarray experiment is conducted properly, adhering to the statistical design principles and addressing the issues on different sources of variability in gene expression measurements discussed so far. We can analyze the microarray data by proper modeling. A concrete example of such a design and experiment is to be given in the next section.

Let $y_{igmspr}$ denote the background corrected, log transformed and normalized probe intensity for the $i$th gene ($i = 1, ..., n_i$), $s$th sample ($s = 1, ..., n_{s(m,g)}$) from the $m$th subject ($m = 1, ..., n_{m(g)}$) in group $g$ ($g = 1, ..., n_g$), $p$th probe ($p = 1, ..., n_{p(i)}$), and $r$th replicate ($r = 1, ..., n_r$) from the experiment.

We assume, for each $i$, $y_{igmspr}$ follows a linear effects model

$$Y_{igmspr} = \mu_i + \tau_{ig} + M_{im(g)} + S_{is(m(g))} + \pi_{p(i)} + \epsilon_{igmspr}, \tag{4}$$

where

$$\begin{aligned}
\mu_i &= \text{mean gene expression for } i\text{th gene} \\
\tau_{ig} &= \text{group } g \text{ effect on } i\text{th gene} \\
M_{im(g)} &= \text{effect of subject } m \text{ in group } g \text{ on } i\text{th gene} \\
S_{is(m(g))} &= \text{effect of } s\text{th sample from } m\text{th subject in group } g \text{ on } i\text{th gene} \\
\pi_{p(i)} &= \text{effect of } p\text{th probe on } i\text{th gene} \\
\epsilon_{igmspr} &= \text{measurement error.}
\end{aligned}$$

Replicate measurement errors are assumed to be independent, identically distributed with variance $\sigma^2_{i\epsilon}$. If they are normally distributed, then estimated group differences have a multivariate normal distribution, from which multiple tests can be derived. If they are not normally distributed, then multiple tests based on resampling of the residuals can be used.

Subject and sample effects can be considered as either fixed or random depending on the purpose of the microarray analysis. If the subject and sample effects are random, their variances are assumed to be $\sigma^2_{iM}$ and $\sigma^2_{iS}$ respectively. In this case, $\text{var}(Y_{igmspr}) = \sigma^2_{i\epsilon} + \sigma^2_{iM} + \sigma^2_{iS}$, the covariance between gene expressions for different replicates from the same sample is $\sigma^2_{iM} + \sigma^2_{iS}$ and the covariance between gene expressions for different samples from the same subject is $\sigma^2_{iM}$.

Model (4) is a *marginal* model in the sense that it does not specify the *joint* distribution of $y_{igmspr}$ across the genes. Modelling expression levels from all genes simultaneously with subject by gene interaction included would generally require too much computer memory.

Other models exist for gene expression data from cDNA spotted arrays and stock Affymetrix arrays (e.g., Wolfinger et al., 2001, Chu et al., 2002, 2004, Smyth, 2004, Lee et al., 2002). However, the model (4) differs significantly from the existing models in that the design underlying it allows for the separate estimation of the subject and sample effects.

Consider, for example, the model Chu et al. (2004) used to analyze the ionizing radiation data in Tusher, Tibshironi, and Chu (2001). Each of two samples from each of four groups (treatment and cell line combinations with two levels each) was hybridized to a stock Affymetrix array. On a gene-by-gene basis, their linear mixed effects model was:

$$Y_{igpl} = \mu_i + \tau_{ig} + \pi_{p(i)} + A_{l(g)} + \epsilon_{igpl} \tag{5}$$

where $Y_{igpl}$ is the log transformed perfect match values for probe-level data from the $i$th gene, $\mu_i$ is the mean gene expression for the $i$th gene, $\tau_{ig}$ is the $g$th group (treatment and cell line combination) effect on the $i$th gene, $\pi_{p(i)}$ is the $p$th probe effect in the $i$th gene, $A_{l(g)}$ is a random array effect, and $\epsilon_{igpl}$ is measurement error.

Comparing model (5) with model (4), we see the array effect $A_l(g)$ in model (5) includes both the subject effect $M_{im(g)}$ and sample effect $S_{is(m(g))}$ in model (4). That is, with stock Affymetrix arrays, sample and subject effects are completely confounded with array effects, and cannot be estimated. Therefore, treatment or group effect is confounded with array effect in this case.

Smyth (2004) applied the following simple linear *fixed* effect model to fit log-transformed

intensities. For the $i$th gene, assume

$$E(\boldsymbol{Y}_i) = X\boldsymbol{\alpha}_i \tag{6}$$

$$var(\boldsymbol{Y}_i) = W_i\sigma_i^2, \tag{7}$$

where $\boldsymbol{Y}_i$ is a vector containing all the transformed intensities from different samples for the $i$th gene, $X$ is the design matrix and $\boldsymbol{\alpha}_i$ is a vector containing all the parameters $\mu$, $\tau$ and $\pi$ for the $i$th gene. $W_i$ is assumed to be a known non-negative definite matrix. It is not entirely clear how to set values for the matrix $W_i$ in real applications.

Models (6) and (4) have difference in the variance matrix of the vector $\boldsymbol{Y}_i$. In particular, the variances of the elements of $\boldsymbol{Y}_i$ in the model (6) are $\sigma_i^2$ times the diagonals of the matrix $W_i$, while they are given by $\sigma_{i\epsilon}^2 + \sigma_{iM}^2 + \sigma_{iS}^2$ (a combination of separate variance components from subject effects, sample effects, and measurement error) in the model (4). Also, the covariances among gene expression levels are modelled differently in (4) and (6). Model (4) specifies the covariance between gene expression levels of different samples from the same subject to be $\sigma_{iM}^2$ and the covariance between gene expression levels of the same sample to be $\sigma_{iM}^2 + \sigma_{iS}^2$. In other words, (4) models the additional covariance due to measurements being from the same sample *additively*. Model (6), on the other hand, assumes both of the covariance between gene expression levels from the same subject and that from the same sample to be multiples of $\sigma_i^2$. These multiples are to be specified in the matrix $W_i$. However, proper specification of the multiples does not seem to be straightforward, especially when the sample and subject effects are additive.

In order to borrow information from the ensemble of genes to assist in estimation of variance of each individual gene, Smyth (2004) assumes a prior distribution on $\sigma_i^2$,

$$\frac{1}{\sigma_i^2} \sim \frac{1}{d_0 s_0^2}\chi_{d_0}^2.$$

With this prior specification, (6) is not a gene-by-gene model any more. The unknown variance $\sigma_i^2$ in model (6) can then be estimated by the posterior mean of $\sigma_i^2$ given $s_i^2$, i.e.

$$\hat{\sigma}_i^2 = \frac{d_0}{d_0 + d_i}s_0^2 + \frac{d_i}{d_0 + d_i}s_i^2.$$

The estimate $\hat{\sigma}_i^2$ shrinks the observed variances $s_i^2$ towards the prior values $s_0^2$ with the extent of shrinkage determined by the relative sizes of the observed and the prior degrees of freedom $d_i$ and $d_0$. This is similar in principle to Tusher, Tibshirani, and Chu (2001)'s idea in SAM of modifying the estimate of $\sigma_i$ by an offset parameter, i.e. $\hat{\sigma}_i = s_i + s_0$. Smyth's offset estimate is motivated by a hierarchical model, whereas $s_0$ in SAM is empirically chosen

to be a particular percentile of all $s_i$ values without a model or an associated distribution theory.

The two-stage ANOVA model in Lee et al. (2002) is basically the same as model (5) except that all effects are assumed to be fixed. They also adjust the mean square error by an offset quantity in testing for significance using $F$-statistics.

# 7 A proof of concept experiment

We executed a microarray experiment to prove the concept that a training experiment can be designed statistically to reliably estimate the variance components of subject, sample, and noise separately, and that marker genes can be selected for the validation study by multiple testing with a properly controlled error rate.

To simulate the comparison of phenotype groups, our experiment compared tissues from normal mice (wild type, labeled group W) with tissues from four groups of mice (labeled groups A, B, C and D) with four different mutations of the microphthalmia transcription factor (Mitf) gene.

Four mice were sampled from each of the five groups. From each mouse, four cRNA samples were prepared after total RNA was isolated and biotin labeled cDNA was synthesized from a spleen tissue sample.

To prove the concept that external reference samples can be used to control for the platform change, we also prepared 16 samples of Universal Mouse Reference Sample (UMRS) from StrataGene.

To demonstrate statistical design of hybridization of samples to microarrays, we utilized NimbleGen microarrays with 12 mini-microarrays on each array. The 96 samples were hybridized to eight arrays, with the samples placed in the mini-microarrays according to the three rows by four columns patterns shown in Figure 4, following the statistical principles of randomization, replication, and blocking.

A total of 99 genes thought to be regulated by the Mitf gene were selected as probes. The probe set for each gene consisted of thirty-two 24-mer probes. To demonstrate the utility of statistically designing microarrays according to the principles of randomization and replication, each probe set was replicated twice in each of the mini-microarrays, and placements of the probes were completely randomized in each mini-microarray.

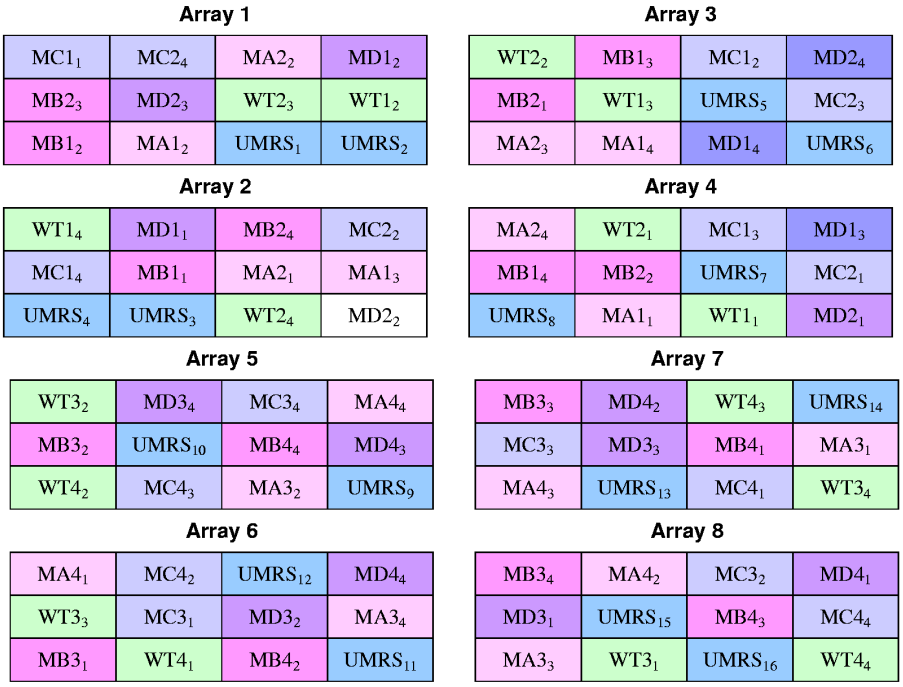


Figure 4: Hybridization design in the proof of concept experiment. (samples from different groups and UMRS are represented by different shades of gray, for example, $MA2_3$ is the third cRNA sample from the second mouse from group A, while $UMRS_2$ is the second UMRS).

## Normalization using an external reference sample

Gene expression measurements from microarrays are subject to array and other processing effects, and are usually "normalized" before group comparisons are made. To "normalize" is to pre-process data to ensure observations from different sources are compatible before inferences are made.

Internal normalization uses samples within a study as controls, while external normalization uses samples external to the study as controls.

Some internal normalization techniques such as quantile normalization have been shown to be reliable within a study. However, how well inferences (such as prognostic algorithms) based on internal normalization carry across different studies is less clear.

External normalization uses reference samples that are homogeneous independent of platforms. Such external normalization can control for platform changes, provided it is as reliable within each platform as proven by internal normalization techniques. We demonstrate the viability of external normalization by showing that internal and external normalizations pro-

duce almost identical results in our study. After correcting probe level measurements for background as described in Irizarry et al. (2003), we applied both normalization techniques, as described below.

With microarrays that allow only one biological sample to be placed on each array, it is unclear whether arrays from different groups should be normalized together or separately, due to confounding of array and group effects.

In our proof of concept experiment, however, the number of samples from each of the six groups is the same across all eight arrays. Specifically, every group of mice (groups A, B, C, D and W) and the external reference sample, UMRS, appear exactly twice on each array. It is thus reasonable to expect the distribution of the probe intensities to be the same across the arrays. We applied quantile normalization to equalize the distributions of the vectors of intensities from the eight arrays.

We propose an array-by-array external normalization process:

1. For each array, first generate a "reference" mini-microarray by averaging, for each probe, the intensities for that probe measured from UMRSs.

2. Then subtract the probe intensities in the reference mini-microarray from the corresponding probe intensities in every other (non-reference) mini-microarray.

To make an analogy to clinical trials, external normalization uses UMRS as a control.

We compared estimated differences between mutated types and wild type (A vs W, B vs W, C vs W and D vs W), after fitting the data normalized by the two techniques to the marginal model (4), using PROC MIXED of the SAS System. (Array 5 data was excluded due to bad quality.) Figure 5 shows that these two normalization techniques produce practically the same results. As quantile normalization is considered reliable, our study shows that normalization via external reference samples is a viable technique for coping with platform change issues.

## Multiple testing for differential expressions

To discover genes differentially expressed between each mutated group (A, B, C, D) and wildtype (W), consider testing the $4 \times 99$ hypotheses

$$\{\{H_{ig} : \ \tau_{ig} = \tau_{iW}, \ g = A, B, C, D\}, \ i = 1, 2, ..., 99\}. \tag{8}$$

in the model (4).

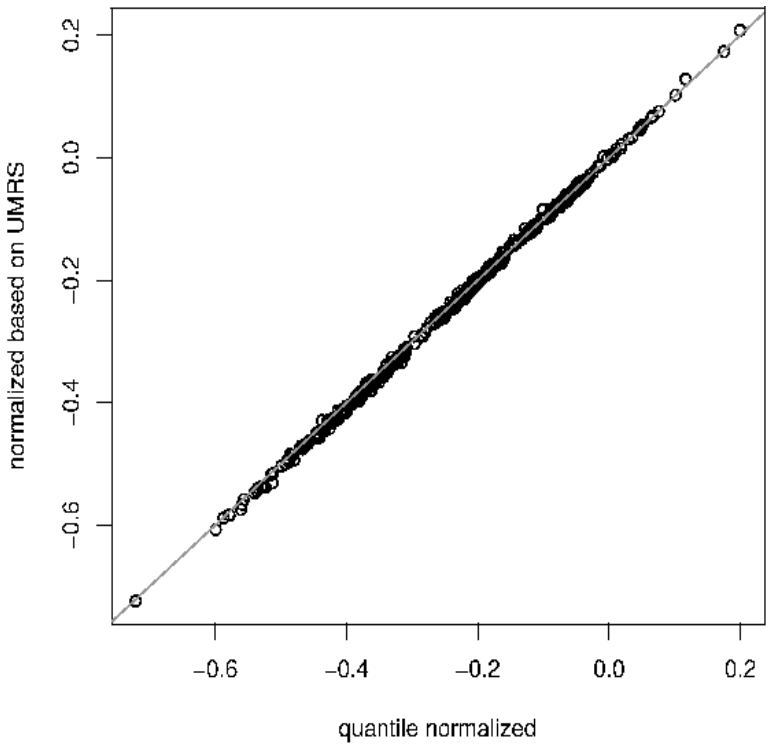


Figure 5: Difference estimates between mutated types and wild type under quantile normalization and UMRS-based normalization.

For a particular gene $i$, estimates of $\tau_{ig} - \tau_{iW}, g = A, B, C, D$, are correlated due to estimating a common $\tau_{iW}$. For a balanced design, if the errors are i.i.d. with normal, this correlation is 0.5, not negligible, and it should be taken into account in multiple testing.

For a particular group $g$, estimates of $\tau_{ig} - \tau_{iW}, i = 1, \ldots, 99$, may have some correlation. Since there are $99 \times 98/2$ correlations, the correlation structure is not easy to infer given the typical amount of data from a microarray experiment. Therefore, one can take either a conservative approach regarding these correlations or a resampling approach to adjust for the correlations.

However, for a pair of $\tau_{ig} - \tau_{iW}$ and $\tau_{jh} - \tau_{jW}$ involving different genes and different groups, $i \neq j$ and $g \neq h$, correlation between estimates is expected to be small because the only dependence comes from the correlation between the estimates of $\tau_{iW}$ and $\tau_{jW}$. Suppose this correlation equals $\rho$. Then, for a balanced design, the correlation between the estimates of $\tau_{ig} - \tau_{iW}$ and $\tau_{jh} - \tau_{jW}$ equals $\rho/2$.

Therefore, a practical approach to controlling the error rate in testing the $4 \times 99$ hypotheses (8) is to take dependence into account in adjusting for multiplicity of testing

$\{H_{ig} : \tau_{ig} = \tau_{iW},\ g = A, B, C, D\}$ for each gene $i$ (in computing raw $p$-values). That is, apply the Bonferroni correction adjusting for multiplicity of the four sets of hypotheses

$$\{H_{ig} :\ \tau_{ig} = \tau_{iW},\ i = 1, 2, ..., 99\},\ g = A, B, C, D, \tag{9}$$

and finally apply an appropriate method to control in testing each set of 99 hypotheses in (9).

Table 7 gives the number of genes discovered to be differentially expressed between each mutated group and wildtype, controlling gFWER at 5%, based on the linear model (4) with both mouse and sample effects as fixed effects. Reported are the results of applications of the Holm method, partition method using Markov's inequality, and the augmentation method for testing each set of 99 hypotheses in (9) at gFWER of 5%/4 under the assumption that the errors are i.i.d. normal. Alternatively, without the normal assumption, the residuals from least squares estimates were resampled for 10,000 times, and the results of the resampling method are also reported.

All methods find that groups $C$ and $W$ are most different, and groups $B$ and $W$ are closest in terms of the measured gene expression levels. It confirms the findings in Figure 3. When FWER was controlled, Holm's procedure and resampling method gave very similar results. When the number of mistakes allowed increased to 5, resampling method came out to be much less conservative than other gFWER-controlled methods, especially for the comparison between group $B$ and group $W$.

Number of rejections by the step-down method controlling the gFWER at 5%, Holm method, partition method using Markov's inequality, augmentation method, and resampling method.

| $m$ | Method | A vs. W | B vs. W | C vs. W | D vs. W |
|---|---|---|---|---|---|
| 0 | Holm | 94 | 18 | 99 | 92 |
| 5 | Markov | 94 | 27 | 99 | 95 |
| 5 | Augmentation | 99 | 23 | 99 | 97 |
| 0 | Resampling | 94 | 17 | 99 | 92 |
| 5 | Resampling | 97 | 48 | 99 | 96 |

# 8 Software Implementation

This section illustrates software implementation with SAS (linear mixed effects modeling) and R (multiple testing) for the modeling-based analyses of microarray data described in previous sections.

## 8.1 SAS Procedures

Sensitivity and specificity of a classification algorithm depend on how variable expressions are between subjects within each group and between samples within each subject, the variability of subject and sample can be estimated by considering them as random effects in modeling expression level data. In this case, the background corrected, log transformed and normalized probe level expressions for each gene are fitted by a linear mixed effect model with subject and sample effects as random:

```
proc mixed data=bcnorm;
class group probe mice sample;
model response= group probe/DDFM=SATTERTH solution;
random mice(group) sample(group mice)/solution;
lsmeans group/diff=control('W') adjust=dunnett;
run;
```

Group effects estimates, and variance components estimates for subject effect, sample effect and measurement error for each gene are saved for sample sizes calculation in validation trial.

On the other hand, to discover group differences, if subject and sample effects can be estimated unbiasedly, then removing them may make group differences reveal themselves more readily. By treating subject and sample as fixed, one may more readily identify differentially expressed genes. In this case, the background corrected and normalized probe level expressions for each gene are fitted by a linear fixed effect model. The residuals for each gene are saved for multiple testing procedure by resampling, without assuming the errors are normally distributed.

## 8.2 R Functions

To discover differentially expressed genes between two groups, multiple testing procedure by resampling the residual is applied to control the gFWER at level 5%/4. It is implemented by the following steps:

1. Resample independently with replacement the residual vectors after linear fixed effect modeling. To account for potential dependence among the measurement errors across genes, the residuals are resampled vector at a time, with each vector consisting of residuals from within each sample. For each re-sampled data set, we compute the test statistic for each gene, which is the difference of the average intensities, averaging within each mouse and then averaging over the mice within each group. After repeated resampling $B$ times, we have an estimated null distribution for the test statistic of each gene.

2. Calculate $p$-values for each gene by comparing the observed test statistics with the estimated null distribution T.mat.

```
p<-(abs(T.mat)- abs(t)>0)%*%rep(1,B)/B
```

3. Build null distribution for the $p$-values. Independently generate another matrix as we did in step 1. The null distribution for $p$-values is then estimated by comparing this new matrix a.T.mat with the matrix T.mat.

```
nullP<- 1- apply(T.mat,2,fn.rawp.T, a.T.mat)

fn.rawp.T<-function(nullT, obsT)
{
return((abs(nullT)- abs(obsT)>0)%*%rep(1,B)/B )
}
```

4. Calculate the adjusted p-value based on the generalized partitioning principal in Xu and Hsu (2007), to control the gFWER with $m = 5$.

```
# sort the matrix nullP so that the corresponding raw
# p-values are increasing
ind<-sort(p,index.return=T)$ix
p.H.sort<-p.H[ind, ]
P.mat.sort<-nullP[ind, ]

# step-down gFWER control
m<-5
adjp<-rep(NA,99)
```

```
# step 0
adjp[1:m]<-0
P.mat1<-P.mat.sort
# step 1
# minp here is actually p(m+1)
minp<-apply(P.mat1,2,function(x) x[(sort(x,index.return=T)$ix)[m+1]])
adjp[m+1]<-sum(minp<p.H.sort[m+1,1])/B
# step 2 and etc.
for (i in (m+2):99)
{
P.mat1<-P.mat1[-1,]
minp<-apply(P.mat1,2, function(x) x[(sort(x,index.return=T)$ix)[m+1]])
adjp[i]<-sum(minp<p.H.sort[i,1])/B
# enforce adjp to be increasing
adjp[i]<-max(adjp[i-1], adjp[i])
}

# present the raw p-values, adj p-values
# and the corresponding gene No.
adjp.H<-cbind(adjp, p.H.sort)
colnames(adjp.H)<-c("adjp","rawp","gene")
adjp.sd.gfwer<-adjp
```

**Acknowledgement** This research was supported in part by NSF Grant No. DMS-0505519 and a grant from the Icelandic Science and Technology Council.

# References


Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society B* **57**: 289–300.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, second edn, Springer-Verlag, New York.

Berger, J. O. and Wolpert, R. L. (1088). *The Likelihood Principle*, 2nd edn, Institute of Mathematical Statistics.

Calian, V., Li, D. and Hsu, J. C. (2008). Partitioning to uncover conditions for permutation tests to control multiple testing error rates, *Biometrical Journal* **50**: 756–766.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*, Duxbury Press.

Chu, T.and Weir, B. and Wolfinger, R. (2002). A systematic statistical lin- ear modeling approach to oligonucleotide array experiment, *Mathematical Biosciences* **176**: 35–51.

Chu, T.and Weir, B. and Wolfinger, R. (2004). Comparison of li-wong and loglinear mixed models for the statistical analysis of oligonucleotide arrays, *Bioinformatics* **20**: 500–506.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing, *Journal of the American Statistical Association* **102**: 93–103.

FDA (2003). *Multiplex Tests for Heritable DNA Markers, Mutations and Expression Patterns: Draft Guidance for Industry and FDA Reviewers*, Center for Devices and Radiological Health (CDRH), U.S. Food and Drug Administration.

FDA (2005a). *Drug-diagnostics co-development concept paper*, CDER/CDRH/CBER/OCP, U.S. Food and Drug Administration.

FDA (2005b). *Pharmacogenomic Data Submission: Guidance for Industry*, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH), U.S. Food and Drug Administration.

FDA (2006). *In Vitro Diagnostic Multivariate Index Assays: Draft Guidance for Industry, Clinical Laboratories, and FDA Staff*, Center for Devices and Radiological Health (CDRH), U.S. Food and Drug Administration.

Genz, A. and Bretz, F. (1999). Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts, *Journal of Statistical Computation and Simulation* **63**: 361–378.

Gordon, A., Glazko, G., Qiu, X. and Yakovlev, A. (2007). Control of the mean number of false discoveries, bonferroni and stability of multiple testing, *Annals of Applied Statistics* **1**: 179–190.

Hommel, G. and Hoffmann, T. (1988). Controlled uncertainty, *in* P. Bauer, G. Hommel and E. Sonnemann (eds), *Multiple Hypotheses Testing*, Springer, Berlin, pp. 154–161.

Hsu, J. C. (1992). The factor analytic approach to simultaneous inference in the general linear model, *Journal of Graphical and Computational Statistics* **1**: 151–168.

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

Lee, M. T., Lu, W., Whitmore, G. and Beier, D. (2002). Models for microarray gene expression data, *Journal of Biopharmaceutical Statistics* **12**: 1–19.

Lehmann, E. L. and Romano, J. (2005). Generalizations of the familywise error rate, *Annals of Statistics* **33**: 1138–1154.

Pollard, K. S. and van der Laan, M. J. (2005). Resampling-based multiple testing: Asymptotic control of type I error and applications to gene expression data, *Journal of Statistical Planning and Inference* **125**: 85–100.

Rao, Y., Lee, Y. and Hsu, J. C. (2009). Determination of sample size for validation study in pharmacogenomics, in preparation.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing, *Annals of Statistics* **35**: 1378–1408.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray, *Statistical Applications in Genetics and Molecular Biology* **3**.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *Proc. Natl. Acad. Sci. U.S.A.* **98**: 5116–5121.

van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives, *Statistical Applications in Genetics and Molecular Biology* **3**.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer, *Nature* **415**: 530–536.

Wolfinger, R., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R. (2001). Assessing gene significance from cdna microarray expression data via mixed models, *Journal of Computational Biology* **8**: 625–637.

Xu, H. and Hsu, J. C. (2007). Using the partitioning principle to control the generalized family error rate, *Biometrical Journal* **49**: 52–67.