META-ANALYSIS FOR PATHWAY ENRICHMENT ANALYSIS AND BIOMARKER DETECTION WHEN COMBINING MULTIPLE GENOMIC STUDIES

by

Kui Shen

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

School of Medicine

This dissertation was presented

by

Kui Shen

It was defended on

April 5, 2010

and approved by

Committee Member:

Eleanor Feingold, Ph.D.
Professor
Department of Human Genetics
Department of Biostatistics
University of Pittsburgh

James Faeder, Ph.D
Associate Professor
Department of Computational Biology
University of Pittsburgh

Kathryn Roeder, Ph.D
Professor
Department of Statistics
Carnegie Mellon University

Dissertation Advisor:

George C. Tseng, Ph.D
Associate Professor
Department of Biostatistics
University of Pittsburgh

Meta-analysis for pathway enrichment analysis and biomarker detection when combining multiple genomic studies

Kui Shen, PhD

University of Pittsburgh, 2010

This thesis focuses on applying meta-analysis methods for combining genomic studies on biomarker detection and pathway enrichment analysis. DNA microarray technology has been maturely developed in the past decade and led to an explosion on publicly available microarray data sets. However, the noisy nature of DNA microarray technology results in low reproducibility across microarray studies. Therefore, it is of interest to apply meta-analysis to microarray data to increase the reliability and robustness of results from individual studies. Currently most meta-analysis methods for combining genomic studies focus on biomarker detection, and meta-analysis for pathway analysis has not been systematically pursued. We investigated two natural approaches of meta-analysis for pathway enrichment (MAPE) by combining statistical significance across studies at the gene level (MAPE_G) or at the pathway level (MAPE_P). Simulation results showed increased statistical power of both approaches and their complementary advantages under different scenarios. We also developed an integrated method (MAPE_I) that incorporates advantages of both approaches. Applications to real data on drug response of a breast cancer cell line, lung and prostate cancer tissues were evaluated to compare the performance of the different methods. MAPE_P has the general advantage of not requiring gene matching across studies. When MAPE_G and MAPE_P show complementary advantages, the integrated version MAPE_I is recommended. A software package named MetaPath, was implemented to perform the MAPE analysis. In addition to developing MAPE

methods, we also applied meta-analysis approach to chemotherapy research to discover robust biomarkers and multi-drug response genes, which have prognostic value and the potential of identifying new therapeutic targets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

As a graduate student in an interdisciplinary program in Computational Biology, I have faced many choices. Should I study stochastic differential equations to model the neuronal system? Should I learn stochastic signal processing to analyze electrocardiography? Or should I devote myself to genomics and microarray analysis? The answer became clear after I took a microarray course taught by Dr. Tseng, who introduced me to the world of statistical analysis of microarray studies. So I would like to express my gratitude to Dr. Tseng, my academic advisor, for mentoring me throughout all my studies and research. I am greatly thankful to our department chair, Dr. Bahar, our student advising committee members, Dr. Camacho and Dr. Zuckerman, and all other faculty members in our program for providing a high quality study and research environment and guiding me to choose the right research topics and the advisor. I also want to express my sincere appreciation for the efforts of my academic committee members, Dr. Faeder, Dr. Feingold, and Dr. Roeder, for their guidance and encouragement of my research. My special thanks to Mr. Gabrin and my colleagues at Precision Therapeutics, Inc. for giving me the chance to conduct an extraordinarily interesting genomic project. Finally, I want to thank my parents and friends for their encouragement, love, and support. I dedicate this work to my wife, Li, my son, Eric, and my daughter, Vivian.

1.0 INTRODUCTION

DNA microarray technology (Kulesh, et al., 1987; Lashkari, et al., 1997; Schena, et al., 1995) provides the ability to detect genome-wide gene expression activities with thousands of probes printed on each high-density chip. It has evolved rapidly in the past decade and has gradually become a standard tool for many biomedical studies. The wide applications of microarray technology have led to an explosion of gene expression profiling studies publicly available. However, the noisy nature of microarray data (Tu, et al., 2002), together with the relatively small sample size in each study, often results in inconsistent biological conclusions (Ein-Dor, et al., 2005). Therefore, methods for synthesizing multiple microarray studies are greatly needed. Meta-analysis, a set of statistical techniques to combine results from several studies, has been recently applied to microarray analysis to increase the reliability and robustness of results from individual studies. Currently, meta-analysis methods for microarray studies are mostly aimed at combining different studies to identify differentially expressed (DE) genes, an analysis at the gene level. However, DE gene analysis has two main shortcomings. First, the identified DE genes may not biologically relate to the phenotype of interest. Second, a gene set from an important pathway may act in concert with moderate activities, which cannot be detected by DE gene analysis, while the pathway may have important biological effects on the phenotype of concern (Subramanian, et al., 2005). To overcome these shortcomings, pathway analysis has been developed, which also has an inherent advantage for work with meta-analysis. It is well-known that the lists of DE genes from independent studies associated with same the phenotype often have little overlap (Ein-Dor, et al., 2005), while pathway analysis often generates improved consistency (Manoli, et al., 2006). This situation motivates us to develop systematic approaches of meta-analysis for pathway enrichment (MAPE), which provides a more robust and powerful tool than standard pathway enrichment analysis. To our knowledge, this is the first study to systematically develop and evaluate meta-analysis methods for pathway analysis in microarray studies.

In addition to the investigation of the meta-analysis method for pathway enrichment analysis, meta-analysis was also applied to the field of chemotherapy research in this thesis for the following two topics: identification of robust chemotherapy response biomarkers and identification of multi-drug response genes in human breast cancer cell lines.

This dissertation is organized as follows: in Chapter 1, meta-analysis and pathway enrichment analysis methods are reviewed. In Chapter 2, two approaches to meta-analysis for pathway enrichment, MAPE_G and MAPE_P, are described; MAPE_G combines statistical significance across studies at the gene level and MAPE_P at the pathway level. Then an integrated method (MAPE_I) is introduced to incorporate the advantages of both MAPE_G and MAPE_P. Simulation results and applications to real data sets are also shown Chapter 2. The implementation and usage of the MetaPathsoftware package are described in Chapter 3. In Chapter 4, meta-analysis was applied to identify robust biomarkers and multi-drug-response genes. Conclusions and discussions are provided in Chapter 5.

1.1 MICROARRAY DATA STANDARDIZATION

DNA microarray technology (Kulesh, et al., 1987; Lashkari, et al., 1997; Schena, et al., 1995) evolved from Southern blotting, a nucleotide hybridization technique developed by Southern in 1975 (Southern, 1975) for detection of a specific DNA sequence in DNA samples. Southern blotting can process only a single or few genes, while microarray technology circumvented this restriction by using thousands of different probes attached to a solid surface. Each microarray probe contains a specific DNA sequence, a short gene segment or other DNA section of interest, to hybridize target cDNA samples under high-stringency conditions. Probe-target hybridization can be quantified by measuring fluorophore-labeled targets to determine relative amounts of DNA sequences in target samples.

Multiple microarray platforms are available, such as cDNA microarray (DeRisi, et al., 1996), Affymetrix (Auer, et al., 2009) and Illumina (Fan, et al., 2006). Due to inconsistent standardization in platform fabrication, microarray data are not directly comparable. To ease the exchange and analysis of microarray data from different platforms, it is necessary to address two issues about microarray data standardization: standard data structure for individual microarray studies and microarray probe ID mapping across microarray platforms.

1.1.1 Microarray data structure

To standardize microarray data structure, a Minimum of Information About a Microarray Experiment (MIAME) project (Brazma, et al., 2001) was proposed, which has six critical elements as follows:

1. The raw microarray data such as CEL files for Affymetrix platform or GPR files for cDNA platform
2. The final data after raw microarray data pre-processing and normalization (Geller, et al., 2003; Quackenbush, 2002; Schadt, et al., 2001; Steinfath, et al., 2001), usually denoted by a numeric data matrix
3. The essential annotation for samples such as experimental factors and their values.
4. The experimental design
5. Array annotation such as gene identifiers and probe oligonucleotide sequences
6. The protocols for laboratory and data processing

MIAME is supported by two major public microarray databases, Gene Expression Omnibus (GEO) (Barrett, et al., 2009; Edgar and Barrett, 2006) and ArrayExpress (Rustici, et al., 2008). In this dissertation, all microarray data sets subjected to our analysis were downloaded from public websites and have been packaged into a MIAME data object using the R language (R Development Core Team, 2005) and the Bioconductor package (Dudoit, et al., 2003; Gentleman, et al., 2004; Kauffmann, et al., 2009; Nie, et al., 2009).

1.1.2 Mapping probe IDs to gene IDs

Because different microarray platforms use their own probe IDs, gene expression values from different platforms cannot be compared directly. Normally, probe IDs from different platforms are mapped to common gene IDs such as Entrez gene IDs or gene symbols for cross-platform comparison (Wheeler, et al., 2003). However, problems arise because one Entrez gene may correspond to multiple probe IDs. For example, 22283 probe IDs in the Affymetrix Hgu133a chip have been mapped to 12998 Entrez genes by the Entrez Gene database (Maglott, et al., 2005) on March 11, 2009. Among all 12998 Entrez genes, 37.7% of them have more than one corresponding probe IDs. Thus a method to map the expression values of probe IDs to gene IDs is needed (Stalteri and Harrison, 2007). In this dissertation, a simple but acceptable method has been adopted (Falcon and Gentleman, 2007). If N probe IDs map to one Entrez ID, we selected the probe ID with the largest interquartile range (IQR) of expression values among all N probe IDs to represent the corresponded Entrez ID.

In conclusion, microarray data standardization and microarray probe ID mapping for cross-platform comparison have been discussed in this subsection. These steps are data pre-processing procedures before performing meta-analysis. In the next subsection, meta-analysis and pathway enrichment analysis methods will be reviewed. For simplicity, we assume genes in multiple microarray studies are matched by gene symbols and no missing value exists.

1.2 META-ANALYSIS METHODS

As early as 1904, Karl Pearson (Pearson, 1904) introduced meta-analysis, a method which combines the results of several studies to generate more powerful statistics than would be provided by analyzing individual studies. Since that time, meta-analysis has been widely applied in epidemiologic research (Annie J. Sasco, et al., 1993; Hettema, et al., 2001; Stroup, et al., 2000).

In the literature, there are two major categories of meta-analysis: one combines statistical significance and the other combines effect sized from individual studies. In the next two subsections, we will introduce popular methods of each category and their applications to microarray studies.

1.2.1 Methods for combining statistical significance

Suppose there are K independent experiments performed to measure a certain effect. θ_k are the unknown parameters that characterize the effect of study k, $k = 1,\ldots,K$. The null hypothesis for the kth experiment is $H_{0k}: \theta_k = 0$. If T_k has a continuous distribution, the significance of a test can be defined as the p-value, which is $p_k = Pr(T_k > t_k|H_{0k})$. When H_{0k} is true, p_k is uniformly distributed. Since the p-value does not depend on the statistical distribution of the data, a test of the combined statistical significance reflected by p-values is a nonparametric test for meta-analysis. It is only dependent on the fact that the p-values are uniformly distributed between 0 and 1 under the null hypothesis.

1.2.1.1 Minimum and maximum p-value statistics

In 1931, Tippett proposed minimum p-value statistics (Tippett, 1931), which can be given by:

$$V^{\min P} = \min p_k, k = 1, ..., K,$$

where p_k is the p-value of in study k. Under the null hypothesis that no genes are differentially expressed, p_k is uniformly distributed on the interval [0, 1]. Therefore, the distribution of V^{minP} under the null hypothesis can be easily derived which is a beta distribution with parameters $\alpha=1$ and $\beta=K$. The test became: H_0 is rejected if $V^{\min P} < 1-(1-\alpha)^{1/k}$, where α is the overall significance level.

The maximum p-value statistic is

$$V^{\max P} = \max p_k, k = 1, ..., K.$$

Similarly, the distribution of V^{maxP} can be derived as a beta distribution with parameters $\alpha=K$ and $\beta=1$. Both minimum and maximum p-value statistics can be considered to be a special case of a more robust *rth* smallest p-value statistics, $Vth=p_{(r)}$ (Wilkinson, 1951).

1.2.1.2 Fisher's statistic

The well-known Fisher's statistic (Mosteller and Fisher, 1948) can be obtained from the following formula,

$$V^{Fisher} = -2\sum_{k=1}^{K} \log(p_k).$$

Under the null hypothesis that no genes are differentially expressed, the distribution of p_k is a uniform distribution on the interval [0, 1]. The distribution of $-log(p_k)$ is then an exponential distribution with parameter β=1, or equivalently, a gamma distribution with parameters $\alpha=1$ and $\beta=1$. Therefore, the distribution of V^{Fisher} is a gamma distribution with parameters $\alpha=k$ and $\beta=1/2$, in other words, chi square distribution with $2k$ degrees of freedom. Fisher's statistic takes advantage of the relationship between the uniform distribution and the chi-square distribution. The test procedure is simple. H_0 is rejected if $V^{Fisher} > C$, where C is the critical value that can be obtained from the upper tail of the chi-square distribution with $2k$ degrees of freedom. Fisher's statistic has been applied in many fields.. Although it has been shown there is no uniformly most powerful test under Gaussian assumptions, Fisher's method has been shown to be powerful under a wide range of alternative hypothesis conditions (Loughin, 2004; Schmid, et al., 1991).

1.2.1.3 Weighted Fisher's statistic

Good (Goods, 1955) extended Fisher's statistics by assigning different positive weights to the K experimental results and proposed the weighted Fisher's statistic

$$V^{WF} = -\sum_{k=1}^{K} w_k \log(p_k),$$

where w_k is the constant weight for the kth study. The weight can be determined based on available prior information such as study quality or expert opinion. Based on Good's work, the exact distribution function of V^{WF} is

$$P(V^{WF} < x) = 1 - \sum_{k=1}^{K} \Lambda_k e^{-x/2w_k},$$

where

$$\Lambda_k = \frac{w_k^{K-1}}{\prod_{\substack{j=1 \\ j \neq k}}^{K} (w_k - w_j)}.$$

Koziol (Koziol and Perlman, 1978) proved if the prior information is available and correct, the weighted Fisher's procedure has an increased power at the alternatives of interest than standard Fisher's procedure. However, there are two issues about Good's work. One is that the exact distribution of weighted Fisher’s statistic will result in ill-conditioned calculations if any of the weights is zero or if two weights are equal. The other is that the choice of weight is somewhat subjective. To solve these problems, Li and Tseng (Li 2008) recently proposed an adaptively weighted statistic, discussed in the next section.

1.2.1.4 Adaptively weighted Fisher’s statistic

The adaptively weighted (AW) Fisher’s statistics was proposed by Li and Tseng (Li 2008) using the following formula

$$V^{OW} = \min_{w \in W} p(u_g(w)),$$
$$u_g(w) = -\sum\nolimits_{k=1}^{K} w_k \log(p_k),$$

where w_k is the weight assigned to the *kth* study and $w = (w_1, \ldots, w_k)$. For simplicity and better biological interpretation, but without loss of generality, the search space is $W = \{w \mid w_i \in \{0,1\}\}$. Compared with weighted Fisher’s statistics, the AW statistic provides a data-driven method to estimate the weight for each study. In addition, the weights used in AW statistic do not have the same limitation as weighted Fisher’s statistics has (the weight cannot be zero and any of two weights cannot be equal). The AW statistic was designed but not limited to combine microarray studies. The adaptively weights provide a natural categorization of the detected DE genes and biological interpretation of whether or not a study contributes to the statistical significance of a gene.

1.2.1.5 Inverse normal statistic

An additional procedure for combining p-values that has widespread use in meta-analysis is the inverse normal method that was proposed by Stouffer (Stouffer, et al., 1949):

$$V^{Z}=\frac{\sum_{k=1}^{K}\Phi^{-1}(p_k)}{\sqrt{K}}.$$

Under the null hypothesis, it is an asymptotically standard normal distribution. H0 is rejected when VZ is larger than the critical value of the standard normal distribution.

The inverse normal statistic also has a weighted version

$$V^{WZ}=\frac{\sum_{k=1}^{K}w_k\Phi^{-1}(p_k)}{\sqrt{K}}.$$

Koziol (Koziol and Perlman, 1978) investigated the power of the inverse normal statistic. He did not recommend the inverse normal procedure since its power is relatively high only in a narrow central wedge of the alternative space.

1.2.2 Methods for combining effect sizes

The methods for combining significance do not provide information concerning the size of the treatment effect. Therefore, when studies have comparable designs and measure the outcome in a similar manner, methods for combining estimates are preferred to the non-parametric methods. Fixed, random and mixed effects models are three major types of statistical analysis for combining estimates. A review of the methods for combining estimates was provided by Hedges (Hedges, 1992). These methods are beyond the scope of this dissertation and the details are not discussed here.

1.2.3 Meta-analysis methods for microarray studies

Many studies reported in the biological literature for combining microarray studies used the naïve method involving widespread use of intersection/union operations or simple counting of appearances in the differentially expressed gene lists obtained from individual studies under certain criteria (e.g. False Discovery Rate= 0.05) (Borovecki, et al., 2005; Cardoso, et al., 2007; Pirooznia, et al., 2007; Segal, et al., 2004). One can quickly note that intersections are often too conservative and unions are anti-conservative, especially when the number of studies increases. Rhode et al. (Rhodes, et al., 2002) was the first to apply Fisher's method to microarray data for a real sense of meta-analysis. They later introduced a weighted form of Fisher's statistic, with the weights determined by the sample size of each study (Ghosh, et al., 2003).

When the studies have a similar design with similar outcomes, combining effect sizes is often preferred to combining significance levels. Choi et al. (Choi, et al., 2003) pointed out that the approach in Rhode et al. "ignored the interstudy variation" and proposed a random effects model under Gaussian assumption. Hu et al. (Hu, et al., 2005) developed a quality measure as weights in the random effects model. For Bayesian approaches, Choi et al. (Choi, et al., 2003) further extended the random effects model to a Bayesian formulation. Similar Bayesian hierarchical models also have been suggested by Tseng et al. (Tseng, et al., 2001) and Conlon et al. (Conlon, et al., 2007) for incorporating different levels of replicate information in cDNA microarray. Conlon et al. (Conlon, et al., 2007) further introduced a Bayesian standardized expression integration model. Shen et al. (Shen, et al., 2004) and Choi et al. (Choi, et al., 2007) proposed a Bayesian mixture model to re-scale and combine data sets.

1.2.4 Two complementary hypothesis settings

for a number of meta-analysis procedures have been discussed in the previous sections. As will be outlined in the following discussion, these statistics are designed to test two complementary hypotheses in general.

Considering the meta-analysis of K gene expression profiling studies, two complementary hypotheses can be defined as:

HS1:

Ho: Gene *g* is not differentially expressed in all k studies (i.e. $\theta_{gk}=0$), $\forall k, k=1,...,K.$

Ha: Gene *g* is differentially expressed in one or some studies (i.e. $\theta_{gk}\neq 0$ for some *k*)

where θ_{gk} denotes the effect size of gene *g* in study *k*. This hypothesis is used to determine which genes are differentially expressed in one or more studies. In many applications, it is also of interest to determine which genes are differentially expressed in all studies. In the latter case, the corresponding hypothesis can be defined as:

HS2:

Ho: Gene *g* is not differentially expressed in one or more studies (i.e. $\theta_{gk}=0$ for some k),

Ha: Gene *g* is differentially expressed in all studies (i.e. $\theta_{gk}\neq 0$ $\forall k, k=1,...,K.$).

Whereas Fisher's , minP and AW statistics are proposed for HS1 problems, maxP and most effect-size models are performed for HS2 problems.

1.2.5 Meta-analysis examples

In this section, an example was given to demonstrate how to apply meta-analysis to microarray studies to identify robust drug-related DE genes and biomarkers by combining two drug response studies on breast cancer cell lines.

1.2.5.1 Cell line's drug response data sets

Liedtke (Liedtke, et al., 2009) and Neve (Neve, et al., 2006) independently measured genome-wide gene expression profiling of breast cancer cell lines using Affymetrix hgu133A platform. Details of both of their data sets are listed in Table 1.1. The raw microarray data files were processed by RMA (Irizarry, et al., 2003), and the data were log2-transformed. Non-specific gene filtering was applied to these data sets using the software package R and Bioconductor (Gentleman, et al., 2004). If x denotes the expression values of probe i, then probes that do not satisfy the following two conditions were filtered out: 1) IQR(x)<0.5; 2) median(x) <log2(100). All probe IDs have been transferred to gene symbols.

The chemosensitivity of the breast cell line to paclitaxel was determined using 50% growth inhibitory concentrations (GI50) data (Liedtke, et al., 2009). According to their chemosensitivity to paclitaxel, the breast cell lines were categorized into two groups: a sensitive group and a resistant group. To calculate the p-values of each gene, the Student's t-test was performed. The maxP statistic was used to combine these two studies. A permutation test was used to evaluate the q-values of genes that, due to the distribution of the maxP statistic, were hard to obtain analytically.

Table 1.1 Summary of drug response data sets.

		Resistant	Sensitive	
Study	Platform	samples	samples	Probe IDs
Liedtke (Liedtke, et al., 2009)	HGU133A	8	8	22,283
Neve (Neve, et al., 2006)	HGU133A	7	8	22,283

1.2.5.2 Details of meta-analysis algorithms

The details of the algorithms that were used to perform the meta-analysis are as follows:

Suppose there are G genes and K studies (K=2 for this case).

I. Individual-study analysis:

a. Compute the Student's t-statistic for the two-group comparison, t_{gk} for gene g and study k

b. Permute the group labels in each study *B* times, and similarly calculate the permuted statistics, $t_{gk}^{(b)}$, where $1 \le g \le G,\ 1 \le k \le K,\ 1 \le b \le B.$

c. Estimate the p-value of t_{gk} as $p_{gk} = \frac{\sum_{b=1}^{B}\sum_{g'=1}^{G} I\left(|t_{g'k}^{(b)}| \ge |t_{gk}|\right)}{B\cdot G}$ and similarly calculate

$$p_{gk}^{(b)} = \frac{\sum_{b'=1}^{B}\sum_{g'=1}^{G} I\left(|t_{g'k}^{(b')}| \ge |t_{gk}^{(b)}|\right)}{B\cdot G}.$$

d. Estimate $\pi_0(k)$, the proportion of non-DE genes, as $\hat{\pi}_0(k) = \frac{\sum_{g=1}^{G} I(p_{gk} \in A)}{G\cdot l(A)}$ (Storey, 2002). We chose A=[0.5, 1] and thus *l*(*A*)=0.5.

e. Estimate the q-value of t_{gk} as $q_{gk} = \frac{\hat{\pi}_0(k)\cdot\sum_{b=1}^{B}\sum_{g'=1}^{G} I\left(|t_{g'k}^{(b)}| \ge |t_{gk}|\right)}{B\cdot\sum_{g'=1}^{G} I(|t_{g'k}| \ge |t_{gk}|)}$. DE genes detected from each individual study are denoted by $G_k = \{g : q_{gk} \le 0.05\}$.

II. Meta-analysis:

a. The maximum p-value statistic (maxP) is used for meta-analysis: $V_g = \max_{1 \le k \le K} p_{gk}$.

Define $V_g^{(b)} = \max_{1 \le k \le K} p_{gk}^{(b)}$.

b. Estimate the p-value of the genes in meta-analysis as $p(V_g) = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(V_{g'}^{(b)} \le V_g\right)}{B \cdot G}$.

c. Estimate π_0, the proportion of non-DE genes in the meta-analysis, as

$\hat{\pi}_0 = \frac{\sum_{g=1}^{G} I(p(V_g) \in A)}{G \cdot l(A)}$. We chose A=[0.5, 1] and thus *l*(*A*)=0.5.

d. Estimate the q-value in the meta-analysis as $q(V_g) = \frac{\hat{\pi}_0 \cdot \sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(V_{g'}^{(b)} \le V_g\right)}{B \cdot \sum_{g'=1}^{G} I(V_{g'} \le V_g)}$. DE

genes detected by the meta-analysis are denoted as $G_{meta} = \{g : q(V_g) \le 0.05\}$.

1.2.5.3 Meta-analysis results

The meta-analysis results are shown in Figure 1.1. For each individual study, 252 and 594 DE genes were identified in the Liedtke and Neve studies, respectively. Using meta-analysis with the maxP statistic, 956 genes were considered to be DE genes. The meta-analysis failed to identify 47 DE genes from the Liedtke study and 143 DE genes from the Neve study (Region VI and region VII in Figure 1.1). This can be explained by the fact that the expression patterns of these genes were not consistent between the Liedtke and Neve studies (the difference in the p-values of these genes was large). Meta-analysis identified 420 DE genes which were not discovered in individual studies.

By checking the literature, we found some DE genes, such as CD44, MSN, and TGFBR2 are related to the cell line subtype and the drug response (Neve, et al., 2006). However, the large number of DE genes makes it hard to consider them individually. Novel methods, referred to as gene set enrichment analysis or pathway enrichment analysis, have been proposed for the analysis of a gene set, rather than individual genes. These methods are reviewed in the next chapter.

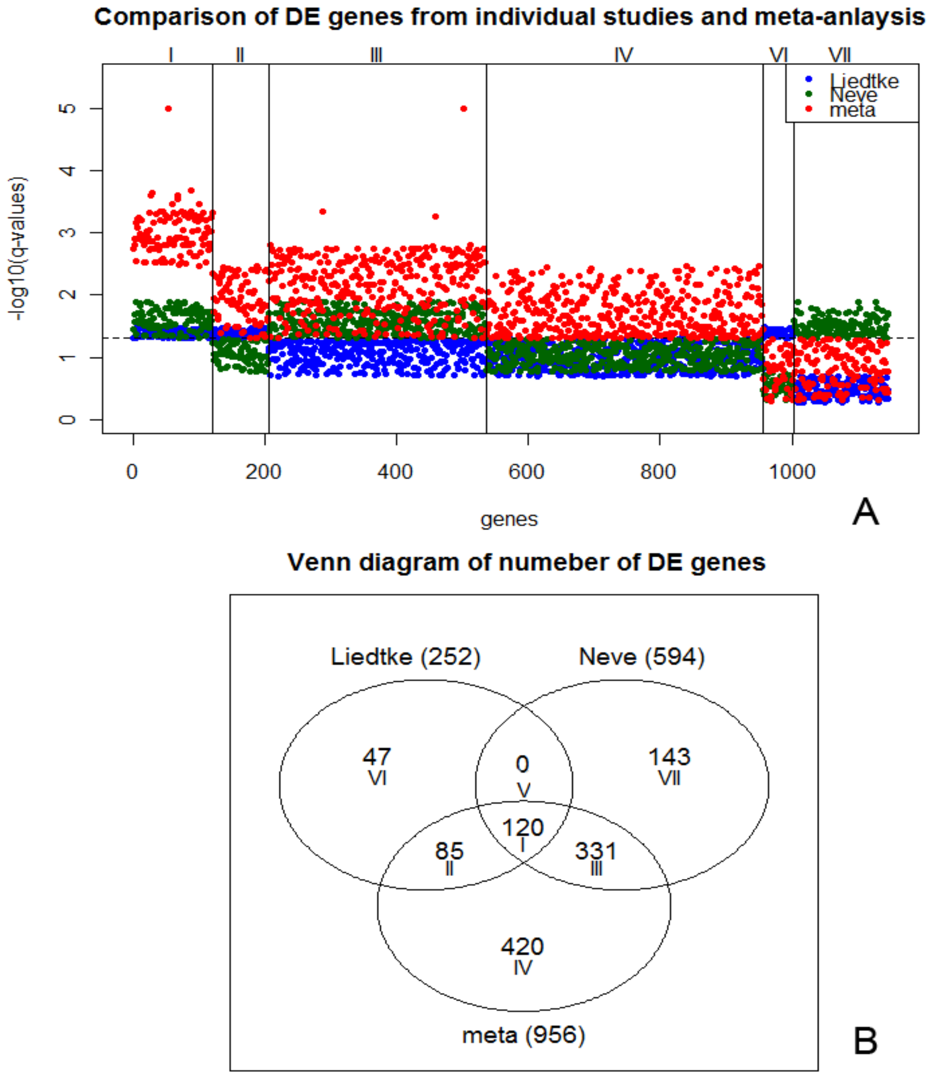


Figure 1.1 Meta-analysis of drug response studies.
In the upper panel, the solid red, green, and dark blue circles represent the –log transformation of q-values of meta-analysis for the Neve and Liedtke studies. The Figure has been divided into seven regions. Region I contains the DE genes that were identified by both individual studies and by meta-analysis. Region II contains DE genes that were identified by the Liedtke study and meta-analysis, but not by the Neve study. Region III contains DE genes that were identified by the Neve study and by meta-analysis, but not by the Liedtke study. Region IV contains DE genes that were identified by meta-analysis, but not by either one of the individual studies. Region V contains DE genes that were identified by the individual studies, but not by meta-analysis. Region VI contains DE genes that were identified by meta-analysis and the Liedtke study, but not by the Neve study. Region VII contains DE genes that were identified by meta-analysis and the Neve study, but not by the Liedtke study. The lower panel shows the Venn diagram of the number of DE genes that were identified by meta-analysis and by the individual studies.

1.3 PATHWAY ENRICHMENT ANALYSIS

In section 1.2, meta-analysis methods that combine gene expression information across studies were reviewed. Gene expression information can be also integrated within a study. Specifically, instead of studying each gene individually, we can also study a gene set. A gene set is a pre-defined set of genes that may have similar locations or functions or form a particular pathway. If genes in a gene set act in concert, this gene set may have important biological effects on the phenotype of concern (Subramanian, et al., 2005). Thus, it is important to test whether a set of genes is coherently associated with the phenotype of interest. This type of analysis is called gene set enrichment analysis or pathway enrichment analysis (Newton, et al., 2007; Subramanian, et al., 2005; Tian, et al., 2005). When gene sets are defined by biological pathways, the term gene set enrichment analysis and pathway enrichment analysis are interchangeable. The common gene set/pathway databases include KEGG, Biocarta, and the gene ontology (GO) databases (Gene Ontology Consortium, 2006; Kanehisa and Goto, 2000). The molecular signatures database (MsigDB) (Subramanian, et al., 2005) is a collection of gene sets (including KEGG, Biocarta and GO) that has five major categories; these are C1: positional gene sets; C2: curated gene sets; C3: motif gene sets; C4: computational gene sets and C5: GO gene sets. The C2 collection contains two sub-categories: canonical pathways (CP) and gene sets that represent gene expression signatures of genetic and chemical perturbations (CGP). Based on the MsigDB version 2.5, CP contains 639 gene sets and CGP contains 1186. In this dissertation, CP and CGP gene set databases were used as our pre-defined gene sets. As CP and CGP are both pathway-related gene sets, we use the term pathway enrichment analysis hereafter. Unless specified otherwise, the C2 collection was used as our pathway database.

Figure 1.2 shows a general diagram for pathway enrichment analysis in an individual microarray study. Suppose a data matrix $\{x_{gs}\}$ ($1 \le g \le G$, $1 \le s \le S$) represents the gene expression intensity of gene g and sample s. Let $\{y_s\}$ ($1 \le s \le S$) represent the phenotype label for sample s, where y_s stands for microarray designs including 1) $y_s \in \{0,1\}$ (two groups comparison); 2) $y_s \in \{0,1,2,\ldots,J\}$ (multiple groups comparison); 3) $y_s \in R$ (time series studies); 4) $y_s \in \{ t_s, c_s\}$(survival analysis; t_s : survival time; c_s: censoring status). For simplicity, we assume that y_s is binary (e.g. 0 represents normal patients and 1 represents tumor patients unless otherwise stated). A pathway database matrix $\{z_{gp}\}$ ($1 \le g \le G$, $1 \le p \le P$) represents the pathway information of P pathways, where z_{gp}=1 when gene g belongs to pathway p and z_{gp}=0 otherwise. The pathway enrichment analysis has two main steps as follows:

Step I. The association scores with phenotype in each gene g are first calculated as $\mathrm{t_g}$, where $\mathrm{t_g}$ can either be Student's t-statistics or one of its variations, such as the moderated t-statistic (Smyth, 2004). Correlations between gene expression values and phenotype can also be used as the association scores.

Step II. The pathway enrichment evidence score v_p is calculated for each pathway p. This is the key step in pathway enrichment analysis. The pathway enrichment evidence score is used to summarize the association scores of all genes in the pathway. Either non-parametric statistics (e.g. Kolmogorov-Smirnov (KS) statistic) or parametric statistics (e.g. mean of t-statistics) can be used to summarize the association scores.

In the following section, we give a brief review of three most commonly used pathway enrichment methods.

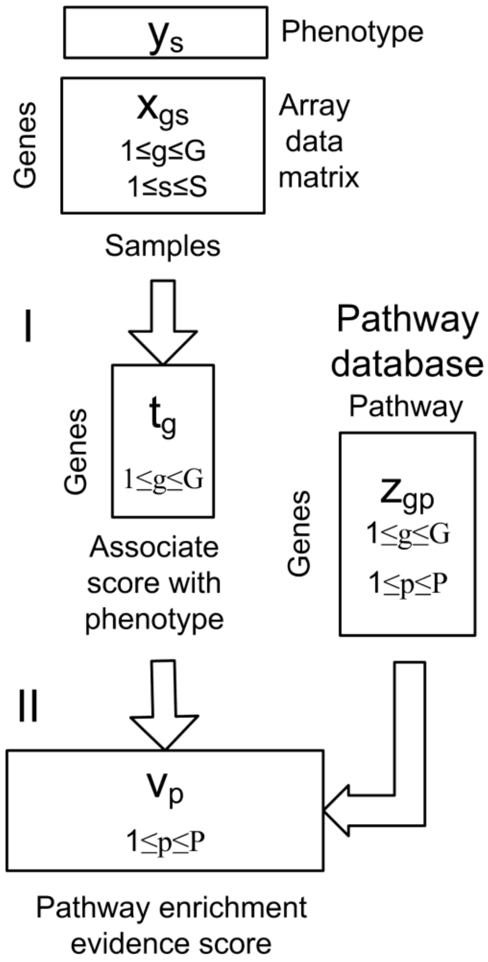


Figure 1.2 Diagram of pathway enrichment analysis.

1.3.1 Fisher's exact test method

The Fisher's exact test method has been widely used in pathway enrichment analysis as a result of its simplicity (Berriz, et al., 2003; Dahlquist, et al., 2002; Draghici, et al., 2003; Zeeberg, et al., 2003; Zhong, et al., 2003). The purpose for Fisher's exact test in this study was to determine whether the ratio of DE genes in a gene set was higher than the ratio outside of the pathway. If the ratio was higher than would be expected by chance, the pathway was referred to as an enriched pathway. The first step in Fisher's exact test method was to identify DE genes, as shown in Step I in Figure 1.2. The number of DE genes both inside and outside of the pathway was then counted as a 2x2 contingency Table (Table 1.2). The p-value for enrichment of a pathway was calculated by testing the independence of the 2x2 contingency Table using Fisher's exact test. The null and alternative hypothesis for the Fisher's exact test is: H_0: $\theta_1 = \theta_2$ and H_1: $\theta_1 > \theta_2$, where θ_1 and θ_2 are the probability of DE genes inside and outside of the pathway. The observed numbers of DE genes inside and outside of pathways are n_{pd} and $n_{p^c d}$ respectively (shown in 1.2). Under the null hypothesis, the conditional distribution of n_{pd} given the marginal totals is the hypergeometric distribution,

$$\frac{\begin{pmatrix} n_p \\ n_{pd} \end{pmatrix}\begin{pmatrix} n_p{}^c \\ n_p{}^c{}_d \end{pmatrix}}{\begin{pmatrix} N \\ n_d \end{pmatrix}}$$

where N, n_d and n_p are fixed numbers. Let N_{pd} and N_d denote the random variables for the observed value n_{pd} and $n_{d.}$ The null hypothesis is rejected when N_{pd} is larger than the critical values. The exact p-value is $P(N_{pd>}\ n_{pd} \mid N_d = n_d)$, which can be calculated from all possible 2 by 2 Tables which have the same marginal totals as the observed one, but having a value of N_{pd} more extreme than n_{pd} (Mehta, et al., 1984).

Table 1.2 2x2 Table for enrichment analysis.

	DE genes	non-DE genes	Total
In the pathway	n_{pd}	n_{pd^c}	n_p
Not in the pathway	n_{p^cd}	$n_{p^cd^c}$	n_{p^c}
Total	n_d	n_{d^c}	N

Though Fisher's exact test method is widely used, its shortcomings are obvious. First, by dividing genes into two categories (DE genes and non-DE genes), it loses information by only counting the number of DE and non-DE genes instead of considering the order of the genes or their p-values. In addition, the selection of the p-value cutoff that is used to define DE and non-DE genes, is always ad-hoc. The shortcomings of Fisher's exact test method can be overcome by the use of a couple of methods. For example, the average t-statistics of genes in a pathway *p* can be used to summarize the gene expression information; this method is outlined in the following section.

1.3.2 Averaging association score method

Let T_p denote the average of t-statistics of all genes in the pathway p, then:

$$T_p = \sum_{g=1}^{G} z_{gp} t_g \Big/ \sum_{g=1}^{G} z_{gp} \, , \quad .$$

where $1 \leq p \leq P$. As there is some difficulty in obtaining the distribution of T_p analytically, a permutation test was applied to obtain the p-value of $T_{p.}$ This method was proposed and discussed in detail by Tian et al. (Tian, et al., 2005). Efron and Tibshirani (Efron and Tibshirani, 2007) provided an improved method, that involved introducing max-mean statistics and a re-standardization procedure.

1.3.3 KS test method

Let *A* and *B* denote the p-values of genes from inside and outside the pathway p, respectively, in which there are *m* genes in the pathway *p* and *n* genes outside of the pathway p. The order statistics for *A* and *B* are: $A_{(1)}$, $A_{(2)}$, ... , $A_{(m)}$ *and* $B_{(1)}$, $B_{(2)}$, ... , $B_{(n)}$. The corresponding empirical distribution functions, $\hat{F}_A(x)$ and $\hat{F}_B(x)$ for *A* and *B*, can be defined as follows:

$$\hat{F}_A(x) = \begin{cases} 0 & \text{if } x < A_{(1)} \\ s/m & \text{if } \mathrm{A}_{(s)} \le x < A_{(s+1)} \text{ for } s = 1, 2, ..., m-1 \\ 1 & \text{if } x \ge A_{(m)} \end{cases}$$

and

$$\hat{F}_B(x) = \begin{cases} 0 & \text{if } x < B_{(1)} \\ s/n & \text{if } \mathrm{B}_{(s)} \le x < X_{(s+1)} \text{ for s} = 1, 2, ..., n-1 \\ 1 & \text{if } x \ge B_{(n)} \end{cases}$$

Let F_A and F_B denote the population distribution for *A* and *B*, respectively. The one-sided two sample KS test can be defined based on the formula:

$$KS = \max_x [F_A(x) - F_B(x)],$$

in which the null hypothesis and the alternative hypothesis are:

$$
\begin{aligned}
&H_0 : F_A(x) = F_B(x) \text{ for all x} \\
&H_1 : F_A(x) \geq F_B(x) \text{ for all x} \\
&\quad\;\; F_A(x) > F_B(x) \text{ for some x}
\end{aligned}
$$

The rejection region can be $KS \geq C_\alpha$

where

$$P(D_{i,j} \geq c_\alpha \mid H_o) \leq \alpha \,.$$

Rejection of H_0 means that *A* is stochastically less than *B* (the CDF of *A* lies above and hence to the left of that for *B*). In another words, the p-values of genes in the pathway p are stochastically less than the p-values of genes outside of pathway p. This indicates that genes in the pathway p have a stronger association with phenotype than genes from outside of the pathway p; thus, the pathway p is of interest. The computational method for calculating $P(KS \geq c_\alpha \mid H_o)$ is provided by Marsaglia et al (Marsaglia, et al., 2003). The KS test method was first applied to gene set enrichment analysis by Subramanian et al (Subramanian, et al., 2005). They also introduced a weighted KS test method and provided the software package GSEA.

1.3.4 Control of false discovery rate and evaluation of q-values

We have reviewed three widely used methods for calculating the pathway enrichment evidence score and its p-value. Considering that the null distribution of the pathway enrichment evidence score is difficult to obtain analytically, a permutation test is typically applied to control the false discovery rate and evaluate the q-value of the pathway. Two basic permutation procedures, sample-wise permutation and gene-wise permutation, have been proposed. These are based on two related, but not equivalent, null hypotheses (Q1 and Q2, respectively) as follows:

Q1: the genes in a gene set have the same pattern associated with the phenotype of interest as the genes outside of the gene set.

Q2: no genes in the gene set have expression patterns associated with the phenotype.

Details about these two null hypotheses are discussed by Tian (Tian, et al., 2005), Geoman (Goeman and Buhlmann, 2007) and Efron (Efron and Tibshirani, 2007). Briefly, Q1 takes the background information (the expression of genes outside of the pathway) into consideration, whereas Q2 does not.

Both of these permutation strategies can work with all three of the aforementioned pathway enrichment methods to evaluate the q-values of pathways. Normally, the false discovery rate is controlled at 5% (this means that among detected pathways, on average 5% are false discoveries). For further investigation, all pathways with a q-value less than 5% are reported as enriched pathways (i.e. $\{p: q(v_p) \leq 5\%\}$).

Reviews and method comparisons of pathway enrichment analyses are available at (Ackermann and Strimmer, 2009; Dorum, et al., 2009; Khatri and Draghici, 2005; Nam and Kim, 2008; Tomfohr, et al., 2005). Our MAPE procedures provided a general statistical framework for performing meta-analysis on pathways. Most of the meta-analysis and enrichment analysis methods could be adopted into our framework. For simplicity, we used the KS test method to demonstrate our MAPE procedures.

1.3.5 Examples of pathway enrichment analysis

Here, we give an example of pathway enrichment analysis for the breast cancer patient's chemotherapy data sets.

1.3.5.1 Breast cancer patient’s chemotherapy data sets

Breast cancer patient’s chemotherapy data sets were provided by Hess et al (Hess, et al., 2006). Tordai et al (Tordai, et al., 2008) have performed pathway enrichment analysis on Hess data using GSEA. To illustrate the advantage of pathway enrichment analysis, we re-analyzed Hess using a slightly different method.

Hess data included 51 estrogen receptor (ER) positive and 82 ER negative breast cancer patients. Before chemotherapy treatment, a fine-needle aspiration biopsy of the cancer was taken from each patient. These needle aspiration samples were prepared for microarray analysis using Affymetrix platform HGU133A. All patients were treated with paclitaxel, followed by 5-fluorouracil, doxorubicin, and cyclophosphamide (TFAC) for a period of six months. After completion of chemotherapy, the pathologic complete response (pCR) of each patient was tested. There are 7 pCR patients in the ER+ group and 27 patients in the ER- group. Because ER+ and ER- patients suffer from two different sub-types of breast cancer, pathway enrichment analysis should be applied to ER+ and ER- patients separately. Our example includes only ER+ patients. Microarray data were pre-processed according to the same procedure as in section 1.3.5. C2 collection of MsigDB (Subramanian, et al., 2005) was used as our pathway database.

1.3.5.2 DE gene analysis

To identify DE genes in the pCR patients and the non-pCR patients, we first performed an unequal variance Student’s t-test. P-values of the genes were adjusted for simultaneous inference using the Benjamini & Hochberg method (Benjamini and Hochberg, 1995). When the adjusted p-value cutoff was set as 0.05, no DE genes were identified. This result is consistent with the findings of Tordai, who applied the beta-uniform mixture (BUM) method to control the FDR (Tordai, et al., 2008).

Although the t-test failed to identify DE genes, it does not follow that there were no real transcriptional difference between pCR patients and non-pCR patients. A set of related genes acting in concert could have a significant effect, even if there was no statistical difference in single genes between both sets of patients. This situation has been discussed in (Subramanian, et al., 2005). For the present chemotherapy study, our pathway enrichment analysis did identify multiple important pathways.

1.3.5.3 Algorithm details

Details of the pathway enrichment algorithm are as follows:

1. Calculate $p(t_g)$, the p-value of gene *g* by Student's t-test, 1≤*g*≤*G*.

2. Compute P_p^{KS}, the p-value of pathway *p*, by one-sided KS test (details in section 1.3.3.)

3. Permute gene labels C times, and calculate the permuted statistics, $P_p^{KS(c)}$, $1 \le c \le C$.

4. Estimate the p-value of pathway *p* as $p(v_p) = \sum_{c=1}^{C}\sum_{p'=1}^{P} I(P_{p'}^{KS(c)} \le P_p^{KS}) \Big/ C \cdot P$ and similarly calculate $v_p^c = \sum_{c'=1}^{C}\sum_{p'=1}^{P} I(P_{p'}^{KS(c')} \le P_p^{KS(c)}) \Big/ C \cdot P$.

5. Estimate π_0, the proportion of non-enriched pathways in the meta-analysis, as $\hat{\pi}_0 = \dfrac{\sum_{p=1}^{P} I(p(v_p) \in A)}{P \cdot l(A)}$. We chose A=[0.5, 1] and thus *l*(*A*)=0.5.

6. Estimate q-value of pathway *p* as

$$q(v_p) = \hat{\pi}_0 \sum_{c=1}^{C}\sum_{p'=1}^{P} I(P_{p'}^{KS(c)} \le P_p^{KS}) \Big/ C \cdot \sum_{p'=1}^{P} I(P_{p'}^{KS} \le P_p^{KS}) \ .$$

1.3.5.4 Pathway enrichment analysis results

Table 1.3 Pathway enrichment analysis for Hess data.

Pathways	Q-values
ZHAN_MM_CD138_PR_VS_REST	0.000
HOFFMANN_BIVSBII_BI_TABLE2	0.000
LEE_TCELLS3_UP	0.000
DOX_RESIST_GASTRIC_UP	0.000
CANCER_UNDIFFERENTIATED_META_UP	0.000
IDX_TSA_UP_CLUSTER3	0.000
BRCA_ER_POS	0.000
ADIP_DIFF_CLUSTER5	0.000
SERUM_FIBROBLAST_CELLCYCLE	0.000
CMV_IE86_UP	0.000
YU_CMYC_UP	0.000
GREENBAUM_E2A_UP	0.000
VERNELL_PRB_CLSTR1	0.000
LE_MYELIN_UP	0.001
OLDAGE_DN	0.002
IRITANI_ADPROX_LYMPH	0.002
CROONQUIST_IL6_STARVE_UP	0.007
CELL_CYCLE	0.007
LI_FETAL_VS_WT_KIDNEY_DN	0.009
P21_ANY_DN	0.014
CELL_CYCLE_KEGG	0.016
BRCA_PROGNOSIS_NEG	0.017
BRENTANI_CELL_CYCLE	0.017
SMITH_HCV_INDUCED_HCC_UP	0.019
HG_PROGERIA_DN	0.026
FLECHNER_KIDNEY_TRANSPLANT_REJECTION_DN	0.026
RUIZ_TENASCIN_TARGETS	0.027
PARP_KO_UP	0.028
SASAKI_TCELL_LYMPHOMA_VS_CD4_UP	0.035
SASAKI_ATL_UP	0.035
VANTVEER_BREAST_OUTCOME_GOOD_VS_POOR_DN	0.036
VANTVEER_BREAST_OUTCOME_GOOD_VS_POOR_UP	0.036
BRCA_PROGNOSIS_POS	0.037
FRASOR_ER_UP	0.038

GOLDRATH_CELLCYCLE	0.038
GAY_YY1_DN	0.039
UVC_TTD_4HR_DN	0.040
SHEPARD_CRASH_AND_BURN_MUT_VS_WT_DN	0.041
HSA00640_PROPANOATE_METABOLISM	0.042
BREAST_DUCTAL_CARCINOMA_GENES	0.042
TAVOR_CEBP_UP	0.046
P21_P53_ANY_DN	0.047

Results of the pathway enrichment analysis are listed in Table 1.3. In our analysis, a total of 42 enriched pathways were identified using the KS test. These pathways are predominately related to cell cycle, cell proliferation, oncogenic pathways and the estrogen receptor-associated gene set. Noticeably, our results indicate that some important oncogenic pathways related to P53 (P21_P53_ANY_DN), MYC (YU_CMYC_UP) may be highly correlated to the chemotherapy response. The most interesting enriched pathway that we detected was the gene module related to doxrubicin resistance in gastric cancer cell lines (DOX_RESIST_GASTRIC_UP). This indicates that there are some common mechanisms for drug response across different tumor types.

2.0 META-ANALYSIS FOR PATHWAY ENRICHMENT ANALYSIS (MAPE)

2.1 MAPE METHODS

In this chapter, we first present the rationale, general framework, and analysis flow charts of two meta-analysis approaches for pathway enrichment: MAPE_G and MAPE_P. We show two example pathways from lung cancer data to demonstrate the complementary advantages of the two methods. Finally, we introduce a simple integrated approach, MAPE_I, to incorporate the advantages of both methods. We then discuss and outline the implementation details.

2.1.1 Framework of MAPE_G and MAPE_P

When combining multiple studies, we assume genes in multiple studies are matched and no missing value exists. Denote by $\{x_{kgs}\}$ $(1\leq k\leq K, 1\leq g\leq G, 1\leq s\leq S_k)$ the expression intensity of gene g and sample s in study k. $\{y_{ks}\}$ $(1\leq k\leq K, 1\leq s\leq S_k)$ and $y_{ks}\in\{0,1\}$ represents the phenotype label for sample s in study k. Figure 2.1A shows the procedure for the MAPE_G method. In Step I, the association scores with phenotype are calculated in each study (i.e. $\{t_{kg}\}$ $(1\leq g\leq G)$). In Step II meta-analysis is performed for biomarker detection and produces a new association score after meta-analysis at the gene level (i.e. $\{u_g\}$ $(1\leq g\leq G)$). In Step III, the pathway enrichment analysis is performed as in Step II in Figure 1.2.The evidence scores $\{v_p\}$, corresponding q-values $\{q(v_p)\}$

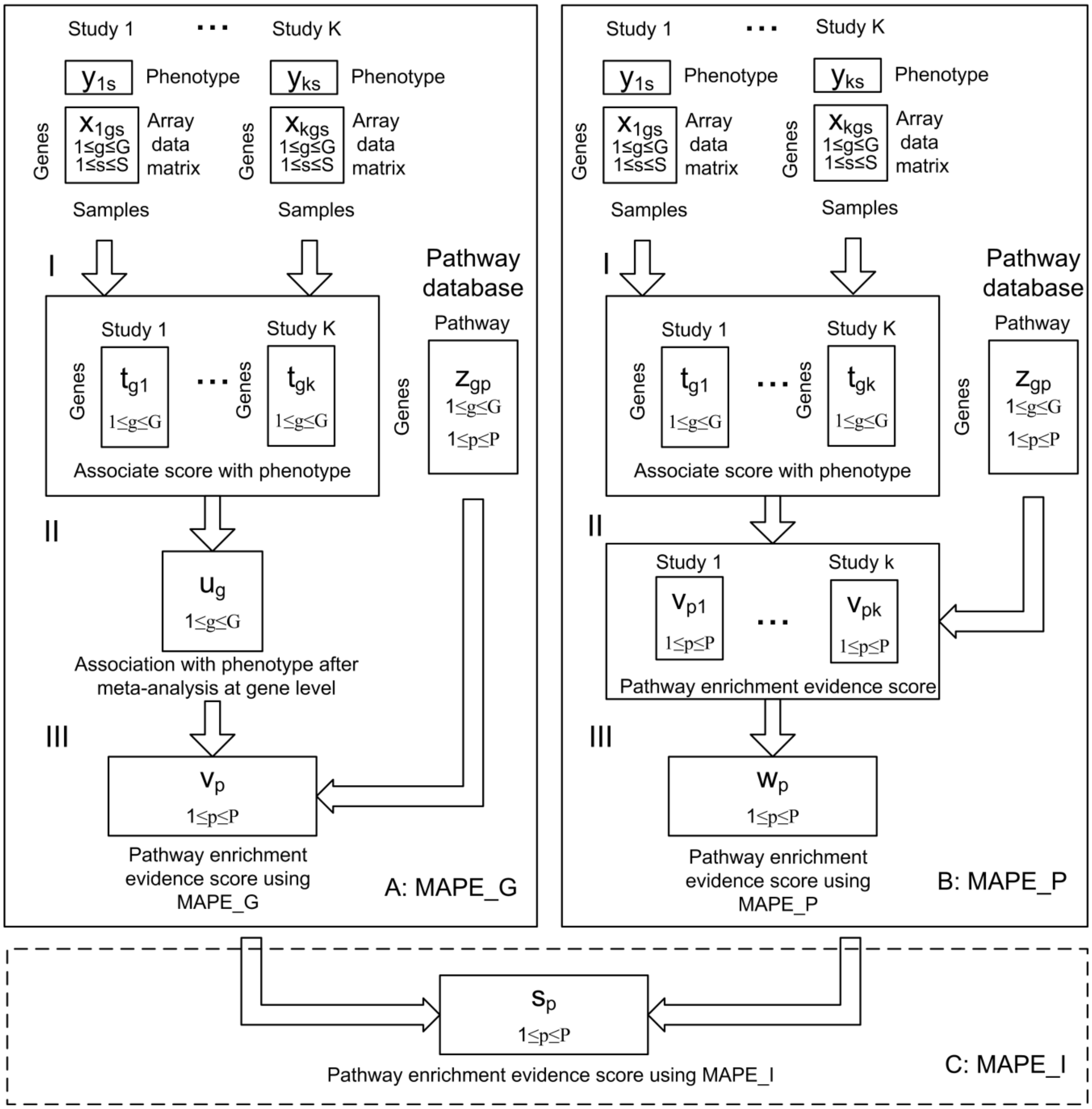


Figure 2.1 The diagram for MAPE_G, MAPE_P, and MAPE_I procedures.

and a list of enriched pathways are then generated. This method can be viewed as a natural combination of meta-analysis for biomarker detection (Step I and II) and pathway enrichment analysis (Step III) in a sequential manner. Rhodes (Rhodes, et al., 2002) has implicitly performed A similar analysis by queried DE genes obtained by meta-analysis in the KEGG database

(Kanehisa and Goto, 2000). For MAPE_G proposed in this study, we replaced the two-stage separated procedures with a unified evaluation of permutation test.

In Figure 2.1B, an alternative procedure for MAPE_P is shown. The Step I of association scores for each study is identical to that in MAPE_G. In Step II, instead of meta-analysis at the gene level, we performed pathway enrichment analysis in each individual study to obtain the study-wise pathway enrichment evidence scores: $\{v_{kp}\}$ ($1 \le k \le K$, $1 \le p \le P$). The meta-analysis on the pathway level was then performed in Step III to assess the combined evidence score and q-values (i.e. (Kuo, et al.) and $\{q(w_p)\}$ ($1 \le p \le P$)).

2.1.2 Complementary advantages of MAPE_G vs. MAPE_P

MAPE_P has an important advantage in that the genes across multiple studies need not be matched to perform meta-analysis as in MAPE_G (Step II of Figure 2.1A). Specifically, we can relax data in Figure 2.1B to $\{x_{kgs}\}$ ($1 \le k \le K$, $1 \le g \le G_k$, $1 \le s \le S_k$) and $\{t_{kg}\}$ ($1 \le g \le G_k$) so that different studies may have a different number of genes and the genes are not matched across studies. The gene matching issue is particularly significant when studies from different microarray platforms are combined. Supplemental Table 1 shows summary statistics of two lung cancer studies that were combined. The Bhat study used the Affymetrix U95A platform and the Beer study used Affymetrix HG6800. Only 5,515 Entrez genes overlapped across the two studies and the MAPE_G method had to drop information from 3,490 out of 9,005 genes that appear in Bhat but not in Beer. When more studies of different array platforms are included, the number of overlapping genes will decrease dramatically. Published studies have also demonstrated weak consistency across studies at the gene level but increased consistency at the pathway level. In general, then, MAPE_P seems to be preferable to MAPE_G.

When we analyzed a combination of two lung cancer studies, however, we identified some examples with better power by MAPE_P and others with MAPE_G. Figure 2.2 shows two example pathways of ALCALAY_AML_NPMC_UP (AANU; genes with increased expression in acute myeloid leukemia bearing cytoplasmic nucleophosmin) and HDACI_COLON_TSABUT_UP (HCTU; genes up-regulated by both butyrate and trichostatin A at any time point up to 48 hrs in SW260 colon carcinoma cells), based on the C2 collection of MsigDB. AANU was identified as an enriched pathway by MAPE_P but not by MAPE_G (Figure 2.2 A and B). In contrast, HCTU was identified by MAPE_G but not by MAPE_P (Figure 2.2 C and D). We performed differential expression analysis by SAM in each study separately (FDR=5%) and found that only 13 genes were identified as DE genes in both studies in the AANU pathway. Thirteen genes were DE in Beer but not in Bhat, and 27 genes were DE in Bhat but not in Beer. We defined a simple concordance index (CI) as the ratio of common DE genes in both studies versus DE genes in at least one of the two studies. The AANU pathway was detected by MAPE_P but not by MAPE_G because the CI is as low as 13/(13+13+27)=0.245. When we pursued meta-analysis at the gene level, very few genes were significant in Step II of Figure 2.2A although the meta-analysis at the pathway level in Step III of Figure 2.2B is quite significant. On the other hand, the high CI in the HCTU pathway (CI=13/(13+1+9)=0.565) increased the statistical power of MAPE_G while MAPE_P did not have enough power to detect this pathway. Such high CI pathways detected only by MAPE_G are usually important because the biomarkers are repeatedly identified in multiple studies. From the two examples above, we conclude that although intuitively MAPE_P has the convenience of not having to match genes across studies, MAPE_G has an advantage in particular situations.

Based on this finding, we developed a simulation scheme (shown in the Results Section) to illustrate conditions when MAPE_G outperforms MAPE_P and vice versa.

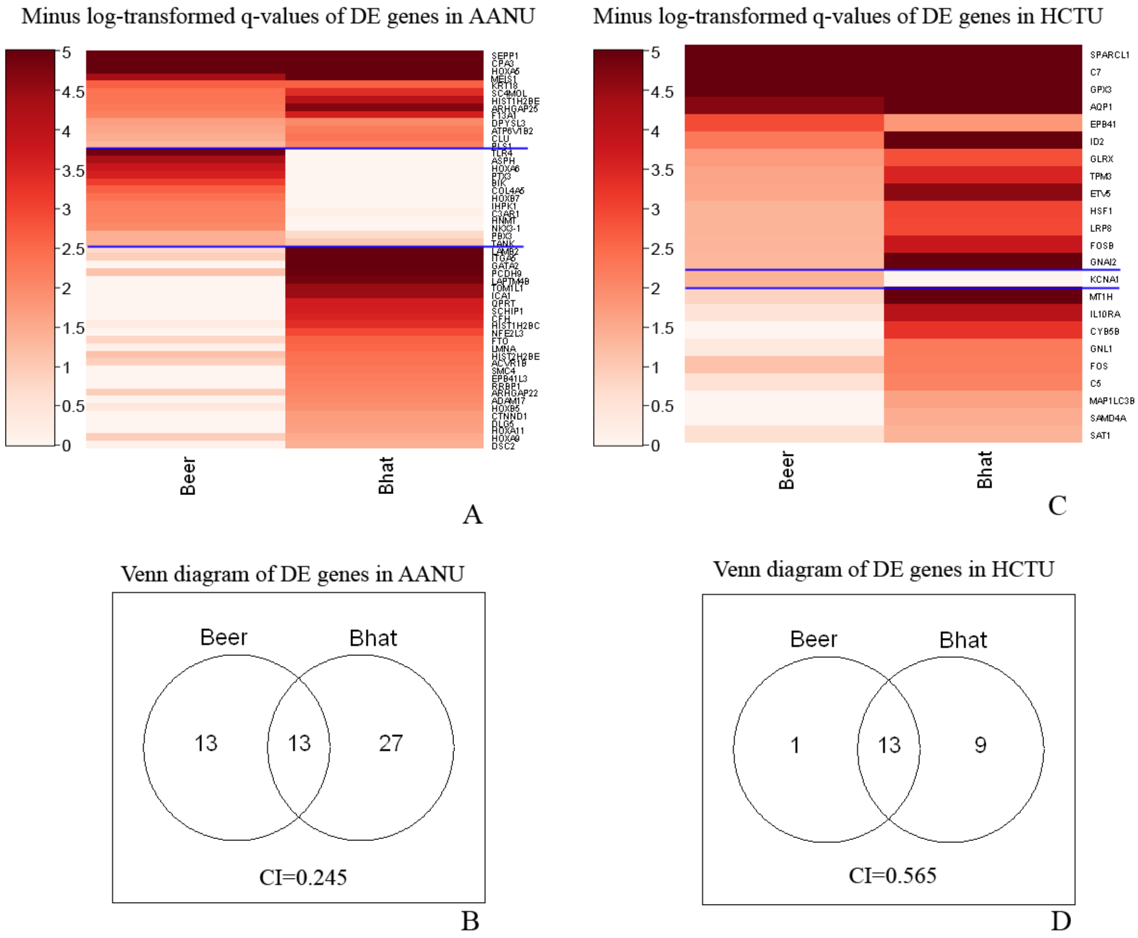


Figure 2.2. Examples of two pathways identified by MAPE_P and MAPE_G in lung cancer studies.

AANU is detected by MAPE_P but not by MAPE_G whereas HCTU is detected by MAPE_G but not MAPE_P. A and C: The heatmaps display log-transformed (base 10) q-values by gradient color. B and D: Venn diagram of biomarkers detected by each individual study (Beer and Bhat).

2.1.3 Framework of MAPE_I

Since pathways detected by both MAPE_G and MAPE_P are of biological interest, we propose a simple integrative method, namely MAPE_I, to incorporate the complementary advantages of both methods (Figure 2.1C). Specifically, we used a minP statistic that takes the minimum p-value from MAPE_G and MAPE_P for each pathway. The statistical inference and control of FDR were similarly performed by permutation analysis.

2.1.4 Implementation strategy

Numerous pathway analysis and meta-analysis methods for microarray data have been described. Most of these methods have pros and cons under different conditions and for different biological goals. Under the general framework shown in Figure 2.1 for MAPE_G, MAPE_P and MAPE_I, we can virtually apply and combine any pathway analysis and meta-analysis method for implementation. There are four major considerations or choices in practice: A. statistics used for association evidence with phenotype (i.e. t_{gk}); B. statistics used for meta-analysis at the gene level (Step II in Figure 2.1A) or the pathway level (Step III in Figure 2.2B); C: statistics used in pathway enrichment analysis (step III in Figure 2.2A and step II in Figure 2.2B); D. permutation test used for statistical inference and FDR control.

A. Statistic selection for association evidence with phenotype: For simplicity, but without loss of generality, we considered t-statistics for a binary phenotype label. For multi-class, continuous, or censored survival phenotype, different test statistics, such as F-statistics, Pearson correlation measure, or statistics from the Cox proportional hazard model, may be used respectively.

B. Statistic selection for meta-analysis: Various meta-analysis statistics, including Fisher's statistic, minimum p-value statistic (minP), and maximum p-value statistic (maxP), have been discussed in the Introduction Section. The best choice of meta-analysis statistic depends on the particular biological goal of interest. Following the convention of Birnbaum (Birnbaum, 1954), two different hypothesis settings may be considered:

$$HS1:\left\{H_0:\text{at least one}\,\theta_{kg}=0,1\le k\le K\ \text{versus}\,H_A:\theta_{kg}\ne 0,\forall 1\le k\le K\right\}$$

$$HS2:\left\{H_0:\theta_{1g}=\cdots=\theta_{Kg}=0\ \text{versus}\,H_A:\text{at least one}\,\theta_{kg}\ne 0,1\le k\le K\right\},$$

where θ_{kg} represents the effect size of gene *g* in study *k*. HS1 corresponds to the biological question: "which genes are consistently differentially expressed in all studies?". In contrast, HS2 detects genes if they are differentially expressed in one or more studies. It can be seen that maxP corresponds to HS1, and Fisher's statistic and minP correspond to HS2. In this paper, we focus on the conservative maxP statistic to identify consistent biomarkers across all microarray studies. Specifically, we will calculate the p-values of evidence scores at the gene level in Step II of Figure 2.1A or at the pathway level in Step III of Figure 2.1B. The maxP statistic for meta-analysis at the gene level is $u_g=\max_{1\le k\le K}p(t_{kg})$ and the pathway level is $w_p=\max_{1\le k\le K}p(v_{kp})$.

C. Statistic selection for the pathway enrichment analysis method: The goal of pathway analysis is to test whether genes in a pathway are coherently associated with the phenotype of interest. Here we demonstrate our MAPE procedures by using the KS test. Any gene set analysis method described above can be adopted into our general framework depicted in Figure 2.1.

D. Control of false discovery and evaluation of q-values: The p-values and q-values of pathway enrichment evidence scores are usually computed by permutation test, considering that the null distribution of gene set statistics is difficult to obtain analytically.

2.1.5 Algorithms details

Algorithms for all three MAPE methods are listed in the following sections.

2.1.5.1 Algorithms for MAPE_P

The basic procedure of MAPE_P is to first calculate the p-value of each pathway in each study. Then, combine the p-values of the pathways across studies by maxP statistics.

I. Pathway enrichment analysis:

1. For each study *k*, calculate $p(t_{gk})$, the p-value of gene *g*, by Student t-test, $1 \le g \le G$.
2. Given a pathway *p*, compute the KS statistic v_{pk} that compares the p-values ($p(t_{gk})$) inside and outside the pathway.
3. Permute gene labels B times, and calculate the permuted statistics, $v_{pk}^{(b)}$, $1 \le b \le B$.
4. Estimate the p-value of KS statistic in pathway *p* and study *k* as $p(v_{pk}) = \sum_{b=1}^{B} \sum_{p'=1}^{P} I(v_{p'k}^{(b)} \ge v_{pk}) \Big/ B \cdot P$ and similarly calculate $p(v_{pk}^{(b)}) = \sum_{b'=1}^{B} \sum_{p'=1}^{P} I(v_{p'k}^{(b')} \ge v_{pk}^{(b)}) \Big/ B \cdot P$.

II. Meta-analysis:

1. The maximum p-value statistic (maxP) is applied for meta-analysis: $w_p = \max_{1 \le k \le K} p(v_{pk})$ and $w_p^{(b)} = \max_{1 \le k \le K} p(v_{pk}^{(b)})$.
2. Estimate p-value of pathway *p* as $p(w_p) = \sum_{b=1}^{B} \sum_{p'=1}^{P} I(w_{p'}^{(b)} \le w_p) \Big/ B \cdot P$. Similarly $p(w_p^{(b)}) = \sum_{b'=1}^{B} \sum_{p'=1}^{P} I(w_{p'}^{(b')} \le w_p^{(b)}) \Big/ B \cdot P$

3. Estimate π_0, the proportion of non-enriched pathways in the meta-analysis, as

$\hat{\pi}_0 = \dfrac{\sum_{p=1}^{P} I(p(w_p) \in A)}{P \cdot l(A)}$. We choose A=[0.5, 1] and thus *l*(*A*)=0.5.

4. Estimate q-value of pathway *p* as $q(w_p) = \hat{\pi}_0 \cdot \sum_{b=1}^{B} \sum_{p'=1}^{P} I(w_{p'}^{(b)} \le w_p) \Big/ B \cdot \sum_{p'=1}^{P} I(w_{p'} \le w_p)$.

$\mathrm{P}_{MAPE\_P} = \{p : q(w_p) \le 0.05\}$ is the enriched pathways obtained by MAPE_P.

2.1.5.2 Algorithms for MAPE_G

Suppose there are K studies and G genes in each study.

I. For a given study *k*, compute the p-value of differential expression of each gene:

1. Compute the t-statistic, t_{gk}, of gene *g* in study *k*, where $1 \le g \le G$, $1 \le k \le K$.
2. Permute group labels in each study *B* times, and calculate the permuted statistics, $t_{gk}^{(b)}$, where $1 \le b \le B$.
3. Estimate the p-value of t_{gk} as $p(t_{gk}) = \sum_{b=1}^{B} \sum_{g'=1}^{G} I(|t_{g'k}^{(b)}| \ge |t_{gk}|) \Big/ B \cdot G$ and p-value of $t_{gk}^{(b)}$ as $p(t_{gk}^{(b)}) = \sum_{b'=1}^{B} \sum_{g'=1}^{G} I(|t_{g'k}^{(b')}| \ge |t_{gk}^{(b)}|) \Big/ B \cdot G$.

II. Meta-analysis:

1. The maximum p-value statistic (maxP) , $u_g = \max_{1 \le k \le K} p(t_{gk})$, is applied for the meta analysis. Similarly, $u_g^{(b)} = \max_{1 \le k \le K} p(t_{gk}^{(b)})$.
2. Estimate the p-value of maxP statistics as $p(u_g) = \sum_{b=1}^{B} \sum_{g'=1}^{G} I(u_{g'}^{(b)} \le u_g) \Big/ B \cdot G$.

III. Enrichment analysis:

1. Given a pathway p, compute v_p, the KS statistic for gene set enrichment based on $p(u_g)$.

2. Permute gene labels B times, and calculate the permuted statistics, $v_p^{(b)}$, 1 ≤b≤B.

3. Estimate the p-value of pathway p as $p(v_p) = {\sum_{b=1}^{B}\sum_{p'=1}^{P} I(v_{p'}^{(b)} \ge v_p)} \Big/ {B \cdot P}$ and similarly calculate $p(v_p^{(b)}) = {\sum_{b'=1}^{B}\sum_{p'=1}^{P} I(v_{p'}^{(b')} \ge v_p^{(b)})} \Big/ {B \cdot P}$.

4. Estimate π_0 , the proportion of non-enriched pathways in the meta-analysis, as $\hat{\pi}_0 = \frac{\sum_{p=1}^{P} I(p(v_p) \in A)}{P \cdot l(A)}$. We choose A=[0.5, 1] and thus *l*(*A*)=0.5.

5. Estimate q-value of pathway p as $q(v_p) = {\hat{\pi}_0 \cdot \sum_{b=1}^{B}\sum_{p'=1}^{P} I(v_{p'}^{(b)} \le v_p)} \Big/ {B \cdot \sum_{p'=1}^{P} I(v_{p'} \le v_p)}$.

   $\mathrm{P}_{MAPE\_G} = \{p : q(v_p) \le 0.05\}$ is the enriched pathways obtained by MAPE_G.

2.1.5.3 Algorithms for MAPE_I

1. Let $s_p = \min(p(v_p), p(w_p))$ and $s_p^{(b)} = \min(p(v_p^{(b)}), p(w_p^{(b)}))$.

2. Estimate the p-value of s_p as $p(s_p) = {\sum_{b=1}^{B}\sum_{p'=1}^{P} I(s_{p'}^{(b)} \le s_p)} \Big/ {B \cdot P}$.

3. Estimate π_0 , the proportion of non-enriched pathways in the meta-analysis, as $\hat{\pi}_0 = \frac{\sum_{p=1}^{P} I(p(s_p) \in A)}{P \cdot l(A)}$. We choose A=[0.5, 1] and thus *l*(*A*)=0.5.

4. Estimate q-value of s_p as $q(s_p) = {\hat{\pi}_0 \cdot \sum_{b=1}^{B}\sum_{p'=1}^{P} I(s_{p'}^{(b)} \le s_p)} \Big/ {B \cdot \sum_{p'=1}^{P} I(s_{p'} \le s_p)}$.

   $\mathrm{P}_{MAPE\_I} = \{p : q(s_p) \le 0.05\}$ is the enriched pathway identified by the method MAPE_I.

2.2 SIMULATION COMPARISON OF MAPE METHODS

We applied a one-pathway simple simulation model to compare the power of MAPE_G and MAPE_P to identify conditions (parameter subspace) in which one method outperforms the other. The result gives us insight into the unique advantages of MAPE_G and MAPE_P. It also argues the necessity of MAPE_I when a mixture of the two types of pathways exists in the data and we are interested in detecting both types of pathways.

Suppose G=500 genes are contained in the genome. The first 100 genes belong to a pathway. Our pathway database has only one pathway (p=1): $\{z_{gp}\}$, z_{gp}=1 when $1 \leq g \leq 100$ and z_{gp}=0 when $101 \leq g \leq 500$. We generate a random binary vector $D=\{d_1,\ldots,d_G\}$ to indicate whether gene g is a DE gene or not. The probability of being a DE gene in the first 100 genes is α and the probability of being a DE gene in all 500 genes is α_0. (i.e. Pr(d_{g}=1)= α if $1 \leq g \leq 100$ and Pr(d_{g}=1)= α_0 if $1 \leq g \leq 500$). We fix α_0=0.1 in our simulation. Intuitively, there is no pathway enrichment if α=0.1 and pathway enrichment exists if α>0.1.

Given the DE gene indicators, two independent array studies are subsequently simulated for meta-analysis. We assume each study contains S=40 samples. The first 20 samples are controls and the next 20 samples are cases (i.e. y_s=0 if $1 \leq s \leq 20$ and y_s=1 if $21 \leq s \leq 40$). When gene g is a DE gene (d_g=1) and for all k, the expression intensities are simulated from $x_{kgs} \sim N(\theta,1)$ if $1 \leq s \leq 20$ and $x_{kgs} \sim N(0,1)$ if $21 \leq s \leq 40$. For a non-DE gene g (d_g=0), the expression intensities are simulated from $x_{kgs} \sim N(0,1)$ $\forall s$ and k. We further assume that the two array studies adopt different array platforms and each of them only covers a portion of genes in the genome. We assume the chance of each gene to be covered by study k is randomly generated with a sampling rate λ_k. The sampled indicator vectors for gene g in study k is denoted by h_{gk}, where h_{gk} =1 if gene g appears in study k and h_{gk} =0 otherwise. In the following, we set λ=Pr(h_{gk}=1)=λ_k (1≤g≤G=1000 and 1≤k≤K=2). As a result, study k contains $G_k = \sum_{g=1}^{G} h_{gk}$ genes in the data matrix, which is a random variable and may be different in each simulation. The overlapped gene set of the two studies contains $G' = \sum_{g=1}^{G} h_{g1} \cdot h_{g2}$. In the implementation of MAPE_P, the original data in both studies with G_1 and G_2 genes can be used. For MAPE_G, the method requires only matched genes and the subset of G′ overlapped gene set in each study will be applied.

The powers of MAPE_P, MAPE_G, and MAPE_I are calculated as follows:

1. Simulate study one and study two with a given parameter vector {θ, α, λ}. Compute the p-value of the gene set enrichment by MAPE_G and MAPE_P methods. We will declare that the gene set is found enriched if the p-value is less than 5%.

2. Repeat step 1 and 2 for B=200 times.

3. Suppose the p-values for MAPE_G and MAPE_P are $p_G^{(b)}$ and $p_P^{(b)}$ respectively, the powers are calculated as $Power_G(\theta, \alpha, \lambda) = \sum_{b=1}^{B} I\left(p_G^{(b)} < 0.05\right)/B$ and $Power_P(\theta, \alpha, \lambda) = \sum_{b=1}^{B} I\left(p_P^{(b)} < 0.05\right)/B$ for each method.

We perform α={0.15, 0.2, 0.25, 0.3}, λ={0.4, 0.6, 0.8, 1} and assign the values to θ_k based on the following 5 scenarios:

1) θ_1 and θ_2 are fixed values and $\theta_1 = \theta_2$; K=2.

We first investigated this simple scenario and θ varies from 0.5 to 4. Specifically, $\theta_1 = \theta_2$ ={0.5, 0.75, 1, 1.5, 2, 4}.

2) θ_1 and θ_2 are fixed values but $\theta_1 \neq \theta_2,$ K=2.

Let θ=[$\theta_1,$ θ_2]. Then the power of MAPE_P and MAPE_G were calculated when θ was assigned to [2,3] and [2,4] respectively.

3) θ_k is fixed and $\theta_1 = \theta_2 = \ldots = \theta_K$, K=4 and 10.

In this scenario, the number of studies was increased to 4 and 10 and θ_k=0.5, 1, 2, 4, k=1,2,…,K.

4) θ_k is a random variable and normally distributed, K=2.

In scenario 1-3, θ_k is a fixed value. In this scenario, θ_k was assigned to random number generated by normal distribution with mean equal to m and standard deviation equal to s, where m={1.5, 2, 4} and s=0.5.

5) K=4 and one of 4 studies is considered as an outlier.

In scenario 1-4, all studies are consistent with each other. In this scenario, 4 studies were generated and one of them was considered as an outliers. We simulated this scenario by two ways:

5.1) θ_k=2, k=1,2,3. θ_4 =.1.

In this case, in the first three studies, θ was set to 2. In the fourth study, θ was set to a smaller value, 0.1, instead.

5.2) $\theta_k \sim N(2,0.05)$, k=1,2,3. $\theta_4 \sim N(2,0.2)$.

In this case, 4 studies were simulated and θ_k was set to 2, k=1, 2, …, 4. Then noise was added to the expression values. In the first three studies, the noise was distributed as $N(0, 0.05)$ and in the fourth study, the noise was stronger and distributed as $N(0, 0.2)$.

A total of B=200 independent simulations are performed for each parameter setting. Intuitively, θ represents the effect size of the DE genes in the data, α represents the strength of pathway enrichment, and λ represents the coverage of an array platform on the genome. The power calculation of MAPE procedures is calculated as the proportion of times the pathway is claimed as an enriched pathway.

The power of MAPE_P, MAPE_G and the power difference of MAPE_P and MAPE_G (i.e. $Power_{MAPE\_P}(\theta,\alpha,\lambda) - Power_{MAPE\_G}(\theta,\alpha,\lambda)$) for scenario 1-5 were shown in Figure 2.3 to 2.8 respectively by gradient colors under different θ, α and λ conditions. The smooth contour plots are performed with a surface smoothing technique using the R package field (Fields Development Team, 2006).

For the results for scenario 1 shown in Figure 2.3, we can clearly see that, when θ is low ($0.5 \le \theta_1 = \theta_2 \le 1$), MAPE_G is more powerful than MAPE_P when the pathway enrichment strength α is low. Specifically, MAPE_G is more powerful than MAPE_P when 1) $\theta_1 = \theta_2 = 0.5$ and α is lower than around 0.25; 2) $\theta_1 = \theta_2 = 0.75$ and α is lower than around 0.19; 3) $\theta_1 = \theta_2 = 1$, α is lower than around 0.18 and $0.4 \le \lambda \le 0.6$. The cutoff of α for MAPE_G dominating MAPE_P is roughly decreasing when θ increases.

When θ is large ($1.5 \le \theta_1 = \theta_2 \le 4$), MAPE_G is more powerful than MAPE_P when the array coverage rate λ ($0.7 \le \lambda \le 1$) is high and the pathway enrichment strength α is low ($0.15 \le \lambda \le 0.2$).

The above observations are consistent with the complementary advantages of MAPE_G vs. MAPE_P discussed in section 2.1.2. When both of the effect size θ and the pathway enrichment strength α are low, the MAPE_P procedure has low power to detect enriched pathway in each individual study thus also has lower power to detect enriched pathway after meta-analysis step. However, MAPE_G procedure combines p-values of genes directly and is able to detect more DE genes than MAPE_P procedure, which makes MAPE_G more powerful. On the other hand, when the effect size θ is large, for a low array coverage rate λ ($0.4 \le \lambda \le 0.7$), the advantage of MAPE_P of not requiring gene matching across studies becomes evident and MAPE_P is more powerful than MAPE_G.

For scenario 2 (Figure 2.4), when $\theta_1 \neq \theta_{2,}$ we got similar results as that for scenario 1.

For scenario 3, when the number of study *K* = 4 (Figure 2.5), MAPE_G is more powerful than MAPE_P when array coverage rate λ is large. When *K*= 10 (Figure 2.6), MAPE_G is more powerful than MAPE_P almost everywhere in the parameter space. In our simulation model, the number of common genes among all studies exponentially decreases with respect to *K*, while the low number of common genes leads to low power of MAPE_G procedure.

For scenario 4 (Figure 2.7) and scenario 5 (Figure 2.8), similar results were found as that for scenario 1.

Our simulation examines the power of a single pathway. In a real application, hundreds to thousands of pathways are analyzed in the pathway database. Both types of pathways for which MAPE_G or MAPE_P have better power will co-exist in an analysis. This motivates our development of an integrated method MAPE_I to incorporate the advantages of the two methods. In the next step, the power of MAPE_I was compared to the power of MAPE_P and MAPE_G for scenario 1-5 (Figure 2.9 to 2.14). The simulation results show that MAPE_I clearly has more robust performance than MAPE_G or MAPE_P.

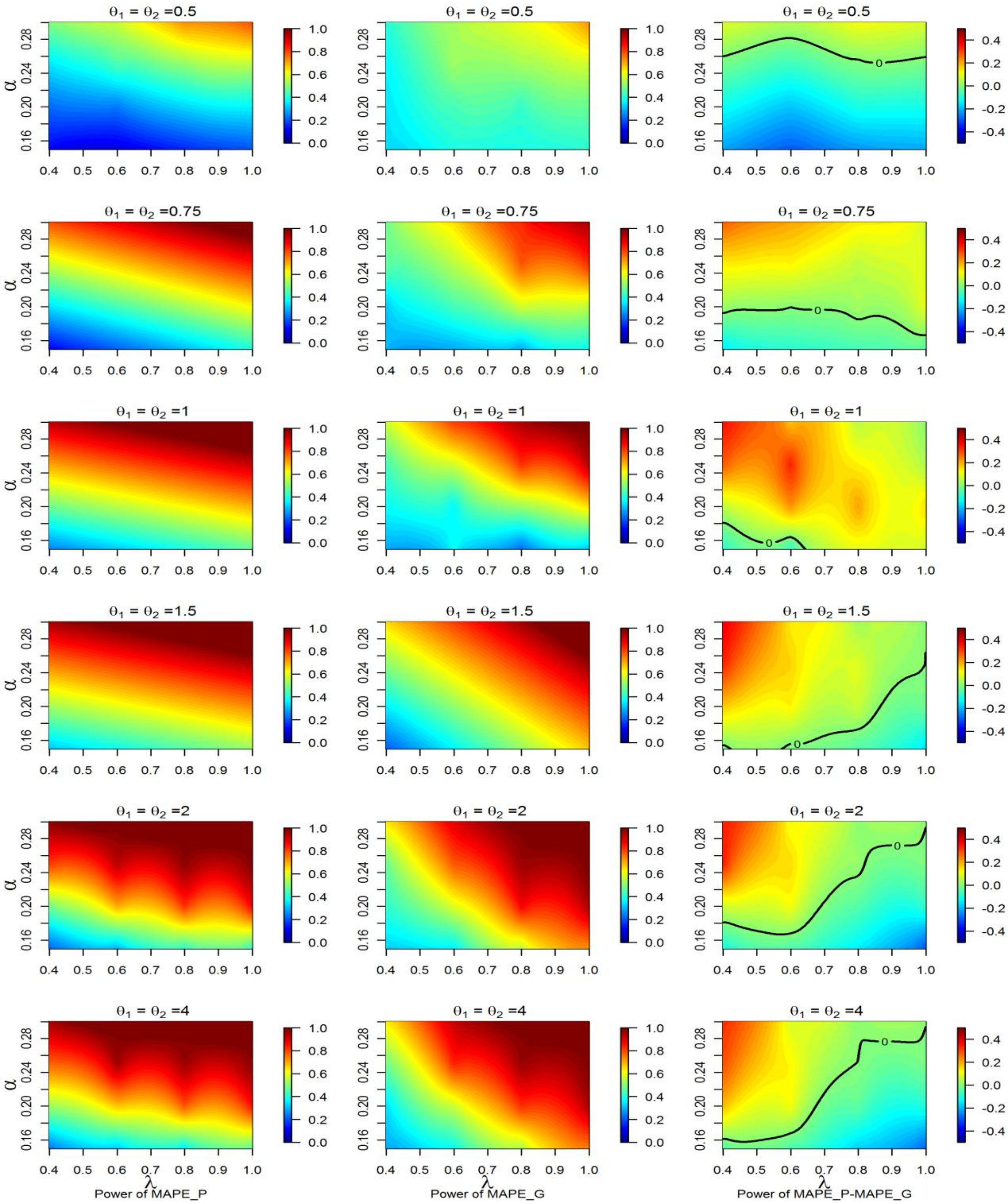


Figure 2.3. Power comparison between MAPE_P and MAPE_G for scenario 1.

The first two columns represent the power of MAPE_P and MAPE_G respectively. The third column represents the difference between the power of MAPE_P and the power of MAPE_G. θ_1 and θ_2 are fixed values and $\theta_1 = \theta_2$.

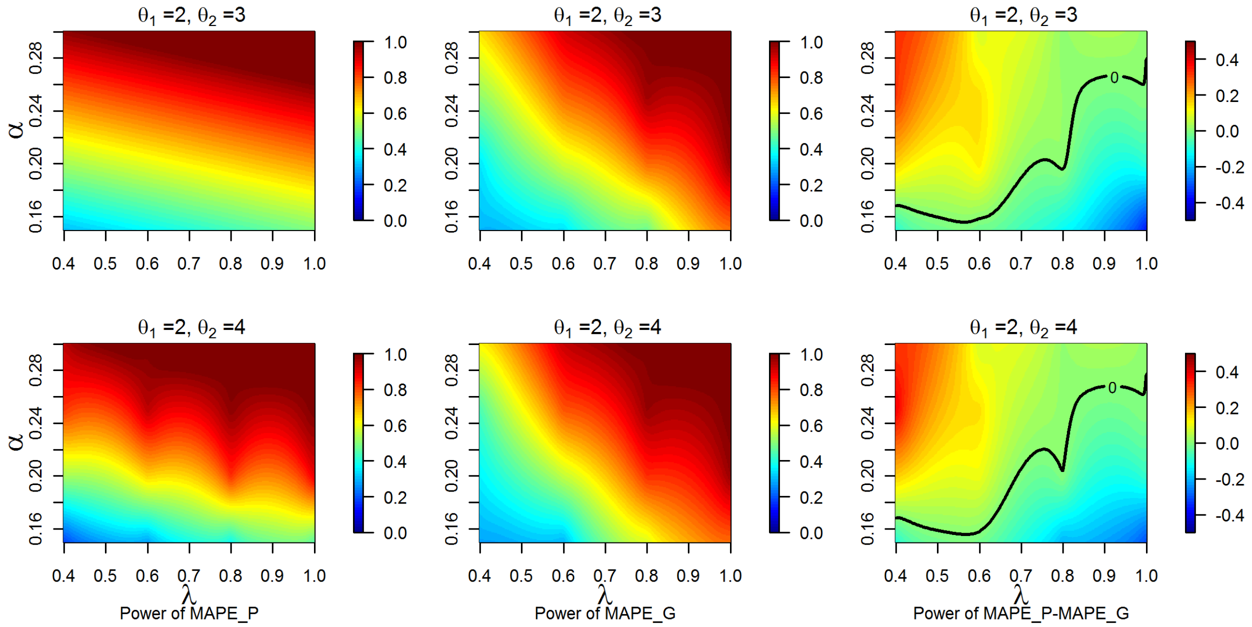


Figure 2.4 Power comparison between MAPE_P and MAPE_G for scenario 2.

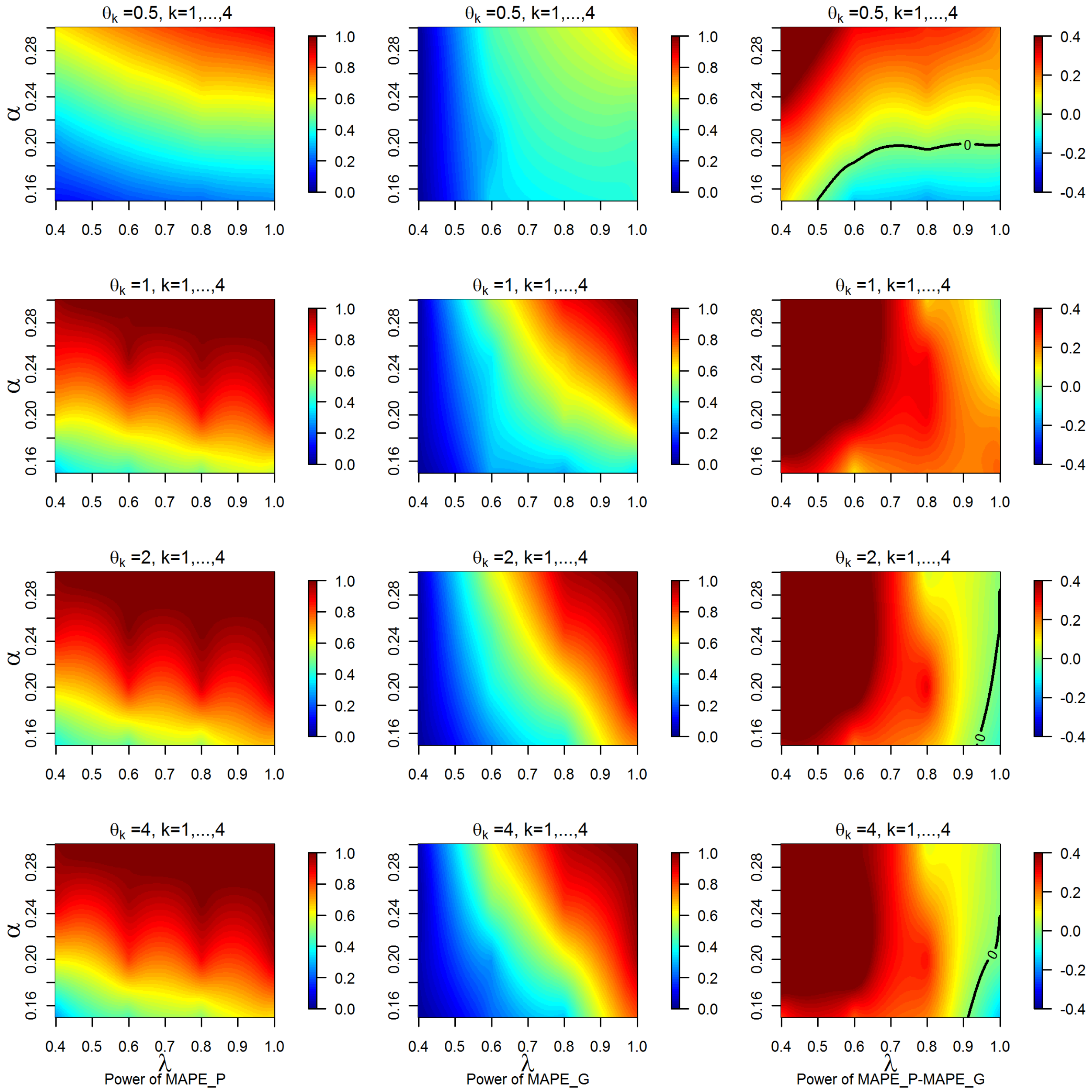


Figure 2.5 Power comparison between MAPE_P and MAPE_G for scenario 3 when *K*=4.

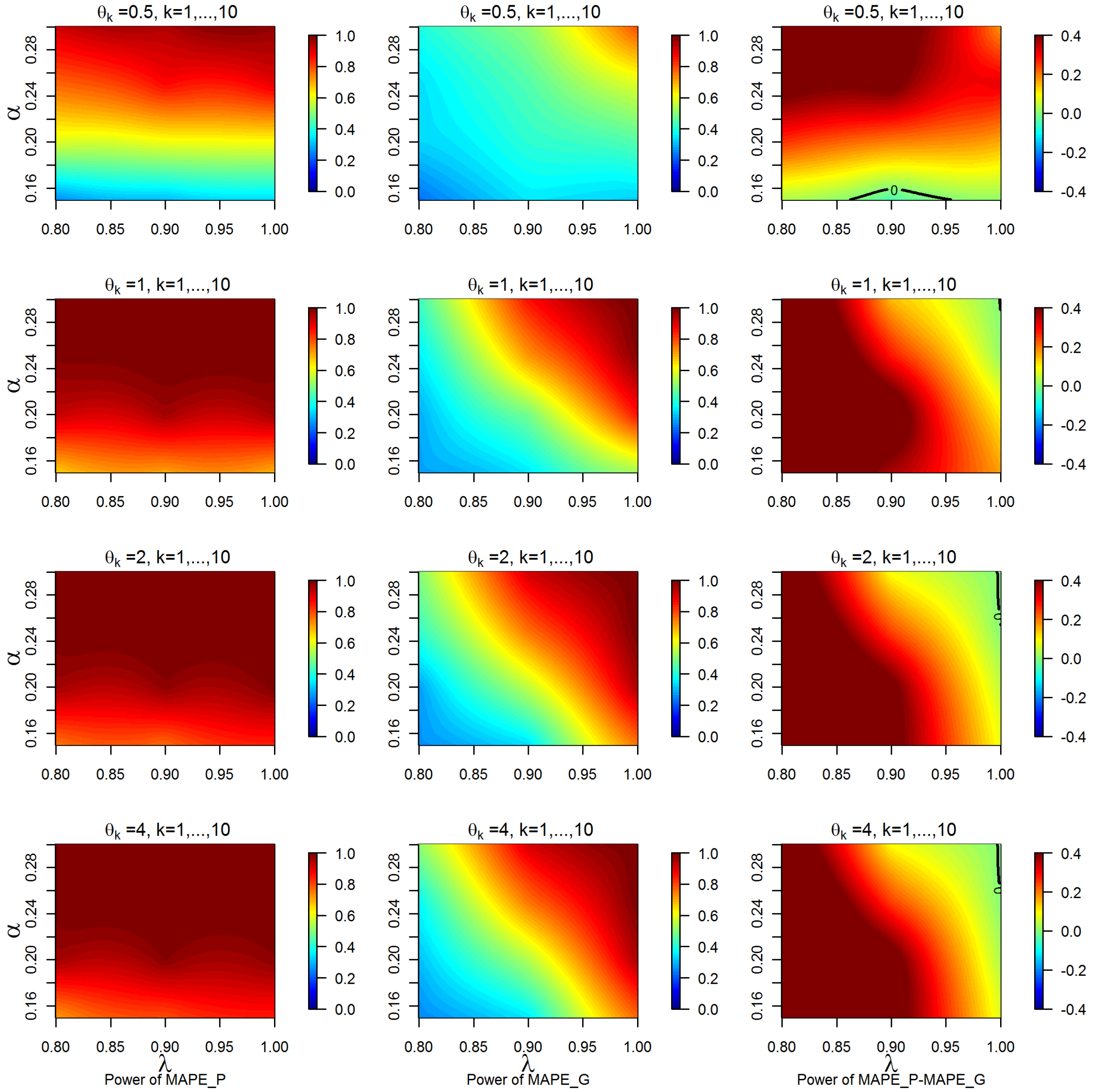


Figure 2.6 Power comparison between MAPE_P and MAPE_G for scenario 3 when *K*=10.

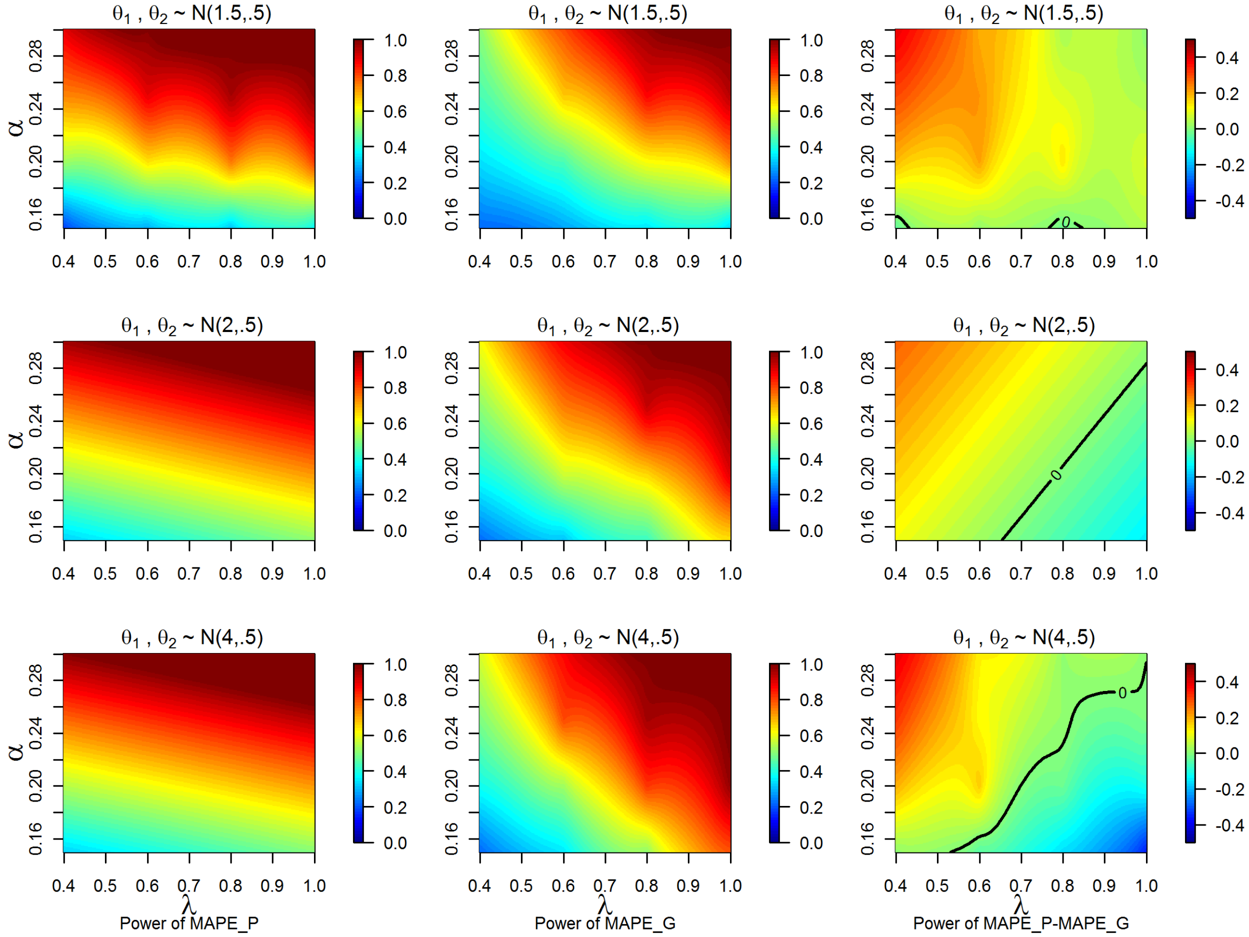


Figure 2.7 Power comparison between MAPE_P and MAPE_G for scenario 4.

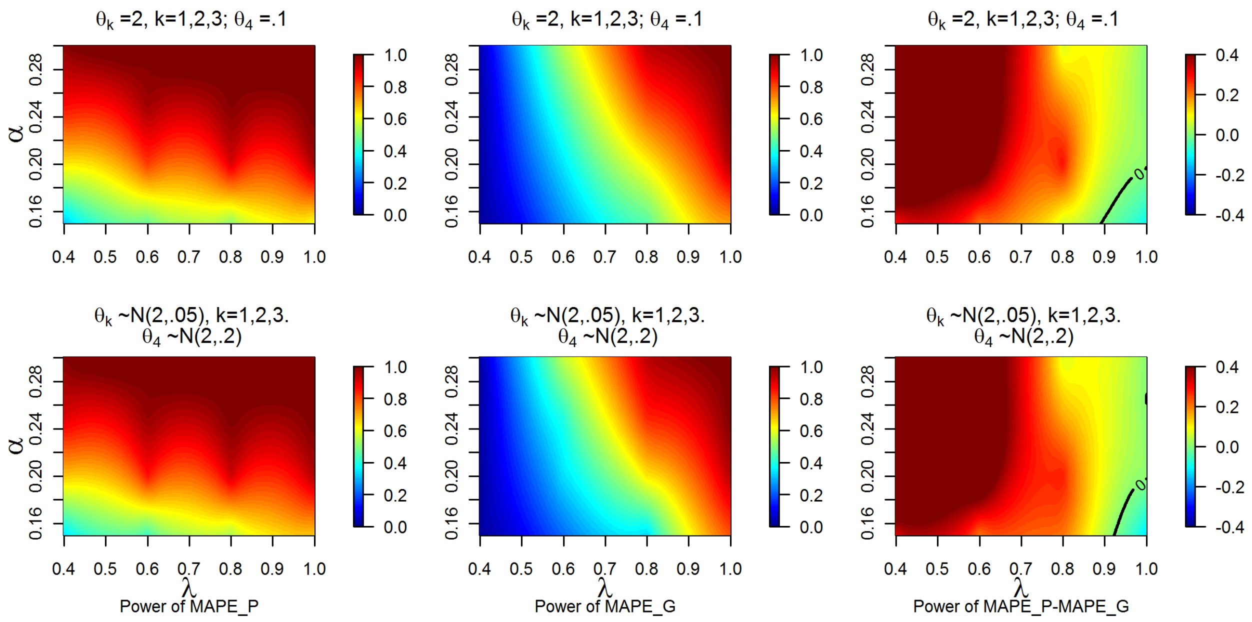


Figure 2.8 Power comparison between MAPE_P and MAPE_G for scenario 5.

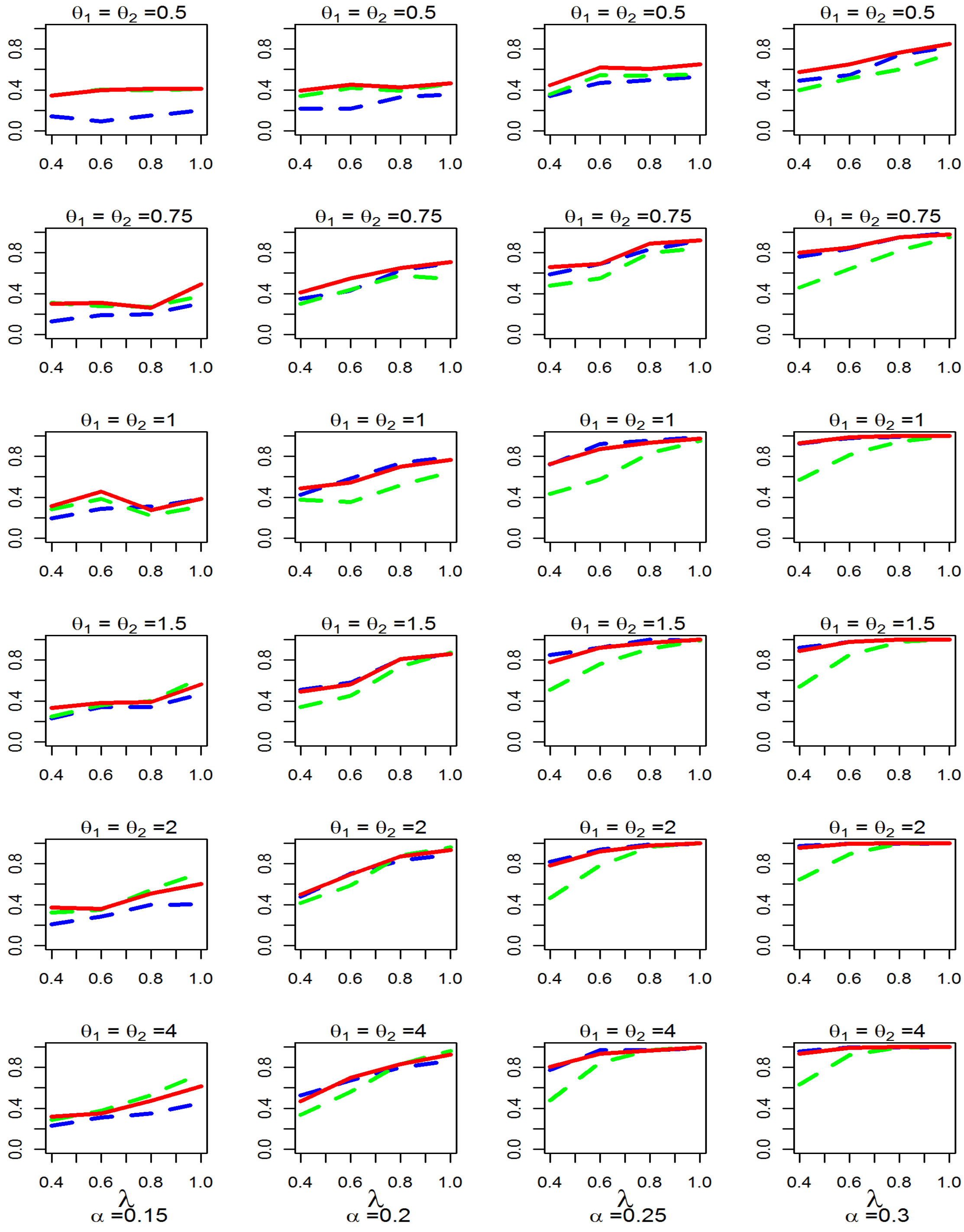


Figure 2.9 Power comparison among MAPE_I, MAPE_P and MAPE_G for scenario 1.

The statistical power of MAPE_P (blue dashed lines), MAPE_G (green dashed lines) and MAPE_I (red solid lines) are displayed (on y-axis) for different λ (on x-axis) and different α (four columns). The result shows that MAPE_I always have the best or near the best statistical power among the three.

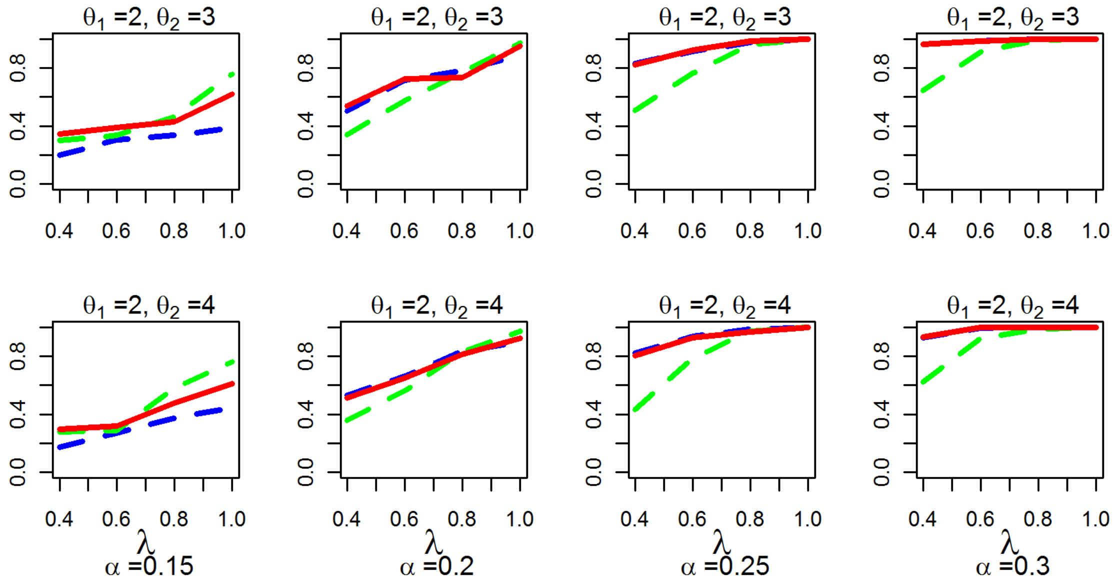


Figure 2.10 Power comparison among MAPE_I, MAPE_P and MAPE_G for scenario 2.

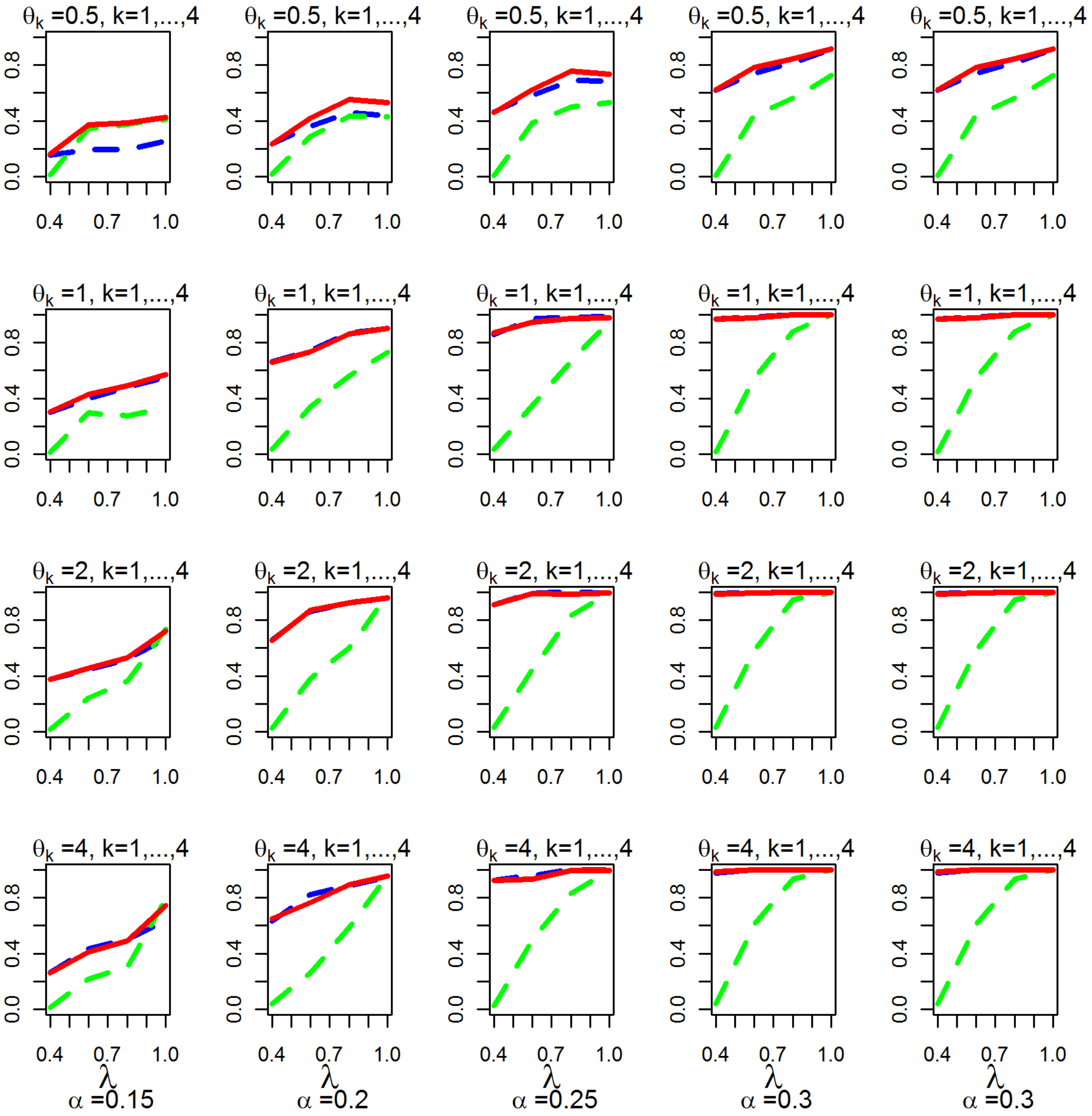


Figure 2.11 Power comparison among MAPE_I, MAPE_P and MAPE_G for scenario 3 when *K*=4.

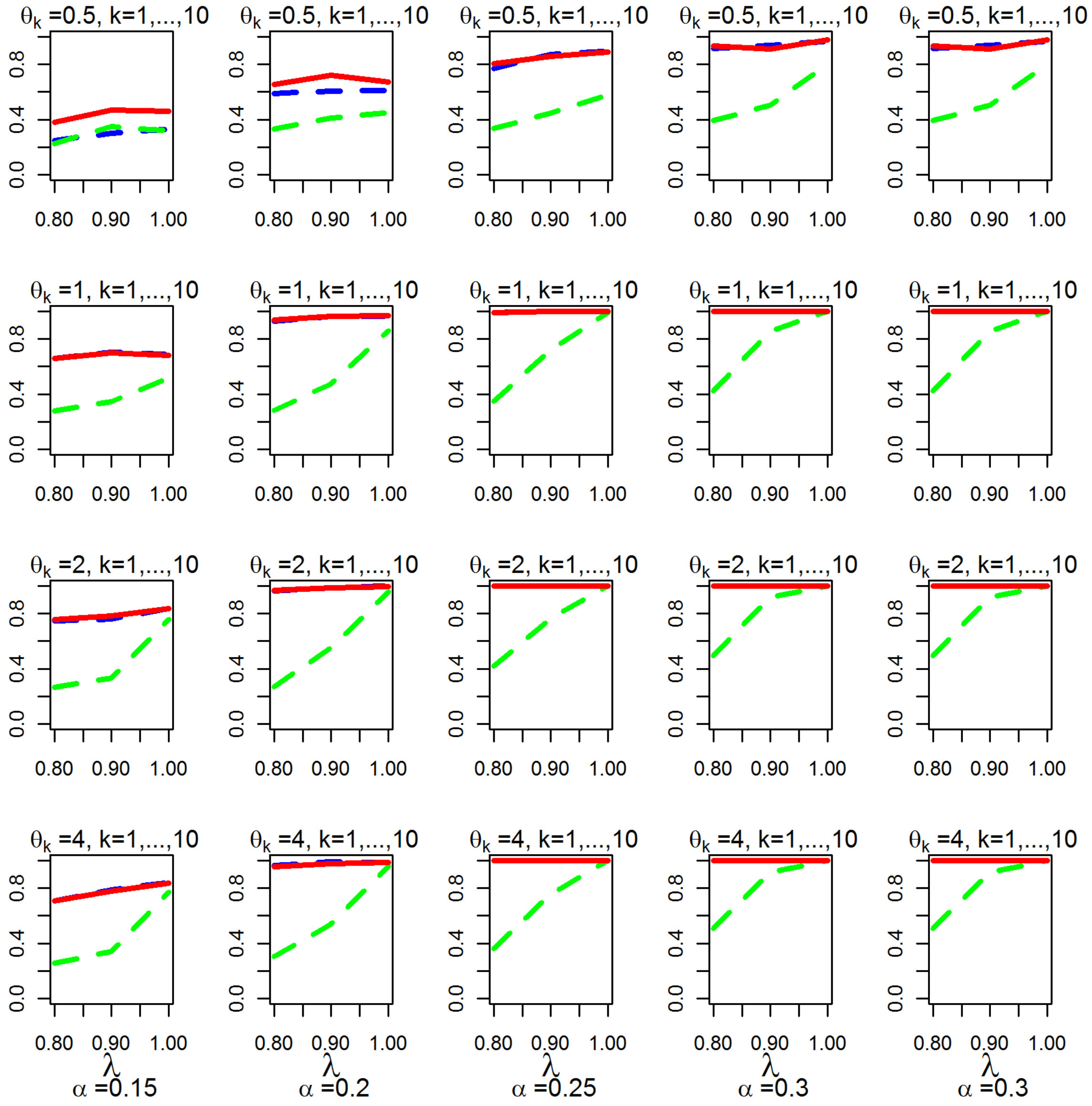


Figure 2.12 Power comparison among MAPE_I, MAPE_P and MAPE_G for scenario 3 when *K*=10.

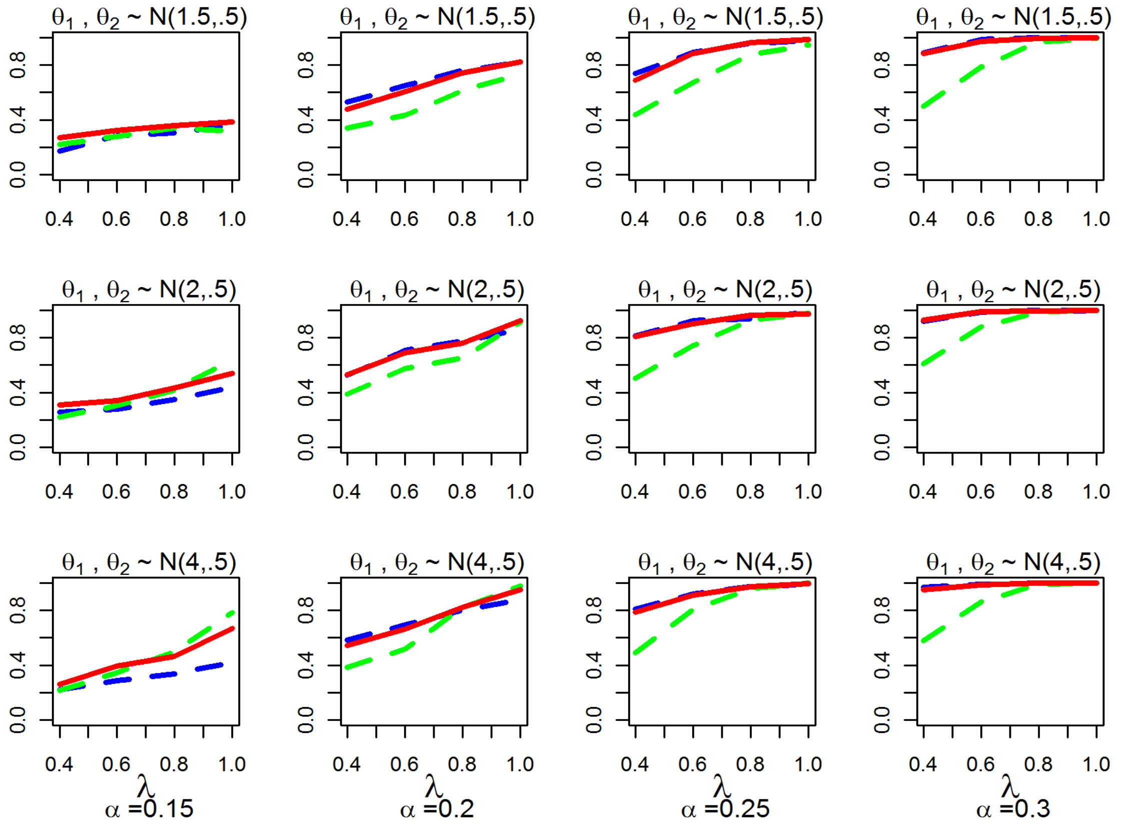


Figure 2.13 Power comparison among MAPE_I, MAPE_P and MAPE_G for scenario 4.

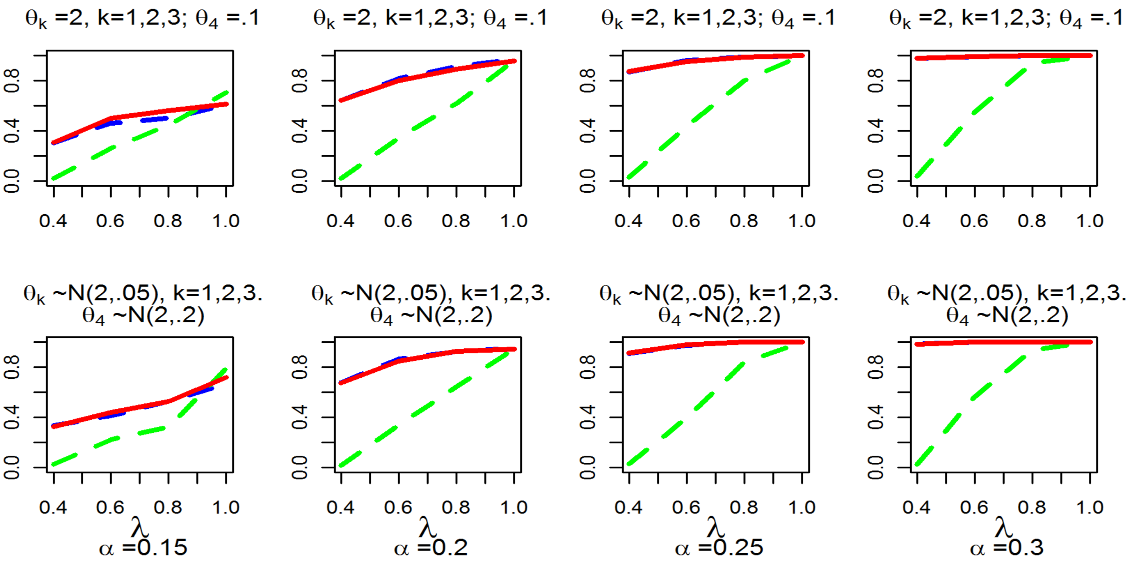


Figure 2.14 Power comparison among MAPE_I, MAPE_P and MAPE_G for scenario 5.

2.3 APPLICATIONS ON REAL MICROARRAY DATA SETS

2.3.1 Application to the drug response studies

In section 1.2.6, gene level meta-analysis has applied on two chemosensitivity studies. In this section, we applied MAPE approaches to the same data sets to identify enriched pathways that are related to drug response to paclitaxel in breast cancer cells lines. In our analysis, when the q-value cutoff was set to 0.15, 60 pathways were identified by MAPE_P, 36 by MAPE_G, and 54 by MAPE_I. If we relax the q-value cutoff of MAPE_I to 0.2, then all the 71 pathways identified by MAPE_P or MAPE_G at cutoff 0.15 were also identified by MAPE_I, showing that MAPE_I is a good way to incorporate and summarize results from MAPE_P and MAPE_G. To demonstrate the advantage of meta-analysis, the result from MAPE_I was compared to individual study pathway analysis (lower plots of Figure 2.15). The Liedtke study identified 28 pathways and the Neve study identified 21 pathways, while MAPE_I detected a total of 54 pathways. Among the 27 pathways detected by MAPE_I but not by either individual study analysis (group IV in Figure 2.15 lower-right Venn diagram), many are known drug-response related pathways, including LEE_MYC_TGFA_UP, EGF_HDMEC_UP. Details of all enriched pathway results are listed in supplemental Table 2. These pathways are predominantly related to cell proliferation, oncogenic pathways, and estrogen receptor-associated gene sets. Noticeably, our results indicate that some important oncogenic pathways related to EGF, MYC and TGFBETA may be highly correlated to chemotherapy response.

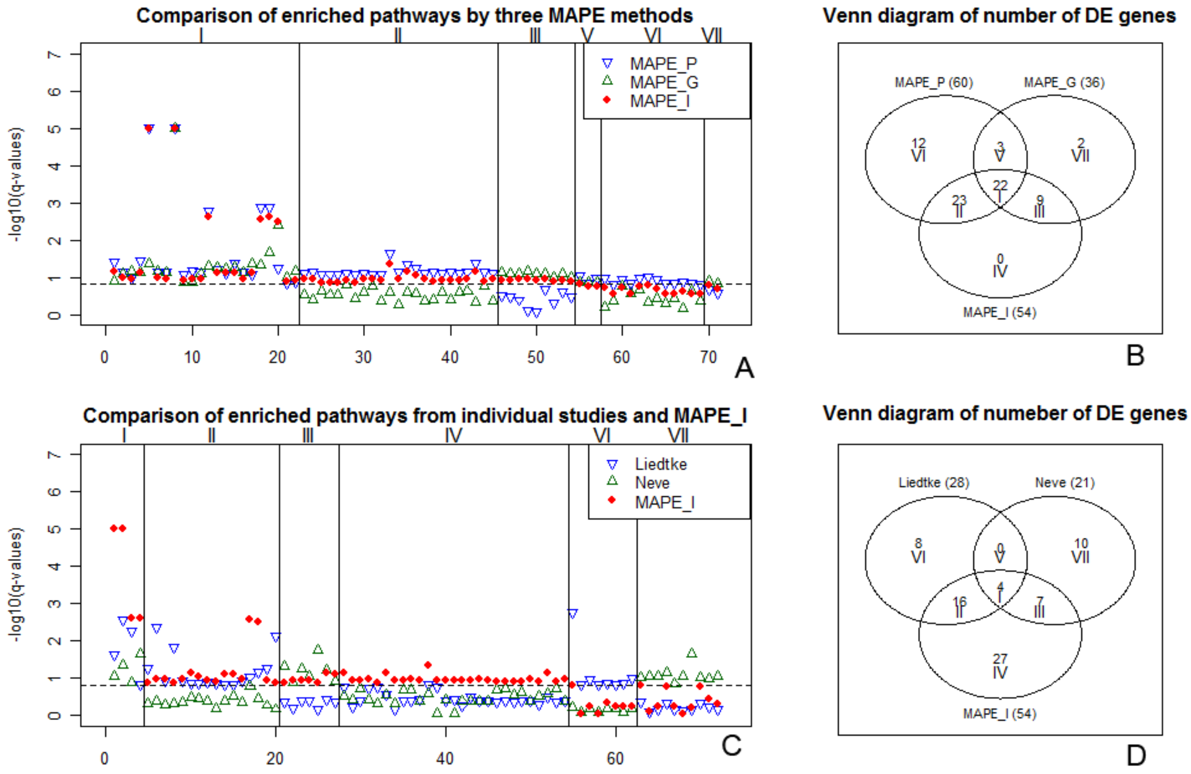


Figure 2.15. MEAP results for drug response studies.

log-transformed (base 10) q-values of pathways detected by MAPE_P (blue), MAPE_G (green) and MAPE_I (red). The Figure has been divided into 7 regions. Region I contains the pathways enriched by all three MAPE methods. Region II contains pathways enriched by MAPE_P and MAPE_I but not MAPE_G. Region III contains pathways enriched by MAPE_G and MAPE_I but not MAPE_P. Region IV contains pathways enriched by MAPE_I but not MAPE_P and MAP_G. Region V contains pathways enriched by MAPE_P and MAPE_G but not MAPE_I. Region VI contains pathways enriched by MAPE_P but not MAPE_I and MAPE_G. Region VII contains pathways enriched by MAPE_G but not MAPE_I and MAPE_P. Upper right: Venn diagram of the pathways detected by MAPE_P, MAPE_G and MAPE_I. Lower left: log-transformed (base 10) q-values of pathways detected by individual study Liedtke (blue), Neve (green) and meta-analysis MAPE_I (red). Lower right: Venn diagram of the pathways detected by Liedtke alone, Neve alone and MAPE_I.

2.3.2 Application to the lung cancer studies

In this section, we applied MAPE methods to two lung cancer studies, details shown in Table 2.1. The raw microarray data sets were processed by procedures similar to those described in section 1.2.6.

Table 2.1. Summary of lung cancer data sets

Study	Platform	Normal samples	Tumor samples	Probe IDs
Bhat (Bhattacharjee, et al., 2001)	HGU95A	16	139	12625
Beer (Beer, et al., 2002)	HG6800	10	86	7129

When the q-value cutoff was set to 0.05, MAPE_P identified 137 enriched pathways and MAPE_G identified 81 (Figure 2.16). There were 63 common enriched pathways detected by both methods. MAPE_I integrates information from both MAPE_P and MAPE_G and identified 114 enriched pathways. The enriched pathways identified by MAPE_I are important. These pathways play important roles in cell migration, cell communication, adhesion, and amino acid metabolism, pathways known to be closely related to tumor progress. The details of the enriched pathways are listed in the Appendix B. Seven pathways detected by MAPE_G and 31 by MAPE_P were not included in the enriched pathway list by MAPE_I. However, this does not indicate that these pathways are not important. If we relax the q-value cutoff of MAPE_I from 0.05 to 0.10, all enriched pathways identified by MAPE_P and MAPE_G were included by MAPE_I. This indicates that MAPE_I, a combination of MAPE_P and MAPE_G, is a good indicator for ranking the pathways.

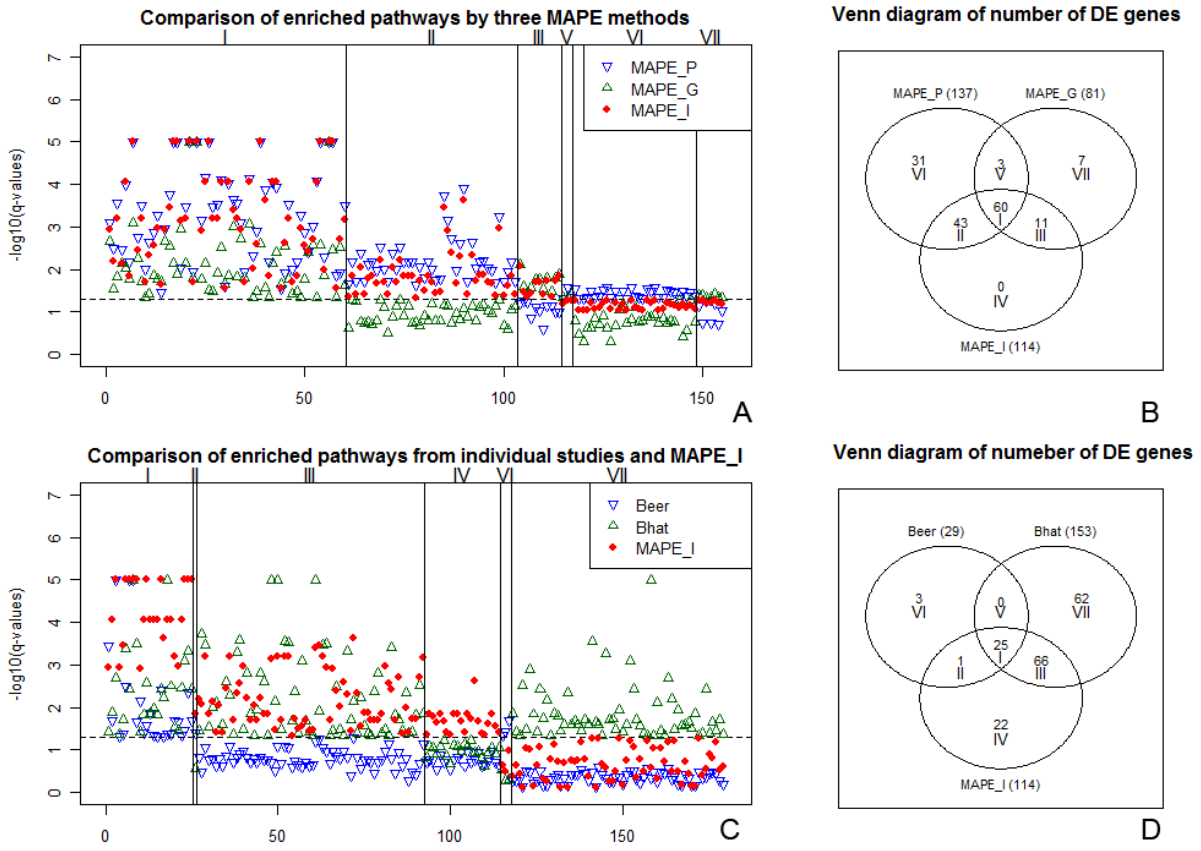


Figure 2.16. MEAP results for lung cancer studies.

Upper left: log-transformed (base 10) q-values of pathways detected by MAPE_P (blue), MAPE_G (green) and MAPE_I (red). The Figure has been divided into 7 regions. Region I contains the pathways enriched by all three MAPE methods. Region II contains pathways enriched by MAPE_P and MAPE_I but not MAPE_G. Region III contains pathways enriched by MAPE_G and MAPE_I but not MAPE_P. Region IV contains pathways enriched by MAPE_I but not MAPE_P and MAP_G. Region V contains pathways enriched by MAPE_P and MAPE_G but not MAPE_I. Region VI contains pathways enriched by MAPE_P but not MAPE_I and MAPE_G. Region VII contains pathways enriched by MAPE_G but not MAPE_I and MAPE_P. Upper right: Venn diagram of the pathways detected by MAPE_P, MAPE_G and MAPE_I. Lower left: log-transformed (base 10) q-values of pathways detected by individual study Beer (blue), Bhat (green) and meta-analysis MAPE_I (red). Lower right: Venn diagram of the pathways detected by Beer alone, Bhat alone and MAPE_I.

2.3.3 Application to the prostate cancer studies

In this section, we applied MAPE methods to two prostate cancer studies, details shown in Table 2.2. The raw microarray data sets were processed by procedures similar to those described in section 1.2.6.

Table 2.2. Summary of prostate cancer data sets

Study	Platform	Normal samples	Tumor samples	Probe IDs
Welsh (Welsh, et al., 2001)	HGU95A	9	25	12625
Singh (Singh, et al., 2002)	HGU95Av2	50	52	12625

When the q-value cutoff was set to 0.05, 57 pathways were identified by MAPE_P, 11 by MAPE_G, and 47 by MAPE_I. If we relax the q-value cutoff of MAPE_I to 0.2, then all the 55 pathways identified by MAPE_P or MAPE_G at cutoff 0.05 were also identified by MAPE_I. The Welsh study identified 28 pathways and the Singh study identified 53 pathways, while MAPE_I detected a total of 47 pathways.

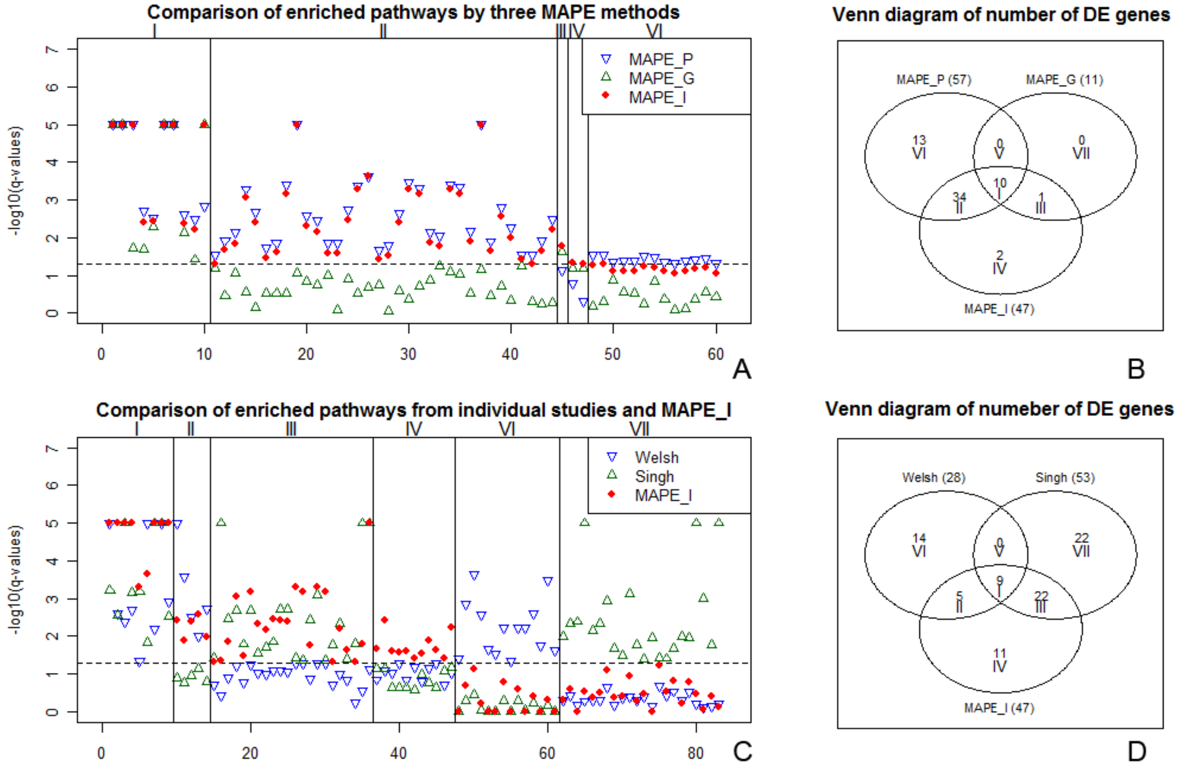


Figure 2.17. MEAP results for prostate cancer studies.

log-transformed (base 10) q-values of pathways detected by MAPE_P (blue), MAPE_G (green) and MAPE_I (red). The Figure has been divided into 7 regions. Region I contains the pathways enriched by all three MAPE methods.

Region II contains pathways enriched by MAPE_P and MAPE_I but not MAPE_G. Region III contains pathways enriched by MAPE_G and MAPE_I but not MAPE_P. Region IV contains pathways enriched by MAPE_I but not MAPE_P and MAP_G. Region V contains pathways enriched by MAPE_P and MAPE_G but not MAPE_I. Region VI contains pathways enriched by MAPE_P but not MAPE_I and MAPE_G. Region VII contains pathways enriched by MAPE_G but not MAPE_I and MAPE_P. Upper right: Venn diagram of the pathways detected by MAPE_P, MAPE_G and MAPE_I. Lower left: log-transformed (base 10) q-values of pathways detected by individual study Welsh (blue), Singh (green) and meta-analysis MAPE_I (red). Lower right: Venn diagram of the pathways detected by Welsh alone, Singh alone and MAPE_I.

3.0 SOFTWARE PACKAGE AND IMPLEMENTATION ISSUES FOR MAPE

In Chapter 2, the statistical framework and algorithms of MAPE analysis have been presented. In this chapter, we discuss the computational and practical issues of MAPE implementation. We first introduce the MetaPath software package for performing the MAPE analysis. Then an example is given to demonstrate how to apply MAPE analysis when the number of studies is large. We also collected a chemotherapy microarray database which is discussed in the end of this Chapter.

3.1 IMPLEMENTATION OF THE METAPATH PACKAGE

3.1.1 Functions of MetaPath package

We developed a software package named MetaPath using the R language (R Development Core Team, 2005) to perform the MAPE_G, MAPE_P and MAPE_I analyses. In addition, the MetaPath package also provided the following useful functions:

A) Data package

We provided a function to package the microarray data according to the Biobase/Bioconductor’s (Gentleman, et al., 2004) standardized data structures to represent microarray data sets.

B) Probe ID mapping.

A function for mapping among probe ID, Gene symbol, Entrez ID (Bruford, et al., 2008) is also available in MetaPath package.

C) Pathway database importing

The pathway database which has the same data structure as molecular signatures database (Subramanian, et al., 2005) can be imported to R environment for further analysis.

D) Pathway enrichment analysis.

Pathway enrichment analysis based on Fisher exact test, t-test, linear regression, KS-test and Wilcoxon rank sum test can be performed.

E) Meta-analysis for genomic biomarkers

MetaPath package also provides functions for meta-analysis to identify DE genes/biomarkers.

F) MAPE

The core function of MetaPath package is MAPE, which performs the MAPE_G, MAPE_P and MAPE_I analysis and generates the reports. Multiple different procedures are available for MAPE analysis. Here we use the MAPE_G analysis procedure to demonstrate the selection for multiple procedures. As shown in Figure 3.1, the first step of MAPE_G is to calculate the association score with each phenotype. Four methods are available to conduct this step for different experimental designs; these include Student's t-test or F-test for two or multiple group comparison experiments, and the correlation/Cox hazard model for time series/survival time studies (R package superpc is used for Cox hazard model estimation (Bair and Tibshirani, 2004)). The second step was for meta-analysis. The MetaPath package includes functions for performing Fisher's statistics, MinP, MaxP and the AW method (the function for the AW method was implemented by Li (Li 2008)). The third step comprised the enrichment analysis.

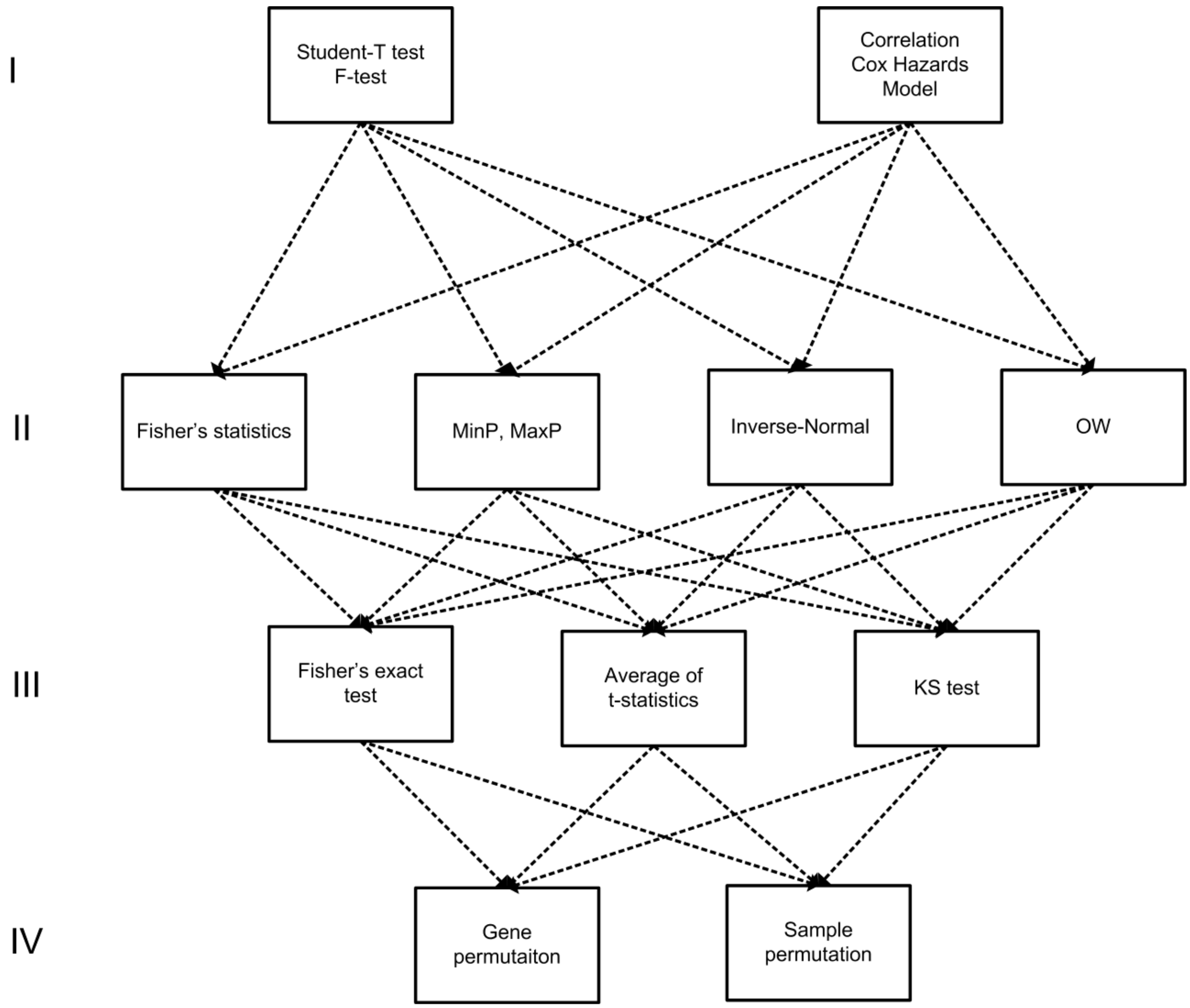


Figure 3.1. Statistical methods for the MAPE_G procedure.

The following methods are provided: 1) Fisher's exact test method; 2) Average of t-statistics method; 3) KS test method. In the fourth step, either a gene-wise permutation or a sample-wise permutation procedure can be used to control the FDR. The method combination of MAPE_P was similar to that of MAPE_G. The users of the MetaPath package can select the appropriate procedure for their own purposes.

In section 1.2 and 1.3, algorithms for performing meta-analysis and pathway enrichment analysis were given for binary phenotype. Here the algorithm for pathway enrichment analysis for continuous phenotype was given. The algorithm for meta-analysis for continuous phenotype at the gene level will be given in Chapter 4.

Details of the pathway enrichment algorithm for continuous phenotype are as follows:

Let x_{gs} denote the gene expression value for gene *g*, sample *s*, *s*, 1≤*g*≤G, 1≤*s*≤S. Let y_s denote the continuous values for phenotype for sample *s*. The regression coefficients β_{1g} for gene *g* was estimated using a standard linear regression model $y_s = \beta_{0g} + \beta_{1g} x_{gs} + \epsilon_{sg}$, where ε is the normal error. Let $t_g = \beta_{1g} / (s_g + s_0)$, where s_g is the standard deviation of β_{1g}. The β_{1g} was calculated by the following formulas:

$$\beta_{1g=} \frac{\sum_{s=1}^{S} y_s (x_{gs} - \bar{x}_g)}{\sum_{s=1}^{S} (y_s - \bar{y}\ )^2},$$

where $\bar{x}_g = \sum_{s=1}^{S} x_{gs} / s,\ \bar{y}\ = \sum_{s=1}^{S} y_s / s$.

$$s_g = \frac{\hat{\sigma}_g}{[\sum_{s=1}^{S} (y_s - \bar{y}\ )^2]^{1/2}} .$$

$$\hat{\sigma}_g = [\frac{\sum_{s=1}^{S} (x_{gs} - \hat{x}_{gs})}{S-2}]^{1/2} .$$

$$\hat{x}_{gs} = \hat{\beta}_{0g} + r_g y_s .$$

$$\hat{\beta}_{0g} = \bar{x}_s - r_g \bar{y}_s .$$

The details for computation of s0 are shown in (Tusher, et al., 2001).

1. Calculate t_g , $1 \le g \le G$.

2. Compute v_p , the enrichment evidence score of pathway *p*, where

$$V_p = \frac{1}{G} \sum_{g=1}^{G} t_g z_{gp}$$

3. Permute sample labels C times, and calculate the permuted statistics, V_p^c , $1 \le c \le C$.

4. Data standardization. Suppose $F_1, \ldots, F_G$ are the empirical cumulative distribution functions of V_g, The data transformation function is

$$\phi_g(\cdot) = \Phi^{-1}\{F_g(\cdot)\}, g = 1, \ldots, G.$$

where $\Phi(\cdot)$ is the cumulative distribution function for standard normal. Data were standardized by $V_p^{(s)} = \phi_g(V_p), V_p^{c(s)} = \phi_g(V_p^{c(s)})$, $1 \le c \le C$, $1 \le g \le G$. For simplicity, we still denote $V_p^{(s)}$ and $V_p^{c(s)}$ by V_p and V_p^c .

5. Estimate the p-value of pathway *p* as $p(v_p) = \sum_{c=1}^{C} \sum_{p'=1}^{P} I(V_{p'}^c \ge V_p) \Big/ C \cdot P$ and similarly calculate $v_p^c = \sum_{c'=1}^{C} \sum_{p'=1}^{P} I(V_{p'}^{c'} \ge P_p^c) \Big/ C \cdot P$

6. Estimate π_0 , the proportion of non-enriched pathways in the meta-analysis, as $\hat{\pi}_0 = \dfrac{\sum_{p=1}^{P} I(p(v_p) \in A)}{P \cdot l(A)}$. We chose A=[0.5, 1] and thus *l*(*A*)=0.5.

7. Estimate q-value of pathway *p* as

$q(v_p)=\hat{\pi}_0\sum_{c=1}^{C}\sum_{p'=1}^{P}I(P_{p'}^{KS(c)}\le P_p^{KS})\Big/C\cdot\sum_{p'=1}^{P}I(P_{p'}^{KS}\le P_p^{KS})$. Pathways whose q-values are less than a pre-defined cutoff are considered as enriched pathways.

3.1.2 Examples for usage of MetaPath package

We present a typical example of usage of the MetaPath package. First, suppose there are *k* studies and all studies have been appropriately pre-processed and all probe IDs have been mapped to gene symbols. For each study, the data sets have been packaged as ExpressionSet objects. All *k* studies have been stored in a list. For example, two lung cancer studies (Table 4) have been packaged into a list entitled lung.cancer.study.

The summary of lung cancer data set can be checked by:

```
> lung.cancer.study
$Beer
ExpressionSet (storageMode: lockedEnvironment)
assayData: 4883 features, 96 samples
  element names: exprs
phenoData
  rowNames: AD10, AD2, ..., LN75  (96 total)
  varLabels and varMetadata description:
    Cluster.ID: the corresponding sample ID
    cluster: the cluster membership
    ...: ...
    testgroup: NA
    (15 total)
featureData
  featureNames: STAT1, GAPDH, ..., STAT5B  (4883 total)
  fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
Annotation:


$Bhat
ExpressionSet (storageMode: lockedEnvironment)
assayData: 5844 features, 155 samples
  element names: exprs
phenoData
  sampleNames: AD262, AD3, ..., AD1  (155 total)
  varLabels and varMetadata description:
    simple_annotation: NA
    CLASS: NA
    Sample: NA
    testgroup: NA
featureData
  featureNames: STAT1, GAPDH, ..., IGF2R  (5844 total)
  fvarLabels and fvarMetadata description: none
experimentData: use 'experimentData(object)'
Annotation:
```

The sample information has been store in the slot CLASS in each study.

```
>lung.cancer.study$Beer$CLASS
 [T] T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T
[39] T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T
[77] T T T T T T T T T T N N N N N N N N N N
```

where T stands for tumor tissue and N stands for normal tissue.

Suppose the pathway database has been transformed to a binary matrix named DB.matrix (METAPATHpackage offers a function to load Msig database and transfer to a binary matrix.)

```
> dim(DB.matrix)
[1]  639 5385
> DB.matrix[1:5,1:2]
```

```
                                           ALDH1A1     ALDH1A2
1_2_DICHLOROETHANE_DEGRADATION                   1           1
1_AND_2_METHYLNAPHTHALENE_DEGRADATION            0           0
41BBPATHWAY                                      0           0
ACE2PATHWAY                                      0           0
ACE_INHIBITOR_PATHWAY_PHARMGKB                   0           0
```

We run MAPE by:

```
>MAPE.obj=MAPE_KS(study=lung.cancerstudy, group=’CLASS’, DB.matrix=DB.matrix, size.min=15, size.max=500, nperm=500, method=’gene.permutation’)
```

Then the Figure 2.17 can be obtained by

```
>MAPE.plot(MAPE.obj)
```

3.1.3 Computational issues of MetaPath package

The MetaPath package is implemented with the R language. R is a scripting language that is not as fast as certain procedural programming languages, such as C. To accelerate the computational time, we carefully implemented the MetaPath package using the following two techniques:

1) Using matrix manipulation

   R has many built-in statistical test procedures, such as the KS test and Fisher's exact test, that can work on only one numeric vector of data values (for example, the expression values of one gene). If we applied the built-in KS test to thousands of genes (thousands of numeric vectors), it would be unfeasibly slow. To solve this problem, we implemented our own KS test/Fisher's exact test based on the matrix manipulation; this greatly reduced the computational time. In addition, we used a binary matrix to denote the pathway database; consequently, most of our MAPE procedures could be implemented by matrix manipulation.

2) Using a sparse matrix

   Although matrix manipulation can accelerate the computational time in the R environment, it requires a substantial amount of memory and the use of a large pathway database. Therefore, we transferred a pathway database to a numeric matrix $\{z_{gp}\}$ ($1 \leq g \leq G$, $1 \leq p \leq P$), to represent the pathway information of P pathways, where $z_{gp}=1$ when gene g belongs to pathway p and $z_{gp}=0$. Due to the existence of many zeros in the pathway database matrix, sparse matrix techniques were adopted in our MetaPath package; these had the dual effects of conserving the memory and reducing the computational time.

3.2 INCLUSION/EXCLUSION CRITERIA

In Chapter 2, for simplicity, we illustrated the MAPE analysis by combining only two studies. A more realistic example which is aimed to integrate large prostate cancer studies was used to discuss the inclusion/exclusion criteria of the MAPE analysis.

We collected 6 prostate cancer studies. A summary of the prostate cancer studies is listed in Table 3.1. Each study has two groups of samples: the normal group and tumor group. There are 3 different platforms for these studies (HGU95A/HUG95AV2, HGU133plus2 and cDNA platform). To make these studies comparable, probe IDs have been mapped to Gene Symbols. The microarray data have been pre-processed by the methods described in section 2.3.1.

Table 3.1 Summary of 6 prostate cancer studies

Study	Platform	Normal samples	Tumor samples	Probe IDs
Welsh (Welsh, et al., 2001)	HGU95A	9	25	12625
Singh (Singh, et al., 2002)	HGU95Av2	50	52	12625
Stuart (Stuart, et al., 2004)	HGU95Av2	50	38	12625
Yu (Yu, et al., 2004)	HGU95Av2	59	66	12625
Varambally (Varambally, et al., 2005)	HGU133plus2	6	7	54675
Lapointe (Lapointe, et al., 2004)	cDNA	41	62	44528

The consistency among all these 6 studies has been checked by our inclusion/exclusion criteria:

1) Sample size requirement. Studies that have fewer than 5 samples in each group are excluded. The array platform needs to measure more than 6,000 gene expression values.

2) Expert screening. Dr. Luo and Dr. Kaminski in University of Pittsburgh reviewed all studies to confirm that they meet high standard and all studies are related for information integration and meta-analysis.

3) Correlation of t-statistics among all studies.

Suppose genes in study k_1 and k_2 have been matched, $1\leq k_1\leq 6$, $1\leq k_2\leq 6$, and there are G common genes in total. We calculated the unequal variance t-statistics for each gene in study k_1 and k_2, denoted by t_{gk1}, t_{gk2}, $1\leq g\leq G$. Then the Pearson correlation between t_{gk1},t_{gk2} was computed to indicate the consistency between the study k_1 and k_2. The pair-wise comparison of consistency among all prostate studies is shown in Table 3.2, which indicates that the Lapointe data set has negative correlation with all other studies. Therefore, we excluded the Lapointe data set from our meta-analysis.

Table 3.2. The pair-wise comparison of consistency among all prostate studies

	Welsh	Singh	Stuart	Yu	Varambally	Lapointe
Welsh	1.00	0.54	0.77	0.62	0.47	-0.15
Singh	0.54	1.00	0.59	0.34	0.33	-0.10
Stuart	0.77	0.59	1.00	0.72	0.42	-0.17
Yu	0.62	0.34	0.72	1.00	0.43	-0.14
Varambally	0.47	0.33	0.42	0.43	1.00	-0.12
Lapointe	-0.15	-0.10	-0.17	-0.14	-0.12	1.00

3.3 MICORARRAY DATABASE FOR CHEMOTHERAPY RESEARCH

We collected drug-response related microarray studies and built a microarray database for chemotherapy research. The specific studies were listed in Table 3.3. In each study, the cancer type, the number of patients, the array platform, the drugs and patient's outcome were listed. For example, the Hess data set has 133 patients. The array platform is Affymetrix U133a. The patients were treated by cyclophosphamide, doxorubicin, fluorouracil and paclitaxel. The pathologic complete response was used the end point to indicate the patient's drug response. The gene expression of patients was measured before chemotherapy treatment. This data set has been widely used as a test set to validate the prediction of patient's clinical outcomes (Garman, et al., 2007; Huang, et al., 2007; Lee, et al., 2010).

This chemotherapy microarray database has great value for bioinformatics researchers in field of chemotherapy research. In Chapter 4, two chemotherapy studies related to identify of robust biomarkers and multi-drug response genes were performed based on this chemotherapy microarray database.

Table 3.3 Chemotherapy microarray database

Indication	1st Author	# Patients	Expression Platform	Drug(s)	Outcome
breast	Modlich	83	U133a	epirubicin cyclophosphamide	clinical response
breast	Hess	133	U133a	cyclophosphamide doxorubicin fluorouracil paclitaxel	pathologic complete response
breast	Chang	24	U95	focetaxel	clinical response
breast	Berthea	60	U133a	epirubicin cyclophosphamide	pathologic complete response
breast	Folgueira	51	cDNA	doxorubicin	clinical response

				cyclophosphamide	
breast	Sorlie		cDNA	paclitaxel	progression-free interval
breast	Lin	24	U133+2	epirubicin docetaxel	pathologic complete response
breast	Korde	21	U133+2	docetaxel capecitabine	clinical response
breast	Pawitan	126	U133a	cyclophosphamide methotrexate 5-fluorouracil	survial
breast	Bonnefoi	66	Affymetrix X3P	fluorouracil epirubicin cyclophosphamide	pathologic complete response
breast	Cleator	43	cDNA	cyclophosphamide doxorubicin	pathologic complete response
breast	ayers	42	cDNA	cyclophosphamide doxorubicin fluorouracil paclitaxel	pathologic complete response
breast	Hannemann	24	cDNA	doxorubicin cyclophosphamide or doxorubicin docetaxel	pathologic complete response
breast	Mina	45	RT-PCR	doxorubicin docetaxel	pathologic complete response
breast	Dressman	37	U133+2	cyclophosphamide methotrexate fluorouracil	pathologic complete response
breast	Paik	651	RT-PCR	tamoxifen cyclophosphamide, methotrexate 5-fluorouracil	distant Free recurrence
ovarian	Spentzos	68	U95a	platinum/taxane based chemotherapy	complete clinical response/remission
ovarian	Berchuck	65	U133a	platin-based combination chemotherapy	survival
rectal carcinomas	Ghadimi	30	cDNA	5-fluorouracil	survival
esophageal	Kihara	20	cDNA	cisplatin 5-fluorouracil	survival
NSCLC[1]	Hsu	59	U133a	cisplatin pemetrexed	clinical response
NSCLC	Kakiuchi	28	cDNA	iressa	pathologic complete response

[1] NSCLC: Non-small cell lung carcinoma

4.0 APPLICATIONS OF MEATA-ANALYSIS METHODS IN CHEMOTHERAPY RESEARCH

In Chapter 3, meta-analysis has been applied to pathway enrichment analysis. In this Chapter, we applied meta-analysis on genes to identify robust genomic biomarkers by combining multiple microarray studies. In Chapter 4.1, robust genomic biomarkers were identified by combining two independent microarray studies on breast cancer cell lines. In Chapter 4.2, genes associated with multiple drug responses were identified by meta-analysis method. These genes have the potential to be the biomarkers to distinguish patients who are unlikely to benefit from current chemotherapeutic drugs.

4.1 IDENTIFICATION OF ROBUST PHARMACOGENOMIC PREDICTORS ASSOCIATED WITH CHEMOTHERAPY TREATMENT IN BREAST CANCER BY META-ANALYSIS

4.1.1 Introduction

Breast cancer remains a significant cause of mortality in women (Jemal, et al., 2008). Even with multiple chemotherapy treatments available, individual patient responses to chemotherapy vary considerably and response rates, in general, remain poor with 30% of early-stage breast cancers recurring (Gonzalez-Angulo, et al., 2007). In an effort to maximize patient response to chemotherapy, pharmacogenomics-based testing is being used a means to identify patients that could benefit from specific chemotherapy treatments (Potti and Nevins, 2008). Recent work has expanded this concept by combining tumor gene expression profiling and clinical outcome data (Bertheau, et al., 2007; Hess, et al., 2006). While this method to date may not be accurate enough to identify specific gene differences between responder and non-responder patient groups (Pusztai, et al., 2007), identified gene signatures can prognosticate on cancer recurrence for specific breast cancer patient subgroups (Hess, et al., 2006; Potti, et al., 2006; Potti and Nevins, 2008; Salter, et al., 2008; Staunton, et al., 2001).

Several recent reviews discuss the strengths and limitations of the methods used to develop pharmacogenomic predictors of response from patient samples and cell lines (Kim, et al., 2009; Marchionni, et al., 2008; Potti and Nevins, 2008; Sotiriou and Pusztai, 2009). One method involves splitting the sample population such that data from a subset of patients are used for the pharmacogenomic predictor discovery and the data from remaining patients are used for its validation. This approach has limited utility when multiple standard-of-care treatments are available for testing (Potti and Nevins, 2008) since large numbers of clinically homogenous patients would be required for validation (Marchionni, et al., 2008). Recently, several groups of researchers have attempted to overcome some of these limitations by using immortalized cell lines as a proxy for patient outcomes in supervised machine-based learning models (Lee, et al., 2007; Potti, et al., 2006; Salter, et al., 2008; Staunton, et al., 2001). While several studies have used NCI-60 drug sensitivity data and Affymetrix gene expression data to develop predictors of response to chemotherapies and to demonstrate the capacity to predict response in patients (Hsu, et al., 2007; Potti, et al., 2006; Potti and Nevins, 2008; Salter, et al., 2008), others have not been able to confirm these results using similar approaches but different methods for measuring in vitro responses (Liedtke, et al., 2009).

The purpose of this study was to identify robust genomic biomarkers associated with chemotherapy treatment by meta-analysis method. We used 15 breast cancer cell lines and chemotherapy response data were generated by exposing these cell lines to various chemotherapy assays to determine in vitro the sensitivity of each cell line to specific chemotherapies (Kornblith, et al., 2004; Kornblith, et al., 2003). For the second part, pharmacogenomic predictors developed from breast cancer cell lines were then validated by using genomic data from independent clinical trials.

4.1.2 Methods

4.1.2.1 Microarray data sets and pre-processing

Three publicly available data sets, Liedtke (Liedtke, et al., 2009), Neve (Neve, et al., 2006), Hoeflich (Hoeflich, et al., 2009), were used to identify robust pharmacogenomic predictors associated with breast cancer cell lines. The raw microarray data were processed by the software package RMA (Bolstad, et al., 2003; Irizarry, et al., 2003; Irizarry, et al., 2003) for the background adjustment and quantitative normalization. The processed data were log2-transformed. Non-specific gene filtering was performed to filter out probes which satisfy one of the following two criterions: 1) Interquartile range (IQR) was less than the median of IQR values of all genes. 2) Median expression values less than 100. The cell line's GI50 was measured by Liedtke et al. (Liedtke, et al., 2009) and used to indicate the cell line's drug sensitivity to the drug paclitaxel.

4.1.2.2 Biomarker identification

Let x_{gsk} denote the gene expression value for gene g, cell line s in study k, s, $1 \le g \le G$, $1 \le s \le S$, $1 \le k \le K$. Let y_{sk} denote the GI50 value for the cell line s in study k. The regression coefficients β_{1gk} for gene g in study k was estimated using a standard linear regression model $y_{sk} = \beta_{0gk} + \beta_{1gk} x_{gsk} + \epsilon_{sgk}$, where ε is the normal error. Let $t_{gk} = \beta_{1gk} / (s_{gk} + s_{0k})$, where s_{gk} is the standard deviation of β_{1gk}. The β_{1gk} was calculated by the following formulas:

$$\beta_{1gk=} \frac{\sum_{s=1}^{S} y_s (x_{gsk} - \bar{x}_{gk})}{\sum_{s=1}^{S} (y_{sk} - \bar{y}_k)^2},$$

where $\bar{x}_{gk} = \sum_{s=1}^{S} x_{gsk} / s$, $\bar{y}_k = \sum_{s=1}^{S} y_{sk} / s$.

$$s_{gk} = \frac{\hat{\sigma}_{gk}}{[\sum_{s=1}^{S} (y_{sk} - \bar{y}_k)^2]^{1/2}} .$$

$$\hat{\sigma}_{gk} = [\frac{\sum_{s=1}^{S} (x_{gsk} - \hat{x}_{gsk})}{S-2}]^{1/2} .$$

$$\hat{x}_{gsk} = \hat{\beta}_{0gk} + r_{gk} y_{sk} .$$

$$\hat{\beta}_{0gk} = \bar{x}_{sk} - r_{gk} \bar{y}_{sk} .$$

The details for computation of s0 are shown in (Tusher, et al., 2001).

The procedure for identification of robust pharmacogenomic predictors was listed as follows:

Suppose there are a total of G genes and K studies (K=3 for this case).

I. Individual-study analysis:

a. Compute the t$_{gk}$ for each gene in each study.

b. Permute the group labels in each study for *B* times, and similarly calculate the permuted statistics, $t_{gk}^{(b)}$, where $1 \le g \le G,\ 1 \le k \le K,\ 1 \le b \le B.$

c. Estimate the p-value of t_{gk} as $p_{gk} = \dfrac{\sum_{b=1}^{B}\sum_{g'=1}^{G} I\left(|\, t_{g'k}^{(b)} \,|\ge|\, t_{gk} \,|\right)}{B \cdot G}$ and similarly calculate

$$p_{gk}^{(b)} = \frac{\sum_{b'=1}^{B}\sum_{g'=1}^{G} I\left(|\, t_{g'k}^{(b')} \,|\ge|\, t_{gk}^{(b)} \,|\right)}{B \cdot G}.$$

d. Estimate $\pi_0(k)$, the proportion of non-DE genes, as $\hat{\pi}_0(k) = \dfrac{\sum_{g=1}^{G} I(p_{gk} \in A)}{G \cdot l(A)}$ (Storey, 2002). We chose A=[0.5, 1] and thus *l*(*A*)=0.5.

e. Estimate the q-value of t_{gk} as $q_{gk} = \dfrac{\hat{\pi}_0(k) \cdot \sum_{b=1}^{B}\sum_{g'=1}^{G} I\left(|\, t_{g'k}^{(b)} \,|\ge|\, t_{gk} \,|\right)}{B \cdot \sum_{g'=1}^{G} I(|\, t_{g'k} \,|\ge|\, t_{gk} \,|)}$. DE genes detected from each individual study are denoted by $G_k = \{g : q_{gk} \le 0.05\}$.

II. Meta-analysis:

a. The maximum p-value statistic (maxP) is used for meta-analysis: $V_g = \max_{1 \le k \le K} p_{gk}$.

Define $V_g^{(b)} = \max_{1 \le k \le K} p_{gk}^{(b)}$.

b. Estimate the p-value of the genes in meta-analysis as $p(V_g) = \dfrac{\sum_{b=1}^{B}\sum_{g'=1}^{G} I\left(V_{g'}^{(b)} \le V_g\right)}{B \cdot G}$

.

c. Estimate π_0, the proportion of non-DE genes in the meta-analysis, as

$\hat{\pi}_0 = \dfrac{\sum_{g=1}^{G} I(p(V_g) \in A)}{G \cdot l(A)}$. We chose A=[0.5, 1] and thus *l*(*A*)=0.5.

d. Estimate the q-value in the meta-analysis as $q(V_g) = \dfrac{\hat{\pi}_0 \cdot \sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(V_{g'}^{(b)} \le V_g\right)}{B \cdot \sum_{g'=1}^{G} I(V_{g'} \le V_g)}$. DE genes detected by the meta-analysis are denoted as $G_{meta} = \{g : q(V_g) \le 0.05\}$.

4.1.2.3 Validation of the pharmacogenomic predictors

Publical available microarray datasets and published literature were reviewed to identify gene expression data useful for validating the pharmacogenomic predictors. An independent breast cancer patient dataset (Hess data) were used to test the accuracy of pharmacogenomic predictors (Hess, et al., 2006). Hess dataset contained expression data generated using the Hgu133A RNA expression array with tumor samples from patients with breast cancer as well as information on the treatments received by each patient and their outcomes. The gene expression profiles of patients were measured before chemotherapy treatment. The patient's complete responses (pCR) were tested after treatment by the drug combination of cyclophosphamidem doxorubicin, fluorouracil and paclitaxel to demonstrate the chemotherapy efficacy.

Supervised principal components regression (Bair and Tibshirani, 2004) was adopted to develop the pharmacogenomic predictor. Suppose a data matrix $\{x_{gs}\}$ ($1 \le g \le G$, $1 \le s \le S$) represents the gene expression intensity of gene *g* and sample *s*. Let $\{y_s\}$ ($1 \le s \le S$) represent the AUC for cell line *s*,. We first calculate t_g, the association score between gene *g* and y_s, $1 \le g \le G$, where $t_g = {}^{r_g}\!/\!{}_{s_g}$; r_g is the linear regression coefficient between x_{gs} and y_s. $1 \le s \le S;$ s_g is the standard error of r_g. Genes were selected if their association score t_g were larger than the threshold, where the threshold was estimated by cross-validation in the training set. A reduced data matrix on these selected genes was formed, and the first principal component based on the reduced data matrix was calculated. The first principal component was used in a regression model to predict the patient's outcome. More details about the supervised principal components regression is available at (Bair, et al., 2006).

4.1.3 Results

255 genes was identified as DE genes whose q-values by meta-analysis less than 0.01. These 255 genes were used as pharmacogenomic predictor and were validated on the expression data from the Hess dataset. The patient's pCR in Hess data was predicted using the supervised principle component regression (Bair, et al., 2006).

. The prediction results were shown in Figure 4.1. When using top 50 genes which have the smallest q-values by meta-analysis, the accuracy was 63.6%, sensitivity was 76.5% and specificity was 59.1%. The area under receiver operator characteristic curves (AU-ROC) was 0.758 (Figure 4.1). We also examined whether this pharmacogenomic predictor was affected by the number of included genes. As the number of genes included in the pharmacogenomic predictor increased, few effects were observed on the accuracy, sensitivity and specificity of the predictor for treatment with paclitaxel (Figure 4.1), indicating a robust predictor.

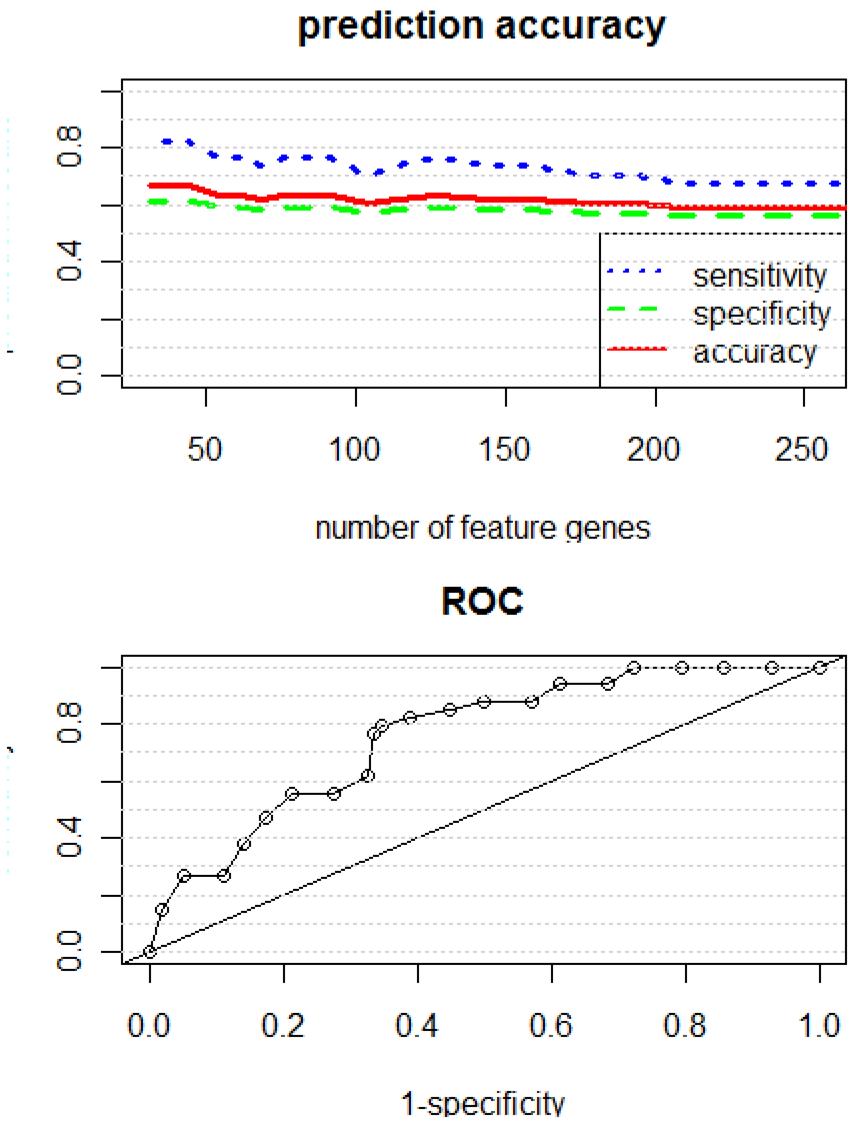


Figure 4.1 Prediction accuracy and the ROC curve.

4.1.4 Conclusions

This study demonstrates use of GI50 as a supervisor to grade the contribution of gene expression in predicting *in vitro* responses of patient-derived primary cultures to various chemotherapy treatment regimens (Kornblith, et al., 2004; Kornblith, et al., 2003). Using the GI50 data on breast cancer cell lines, we were able to identify pharmacogenomic predictors of patient response to several standard-of-care chemotherapies for breast cancer. These pharmacogenomic predictors were validated by the use of an independent genomic datasets, which also contained data on patient treatments and outcomes. Our pharmacogenomic predictors had sufficiently high accuracy, sensitivity and specificity to warrant further testing. Importantly, our multigene predictors remained stable even as the number of genes included in the predictor increases, suggesting that GI50 trained predictors may provide indications of chemosensitivity and chemoresistance that are specific to the chemotherapy treatment tested and are not a result of general chemotherapy sensitivity (Pusztai, et al., 2007). Thus, our study indicates that use of the ChemoFx results as the supervisor is feasible to identify multigene predictors of responses to chemotherapy for breast cancer.

Two methods have been adopted to develop pharmacogenomic predictors, one based on pharmacogenomic data from patients while the other one is based on cell lines. The first method involves splitting data from an existing cohort into separate test and validation sets; however, this method restricts the strength of the pharmacogenomic predictors because of the large number of cases required for each set. The second method involves the use of established cell lines to train data to identify potential pharmacogenomic predictors of chemosensitivity and resistance and then validating the pharmacogenomic predictors using data from a patient cohort (Liedtke, et al., 2009; Potti, et al., 2006; Salter, et al., 2008; Staunton, et al., 2001). The advantage to this approach is that the use of cell lines is much faster and less costly to perform than the use of data from a prospectively collected patient cohort.

Potti et al (Potti, et al., 2006) first reported the use of NCI-60 cell lines to develop pharmacogenomic predictors; however, their results could not be replicated by an independent group (Liedtke, et al., 2009). NCI-60 cell lines have various histological origins, which may introduce a confounding variable in the development of the pharmacogenomic predictor. In the current report, we demonstrate the ability to use cell lines trained using the GI50 assay to predict patient responses. The use of the GI50 assay allowed for the selection of malignant cells within each cell line and therefore supports the concept of using cell lines of identical histological origin to develop predictors of patient chemotherapy response.

Thus, our data are quite promising for the feasibility of using the in vitro drug responses for the identification of pharmacogenomic predictors of response to chemotherapy treatment for breast cancer patients. Future studies will examine the use of drug responses from primary cultures of patient tumors to develop pharmacogenomic predictors of breast cancer patient responses to chemotherapy treatment.

4.2 IDENTIFICATION OF MULTI-DRUG RESPONSE GENES BY META-ANALYSIS IN HUMAN BREAST CANCER CELL LINES

A major obstacle in the effective treatment of cancer with chemotherapeutic agents is the phenomenon of multidrug resistance. In breast cancer patients, multiple chemotherapy drugs have been widely used. Standards of care have involved various neoadjuvant approaches to chemotherapy and surgical resection with the greatest success occurring when tumor tissue is surgically removed and patients are subsequently treated with chemotherapy. Success rates with primary breast cancer, caught early, are now approaching 80% (Haigh, et al., 2000). However, chemotherapeutic agents alone have an efficacy of about 50% (Buzdar, et al., 2005). Additionally, chemotherapeutic agents are less effective in treating recurrent disease. A contributing factor is the resistance to current chemotherapeutic drugs. Moreover, many tumor cells resistant to one drug often have different degrees of resistance to other chemotherapeutic drugs. This phenomenon is commonly referred as multidrug resistance (MDR) (Chang, et al., 2003; Gianni, et al., 2005; Hess, et al., 2006; Iwao-Koizumi, et al., 2005; Liedtke, et al., 2009; Paik, et al., 2006; van de Vijver, et al., 2002; Wang, et al., 2005). Understanding the molecular mechanisms of MDR has important biological significance and potential clinical utility. It is important to identify patients who will not respond to current chemotherapeutic drugs and avoid giving them unnecessary treatment. Furthermore, understanding the mechanisms of MDR will further facilitate drug selection studies, and perhaps identify new therapeutic targets.

Cancer cell lines have been extensively used for investigating mechanisms of drug response. MDR genes are identified by integrating gene expression profiles and drug response patterns. To date, many research groups have studied MDR in NCI-60 cells because their gene expressions have been well characterized and they have been examined for resistance to numerous drugs (Dan, et al., 2002; Kang, et al., 2004; Mariadason, et al., 2003; Staunton, et al., 2001). Since NCI-60 is composed of cells with different origins, such as breast, prostate, lung, colorectal, renal, ovarian, prostate, lung, leukaemias, melanomas and neural system, the mechanisms identified by these studies are presumably independent of tumor cell histology. Other investigations have focused on specific cancer cell lines including gastric (Kang, et al., 2004) , and colon cancer (Mariadason, et al., 2003). However, no studies have yet been done in breast cancer cell lines. Given the multidrug resistance seen in breast cancer patients, identifying MDR genes in breast cancer patients may have considerable clinical implications. In this paper we used the GI50 to determine the sensitivity of 16 well-studied breast cancer cell lines to 4 chemotherapy agents commonly used to treat breast cancer patients: paclitaxel, cyclophosphamide, fluorouracil and doxorubicin. Meta-analysis method was applied to identify genes that are related to multidrug resistance in breast cancer associated with chemotherapy treatment.

4.2.1 Materials and method

4.2.1.1 Microarray data sets and pre-processing

A publicly available data set (Neve, et al., 2006) was used to identify MDR genes associated with four drugs: paclitaxel, cyclophosphamide, fluorouracil and doxorubicin in breast cancer cell lines. The raw microarray data were processed by the software package RMA (Bolstad, et al., 2003; Irizarry, et al., 2003; Irizarry, et al., 2003) for the background adjustment and quantitative normalization. The processed data were log2-transformed. Non-specific gene filtering was performed to filter out probes which satisfy one of the following two criterions: 1) Interquartile range (IQR) was less than the median of IQR values of all genes. 2) Median expression values less than 100. The 19 breast cancer cell line's GI50 was measured by Liedtke et al. (Liedtke, et al., 2009) and used to indicate the cell line's drug sensitivity to the drug paclitaxel, cyclophosphamide, fluorouracil and doxorubicin.

4.2.1.2 Identification of genes related to multidrug response

To analyze how gene expression is related to multi drug response in breast cell lines, meta-analysis was performed to identify genes which response to at least 3 drugs in breast cell lines. The details of the algorithms that were used to perform the meta-analysis are as follows:

Let x_{gsk} denote the gene expression value for gene g, cell line s for drug k, s, $1 \leq g \leq G$, $1 \leq s \leq S$, $1 \leq k \leq K$. Let y_{sk} denote the GI50 value for the cell line s for drug k. The regression coefficients β_{1gk} for gene g in study k was estimated using a standard linear regression model $y_{sk} = \beta_{0gk} + \beta_{1gk} x_{gsk} + \epsilon_{sgk}$, where ε is the normal error. Let $t_{gk} = \beta_{1gk} / (s_{gk} + s_{0k})$, where s_{gk} is the standard deviation of β_{1gk}. The β_{1gk} was calculated by the same formulas in section 4.1.2.2.

The procedure for identification of MDR genes is similar as the procedures to identify robust biomarkers in section 4.1.2.2. The difference is that the rth rank statistic is used instead of the maxP statistic to identify genes response to at least 3 drugs. Details of the algorithm were listed as follows:

Suppose there are a total of *G* genes and *K* drugs (*K*=4 for this case).

III. Individual-study analysis:

a. Compute the t_{gk} for each gene for each drug.

b. Permute the group labels in each study for *B* times, and similarly calculate the permuted statistics, $t_{gk}^{(b)}$, where $1 \leq g \leq G,\ 1 \leq k \leq K,\ 1 \leq b \leq B$.

c. Estimate the p-value of t_{gk} as $p_{gk} = \dfrac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(|t_{g'k}^{(b)}| \geq |t_{gk}|\right)}{B \cdot G}$ and similarly calculate

$$p_{gk}^{(b)} = \frac{\sum_{b'=1}^{B} \sum_{g'=1}^{G} I\left(|t_{g'k}^{(b')}| \geq |t_{gk}^{(b)}|\right)}{B \cdot G}.$$

d. Estimate $\pi_0(k)$, the proportion of non-DE genes, as $\hat{\pi}_0(k) = \dfrac{\sum_{g=1}^{G} I(p_{gk} \in A)}{G \cdot l(A)}$ (Storey, 2002). We chose A=[0.5, 1] and thus $l(A)$=0.5.

e. Estimate the q-value of t_{gk} as $q_{gk} = \dfrac{\hat{\pi}_0(k) \cdot \sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(|t_{g'k}^{(b)}| \geq |t_{gk}|\right)}{B \cdot \sum_{g'=1}^{G} I(|t_{g'k}| \geq |t_{gk}|)}$. DE genes detected from each individual study are denoted by $G_k = \{g : q_{gk} \leq 0.05\}$.

IV. Meta-analysis:

a. The r-th rank statistic is used for meta-analysis: $V_g = p_{gk(3)}$. Define $V_g^{(b)} = p_{gk(3)}^{(b)}$.

b. Estimate the p-value of the genes in meta-analysis as $p(V_g) = \dfrac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(V_{g'}^{(b)} \le V_g\right)}{B \cdot G}$

.

c. Estimate π_0, the proportion of non-DE genes in the meta-analysis, as

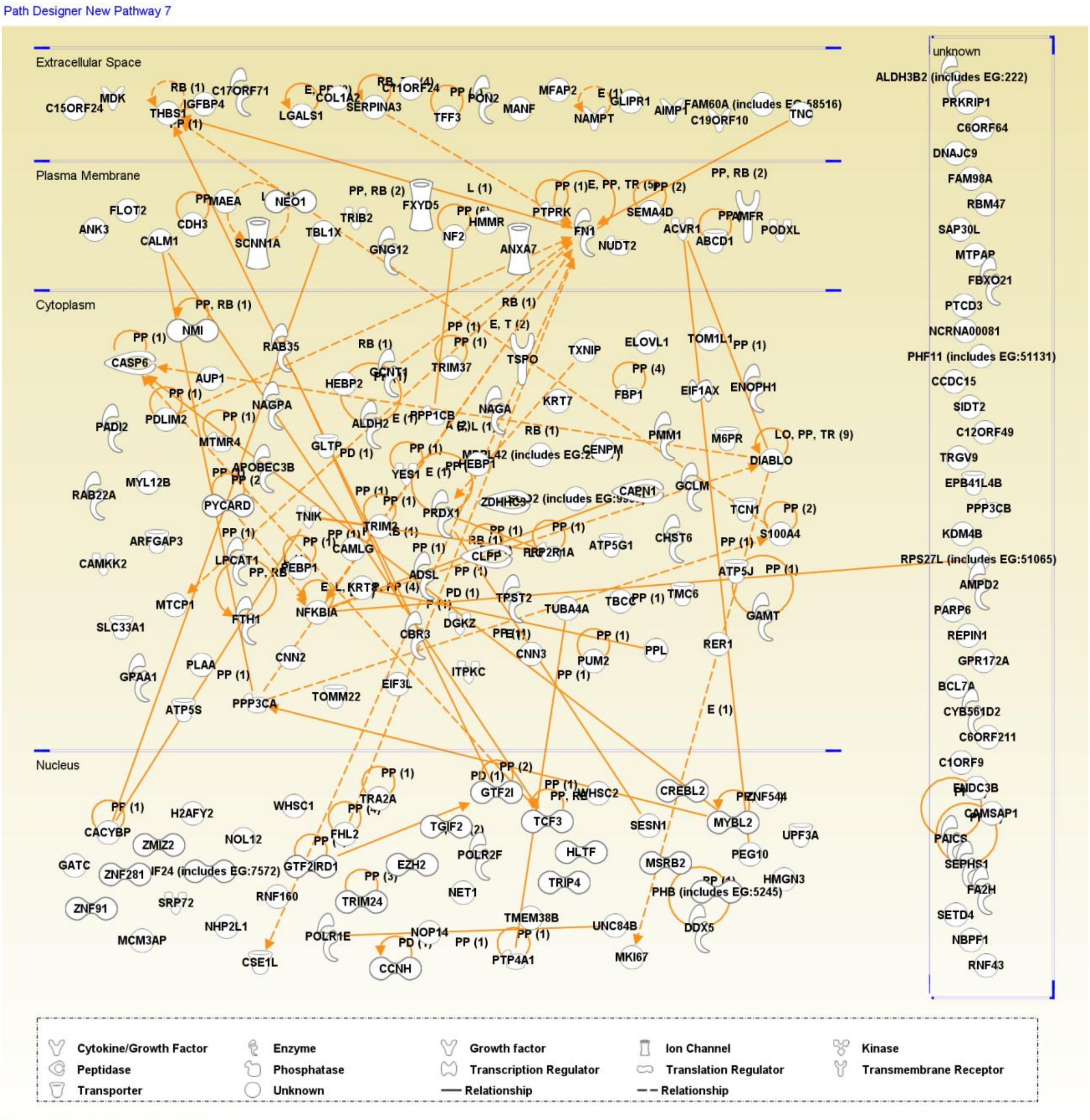
$\hat{\pi}_0 = \dfrac{\sum_{g=1}^{G} I(p(V_g) \in A)}{G \cdot l(A)}$. We chose A=[0.5, 1] and thus *l*(*A*)=0.5.

d. Estimate the q-value in the meta-analysis as $q(V_g) = \dfrac{\hat{\pi}_0 \cdot \sum_{b=1}^{B} \sum_{g'=1}^{G} I\left(V_{g'}^{(b)} \le V_g\right)}{B \cdot \sum_{g'=1}^{G} I(V_{g'} \le V_g)}$.

MDR genes detected by the meta-analysis are denoted as $G_{meta} = \{g : q(V_g) \le 0.05\}$.

4.2.2 Results and discussions

Through pharmacogenomic analysis, 200 genes were identified to be related to multidrug resistance in breast cancer cell lines. The function categories and locations of these MDR genes were shown in Figure 4.2 and Table 4.1. Functional analysis by Ingenuity Pathway Analysis (Ingenuity Systems) software indicates these genes execute the function as kinase, transcription regulator, translation regulator, transmembrane receptor and transporter.



© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

Figure 4.2 MDR genes associated with drug paclitaxel, cyclophosphamide, fluorouracil and doxorubicin in breast cancer cell lines.

Table 4.1 Categories and locations of MDR genes.

Location	Type(s)	Total
Cytoplasm	enzyme	20
	kinase	5
	other	30
	peptidase	4
	phosphatase	5
	transcription regulator	2
	translation regulator	1
	transmembrane receptor	1
	transporter	10
Cytoplasm Total		78
Extracellular Space	cytokine	3
	enzyme	2
	growth factor	1
	other	13
Extracellular Space Total		19
Nucleus	enzyme	3
	kinase	1
	other	20
	phosphatase	1
	transcription regulator	17
	transporter	2
Nucleus Total		44
Plasma Membrane	enzyme	2
	ion channel	3
	kinase	3
	other	9
	phosphatase	2
	transcription regulator	1
	transmembrane receptor	1
	transporter	1
Plasma Membrane Total		22
unknown	enzyme	7
	other	27
	phosphatase	1
	transporter	1
unknown Total		36
(blank)	(blank)	
(blank) Total		
Grand Total		199

Current treatment guidelines recommend a consideration of chemotherapy for a majority of cancer patients; however, it is helpful to distinguish those patients who are not good candidates for chemotherapy. MDR genes have the potential to be such a biomarker. Although various clinical factors, including ER, PR, and grade have been related to multidrug response, MDR genes as a biomarker can provide additional information. Therefore, integrating clinical information and MDR information may assist us to better identify patients who are candidates for chemotherapy.

To date, both tumor tissue and cancer cell lines have been used for drug response studies. Several studies have been performed using tumor tissue from breast cancer patients, and gene expression profiles associated with clinical outcome have been identified. However, there are major drawbacks to using patient tumor tissue for these studies. These drawbacks include a limited source of tissue and the long time necessary to assess clinical outcome. Using cell lines has the advantage of overcoming these obstacles.

5.0 CONCLUSIONS AND FUTURE DIRECTIONS

In this thesis, we applied meta-analysis methods for combining genomic studies on pathway enrichment analysis and biomarker detection. In Chapter 2, we formulated a framework of two meta-analysis approaches for pathway enrichment analysis, namely MAPE_G, which combines statistical significance at the gene level, and MAPE_P, which combines at the pathway level. In general, MAPE_P has the advantage of not requiring gene matching across studies and is often more powerful. MAPE_G is, however, usually more powerful if the majority of genes across studies can be properly matched. We proposed an automated integrated approach, namely MAPE_I, to accommodate the advantages of MAPE_G and MAPE_P and to capture all pathways of potential biological interest. Our simulation study characterized conditions when and how MAPE_G and MAPE_P outperform one another and verified the robust performance of MAPE_I. Applications to breast cancer cell line drug response and lung cancer demonstrated similar conclusions and identified previously verified pathways related to drug response and carcinogenesis. Meta-analysis identified more pathways than individual studies. The MAPE_I procedure integrated results from MAPE_P and MAPE_G. To our knowledge, this is the first study to systematically investigate and develop meta-analysis approaches for pathway enrichment analysis.

In Chapter 3, a software package, MetaPath, was implemented to perform MAPE analysis. MetaPath provided functions to perform MAPE analysis on microarray data with binary, continuous responses and survival data. The sparse matrix technique has been adopted in MetaPath package to speed up the computation of MAPE analysis. The MetaPath package was written using R language and can be installed in R environment. In addition to MetaPath package, the practical issues of MAPE analysis were also discussed. The inclusion/exclusion criteria of the MAPE analysis has been proposed to avoid low-quality studies in meta-analysis.

In Chapter 4, we first applied meta-analysis to identify robust genomic biomarkers related to chemotherapy response in breast cancer cell lines. We demonstrated the feasibility of using the in vitro breast cancer cell line's drug responses to predict the response to chemotherapy treatment for breast cancer patients. Then we applied meta-analysis to detect multi-drug response genes in human breast cancer cell lines. These genes have the potential to be the biomarkers to distinguish patients who are unlikely to benefit from current chemotherapeutic drugs.

Our future work will focus on the following two directions:

1. Hierarchical MAPE analysis for pathways (MAPE_H)

The hierarchical MAPE scheme will combine genomic studies with similar characteristics at the first hierarchy and with potentially different but related characteristics at the second hierarchy. We hypothesize that the hierarchical MAPE will more flexibly integrate information from a wide range of genomic studies to meaningfully answer biological questions. An example of MAPE_H analysis was shown in Figure 5.1. In Figure 5.1A, the first step (Step I in Figure 5.1A) is aimed to identify pathways related to one particular drug's response and the second step (Step II in Figure 5.1A) is aimed to identify pathways related to some different drugs. Specifically, in Step IA, MAPE_I is applied to identify pacilitaxel related pathway by combining two similar genomic studies while Step 1B is to discover doxorubicin related pathways. The goal of Step IA and IB is to find consistent enriched pathways across two homogeneous studies, thus

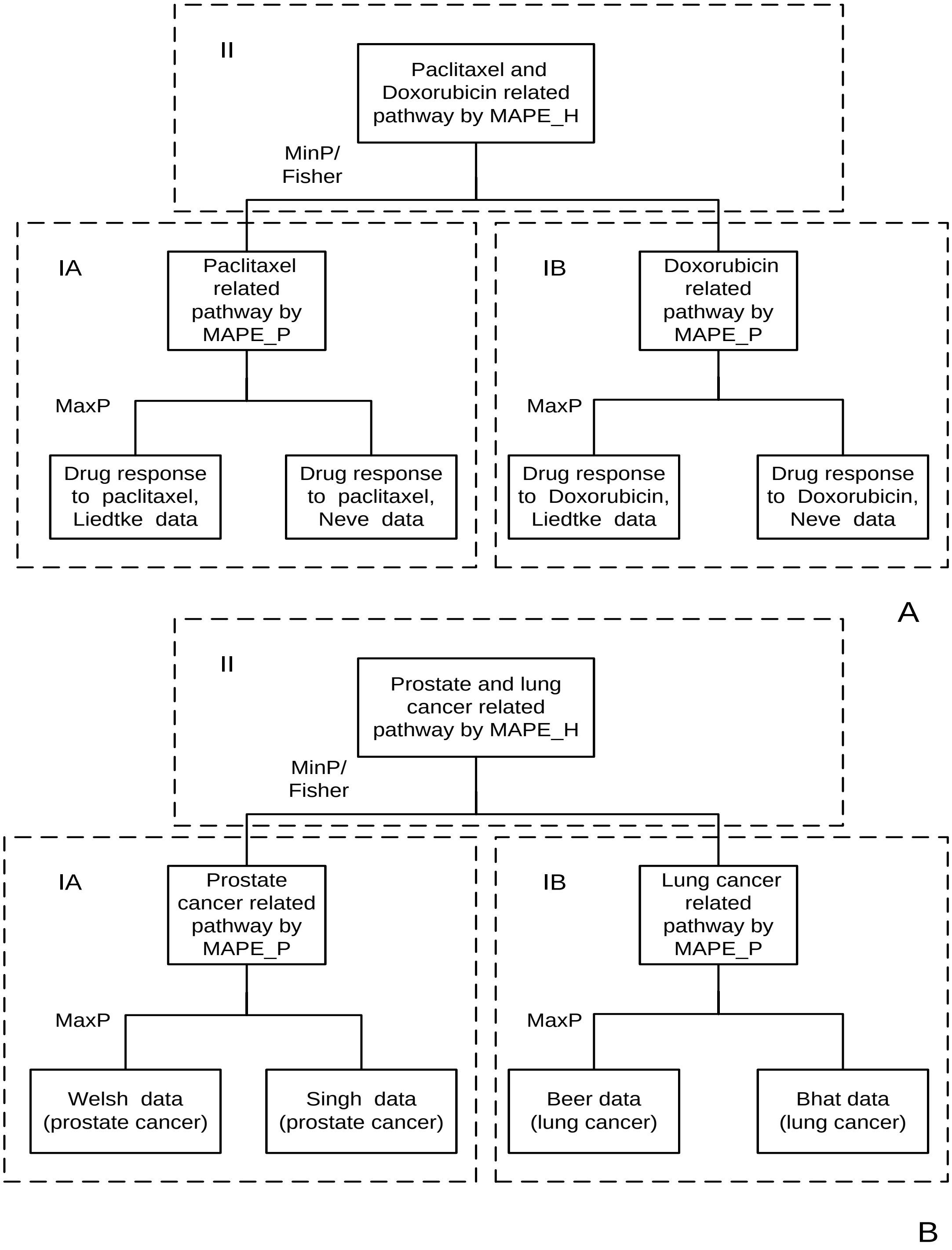

II
Paclitaxel and Doxorubicin related pathway by MAPE_H
MinP/ Fisher
IA
Paclitaxel related pathway by MAPE_P
IB
Doxorubicin related pathway by MAPE_P
MaxP
MaxP
Drug response to paclitaxel, Liedtke data
Drug response to paclitaxel, Neve data
Drug response to Doxorubicin, Liedtke data
Drug response to Doxorubicin, Neve data
A
II
Prostate and lung cancer related pathway by MAPE_H
MinP/ Fisher
IA
Prostate cancer related pathway by MAPE_P
IB
Lung cancer related pathway by MAPE_P
MaxP
MaxP
Welsh data (prostate cancer)
Singh data (prostate cancer)
Beer data (lung cancer)
Bhat data (lung cancer)
B

Figure 5.1 Examples for MAPE_H analysis.

the maxP statistics is adopted. In Step II, another level of meta-analysis is applied to discover the either pacilitaxel or doxorubicin related pathways. For this purpose, either minP or Fisher's statistic can apply. Similar analysis can be performed on prostate and lung cancer studies. In

Figure 5.1B, the first level of meta-analysis is to identify enriched pathways for prostate or lung cancer studies. The second level analysis is to combine lung and prostate cancer studies to investigate the pathways respective to both of the cancer types.

2. Evaluation and comparison of parameters/methods in MAPE procedures.

As was discussed in the Chapter 2 and 3, many meta-analysis techniques and pathway enrichment analysis methods have been developed in the past few years. This paper provides an initial investigation of a unified framework. Conceptually, any meta-analysis technique and pathway enrichment method can be combined under the proposed framework. Among the many available methods in both areas, evaluation of different method selection and the choice of a best method is the future direction.

APPENDIX A

QVALUES OF ENRICHED PATHWAYS FOR DRUG RESPONSE STUDY

Q-values of enriched pathways detected by individual studies and MAPE methods in drug response data (column 3-7: q-value threshold 0.05 and significant q-values marked in red) and categories (column 8-9) that correspond to Figure 6 in the manuscript. “Categories comparing MAPE_P, MAPE_G & MAPE_I” correspond to the categories in Figure 2.15A and 2.15B. “Categories comparing Liedtke, Neve & MAPE_I” correspond to the categories in Figure 2.15C and 2.15D. CA: Categories comparing MAPE_P, MAPE_G & MAPE_I. CB: Categories comparing Liedtke, Neve & MAPE_I”

Pathway	Descriptions	Liedtke	Neve	MAPE_P	MAPE_G	MAPE_I	CA	CB
CELL_MOTILITY	Any process involved in the controlled movement of a cell.	0.167	0.291	0.038	0.126	0.071	I	IV
CORDERO_KRAS_KD_VS_CONTROL_UP	Genes upregulated in kras knockdown vs control in a human cell line	0.004	0.411	0.072	0.076	0.105	I	II
LEE_CIP_UP	Genes up-regulated in hepatoma induced by ciprofibrate	0.113	0.514	0.108	0.067	0.107	I	II
IRITANI_ADPROX_VASC	BLOOD VASCULAR EC	0.263	0.273	0.036	0.073	0.072	I	IV
LI_FETAL_VS_WT_KIDNEY_UP	These are genes identified by simple statistical criteria as differing in their mRNA expresssion between WTs and fetal kidneys LOW	0.024	0.088	0.000	0.042	0.000	I	I
GAMMA_UNIQUE_FIBRO_DN	Down-regulated at any timepoint by treatment of human fibroblasts with gamma radiation, but not by UV lght or 4-NQO	0.395	0.201	0.073	0.066	0.103	I	IV
HSC_LTHSC_SHARED	Up-regulated in mouse long-term functional hematopoietic stem cells from both adult bone marrow and fetal liver (Cluster i, LT-HSC Shared)	0.117	0.428	0.071	0.070	0.109	I	II
TGFBETA_ALL_UP	Upregulated by TGF-beta treatment of skin fibroblasts, at any timepoint	0.003	0.044	0.000	0.000	0.000	I	I
ADIP_VS_PREADIP_DN	Downregulated in mature murine adipocytes (7 day differentiation) vs. preadipocytes (6 hr differentiation)	0.328	0.412	0.082	0.136	0.114	I	IV
LVAD_HEARTFAILURE_UP	Upregulated in the left ventricle myocardium of patients with heart failure following implantation of a left ventricular assist device	0.126	0.400	0.064	0.136	0.113	I	II
EGF_HDMEC_UP	Up-regulated in human dermal (foreskin) microvascular endothelial cells that were stimulated to proliferate with prolonged EGF treatment, versus non-stimulated quiescent controls.	0.413	0.395	0.070	0.080	0.111	I	IV
TGFBETA_EARLY_UP	Upregulated by TGF-beta treatment of skin fibroblasts at 30 min (clusters 1-3)	0.005	0.133	0.002	0.048	0.002	I	I

BRCA_ER_POS	Genes whose expression is consistently positively correlated with estrogen receptor status in breast cancer - higher expression is associated with ER-positive tumors	0.376	0.060	0.062	0.052	0.073	I	III
CMV_8HRS_DN	Downregulated at 8hrs following infection of primary human foreskin fibroblasts with CMV	0.149	0.414	0.075	0.056	0.076	I	II
CMV_24HRS_DN	Downregulated at 24hrs following infection of primary human foreskin fibroblasts with CMV	0.149	0.292	0.041	0.053	0.075	I	II
HSC_LTHSC_FETAL	Up-regulated in mouse long-term functional hematopoietic stem cells from fetal liver (LT-HSC Shared)	0.117	0.428	0.071	0.070	0.109	I	II
TGFBETA_LATE_UP	Upregulated by TGF-beta treatment of skin fibroblasts only at 1-4 hrs (clusters 4-6)	0.429	0.117	0.082	0.044	0.075	I	III
AGEING_KIDNEY_SPECIFIC_UP	Up-regulation is associated with increasing age in normal human kidney tissue from 74 patients, and expression is higher in kidney than in whole blood	0.096	0.159	0.001	0.045	0.003	I	II
AGEING_KIDNEY_UP	Up-regulation is associated with increasing age in normal human kidney tissue from 74 patients	0.144	0.022	0.001	0.022	0.002	I	I
CMV_ALL_DN	Downregulated at any timepoint following infection of primary human foreskin fibroblasts with CMV	0.069	0.364	0.055	0.004	0.003	I	II
ESR_FIBROBLAST_UP	Up-regulated in the environmental stress response in human fibroblasts (regulated similarly by gamma and UV rediation and 4-NQO)	0.514	0.303	0.137	0.098	0.126	I	IV
HSA01430_CELL_COMMUNICATION	Genes involved in cell communication	0.055	0.536	0.128	0.070	0.118	I	II
PASSERINI_ADHESION	Genes associated with cellular adhesion that are differentially expressed in endothelial cells of pig aortas from regions of disturbed flow (inner aortic arch) versus regions of undisturbed laminar flow (descending thoracic aorta).	0.387	0.189	0.074	0.284	0.114	II	IV
HADDAD_HSC_CD10_UP	Genes upregulated in human hematopoietic stem cells of the line CD45RA(hi) Lin- CD10+, which are biased toward developing into B cells, versus CD45RA(int) CD7- and CD45RA(hi) CD7+.	0.188	0.392	0.070	0.394	0.107	II	IV
BREAST_CANCER_ESTROGEN_SIGNALING	Genes preferentially expressed in breast cancers, especially those involved in estrogen-receptor-dependent signal transduction.	0.055	0.466	0.081	0.231	0.136	II	II
CELL_ADHESION	The attachment of a cell, either to another cell or to the extracellular matrix, via cell adhesion molecules.	0.434	0.046	0.083	0.299	0.137	II	III
PASSERINI_PROLIFERATION	Genes associated with cellular adhesion that are differentially expressed in endothelial cells of pig aortas from regions of disturbed flow (inner aortic arch) versus regions of undisturbed laminar flow (descending thoracic aorta).	0.015	0.468	0.080	0.301	0.135	II	II
LEI_MYB_REGULATED_GENES	Myb-regulated genes	0.389	0.057	0.077	0.160	0.117	II	III
HADDAD_HPCLYMPHO_ENRICHED	Genes enriched in CD45RAhiLin-CD10+ vs CD45RAintCD7- and CD45RAhiCD7hi HPCs	0.174	0.465	0.084	0.356	0.137	II	IV
KUMAR_HOXA_DIFF	Genes that were significantly different between wild-type, preleukemic, and leukemic mice	0.389	0.086	0.077	0.252	0.111	II	III
LINDSTEDT_DEND_DN	Genes down-regulated in maturing DC	0.396	0.206	0.080	0.171	0.114	II	IV
GH_GHRHR_KO_24HRS_UP	Up-regulated at least 2-fold 24 hours following injection of human growth hormone (GH) into mice lacking functional GHRHR (lit/lit), and with no detecTable endogenous GH	0.358	0.414	0.084	0.436	0.116	II	IV
BRG1_ALAB_UP	Up-regulated at 18 and 24 hours following adenovirus-mediated expression of BRG1 in ALAB breast cancer cells with mutant, inactive BRG1	0.152	0.260	0.022	0.258	0.045	II	IV
GH_GHRHR_KO_6HRS_UP	Up-regulated at least 2-fold 6 hours following injection of human growth hormone (GH) into mice lacking functional GHRHR (lit/lit), and with no detecTable endogenous GH	0.408	0.394	0.070	0.550	0.110	II	IV
POD1_KO_DN	Down-regulated in glomeruli isolated from Pod1 knockout mice, versus wild-type controls	0.142	0.319	0.045	0.256	0.072	II	II

ELONGINA_KO_UP	Upregulated in MES cells from elongin-A knockout mice	0.135	0.356	0.055	0.272	0.090	II	II
STRESS_ARSENIC_SPECIFIC_UP	Genes up-regulated 4 hours following arsenic treatment that discriminate arsenic from other stress agents	0.394	0.415	0.076	0.421	0.108	II	IV
EMT_UP	Up-regulated during the TGFbeta-induced epithelial-to-mesenchymal transition (EMT) of Ras-transformed mouse mammary epithelial (EpH4) cells (EMT is representative of late-stage tumor progression and metastasis)	0.419	0.202	0.072	0.388	0.125	II	IV
DSRNA_UP	Upregulated by dsRNA (polyI:C) in IFN-null GRE cells	0.392	0.194	0.079	0.250	0.122	II	IV
IDX_TSA_DN_CLUSTER5	Strongly down-regulated at 2-96 hours during differentiation of 3T3-L1 fibroblasts into adipocytes with IDX (insulin, dexamethasone and isobutylxanthine), vs. fibroblasts treated with IDX + TSA to prevent differentiation (cluster 5)	0.416	0.279	0.073	0.398	0.119	II	IV
BAF57_BT549_UP	Up-regulated following sTable re-expression of BAF57 in Bt549 breast cancer cells that lack functional BAF57	0.387	0.236	0.078	0.240	0.123	II	IV
FSH_OVARY_MCV152_UP	Up-regulated in ovarian epithelial cells (MCV152) 72 hours following FSH treatment, compared to untreated	0.396	0.412	0.072	0.238	0.109	II	IV
HSA00564_GLYCEROPHOSPHOLIPID_METABOLISM	Genes involved in glycerophospholipid metabolism	0.316	0.268	0.041	0.450	0.070	II	IV
HSA04060_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	Genes involved in cytokine-cytokine receptor interaction	0.420	0.193	0.072	0.173	0.124	II	IV
HSA05222_SMALL_CELL_LUNG_CANCER	Genes involved in small cell lung cancer	0.385	0.412	0.074	0.419	0.106	II	IV
LEE_MYC_TGFA_UP	Genes up-regulated in hepatoma tissue of Myc+Tgfa transgenic mice	0.608	0.383	0.315	0.076	0.118	III	IV
LU_IL4BCELL	Genes induced in peripheral B cells by 4 hours of incubation with the cytokine IL-4.	0.627	0.131	0.342	0.077	0.118	III	III
TAKEDA_NUP8_HOXA9_10D_DN	Effect of NUP98-HOXA9 on gene transcription at 10 d after transduction Down	0.666	0.489	0.400	0.080	0.111	III	IV
UVC_TTD_4HR_DN	Down-regulated at 4 hours following treatment of XPB/TTD fibroblasts with 3 J/m^2 UVC	0.175	0.902	0.784	0.069	0.114	III	IV
UVC_TTD_ALL_DN	Down-regulated at any timepoint following treatment of XPB/TTD fibroblasts with 3 J/m^2 UVC	0.363	0.896	0.839	0.078	0.118	III	IV
CMV_HCMV_TIMECOURSE_14HRS_DN	Down-regulated in fibroblasts following infection with human cytomegalovirus (at least 3-fold, with Affymetrix change call, in at least two consecutive timepoints), with maximum change at 14 hours	0.555	0.401	0.206	0.078	0.110	III	IV
BRCA_ER_NEG	Genes whose expression is consistently negatively correlated with estrogen receptor status in breast cancer - higher expression is associated with ER-negative tumors	0.673	0.017	0.498	0.099	0.132	III	III
HSC_LTHSC_ADULT	Up-regulated in mouse long-term functional hematopoietic stem cells from adult bone marrow (LT-HSC Shared + Adult)	0.139	0.634	0.236	0.081	0.120	III	II
HSA04510_FOCAL_ADHESION	Genes involved in focal adhesion	0.007	0.674	0.339	0.100	0.131	III	II
KENNY_WNT_DN	Genes down-regulated by Wnt in HC11 (mammary epithelial cells)	0.443	0.093	0.090	0.138	0.153	V	VII
UVC_HIGH_ALL_DN	Down-regulated at any timepoint following treatment of WS1 human skin fibroblasts with UVC at a high dose (50 J/m^2) (clusters d1-d9)	0.156	0.547	0.141	0.138	0.180	V	NA
IRS1_KO_ADIP_DN	Down-regulated in brown preadipocytes from Irs1-knockout mice, which display severe defects in adipocyte differentiation, versus wild-type controls	0.459	0.094	0.102	0.135	0.174	V	VII
ASTON_DEPRESSION_DN	Genes downregulated in major depressive disorder (p < 0.05, fold change > 1.4, mean average difference > 150 in at least one of the groups, called present in greater than 20% of all samples)	0.380	0.515	0.108	0.640	0.192	VI	NA
HINATA_NFKB_UP	Genes upregulated by NF-kappa B	0.522	0.433	0.148	0.414	0.272	VI	NA

ST_INTEGRIN_SIGNALING_PATHWAY	Integrins are transmembrane receptors that mediate cell growth, survival, and migration by binding to ligands in the extracellular matrix.	0.243	0.514	0.109	0.168	0.187	VI	NA
BRENTANI_CELL_ADHESION	Cancer related genes involved in cell adhesion and metalloproteinases	0.527	0.176	0.146	0.275	0.272	VI	NA
CELL_ADHESION_MOLECULE_ACTIVITY	Obsolete by GO - mediates the adhesion of the cell to other cells or to the extracellular matrix.	0.455	0.071	0.101	0.214	0.172	VI	VII
GUO_HEX_DN	Down-regulated genes in day-6 Hex/ embryoid bodies	0.452	0.413	0.096	0.446	0.160	VI	NA
MOREAUX_TACI_HI_VS_LOW_UP	Genes overexpressed in TACI high patients	0.478	0.262	0.112	0.369	0.199	VI	NA
TAKEDA_NUP8_HOXA9_8D_DN	Effect of NUP98-HOXA9 on gene transcription at 8 d after transduction Down	0.527	0.394	0.147	0.501	0.273	VI	NA
BASSO_GERMINAL_CENTER_CD40_UP	CD40 up-regulated genes	0.528	0.520	0.146	0.362	0.269	VI	NA
CMV_24HRS_UP	Upregulated at 24hrs following infection of primary human foreskin fibroblasts with CMV	0.516	0.385	0.132	0.670	0.237	VI	NA
TCELL_ANERGIC_UP	Genes up-regulated in anergic mouse T helper cells (A.E7), versus non-anergic stimulated controls	0.523	0.534	0.145	0.237	0.266	VI	NA
HSA04670_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	Genes involved in Leukocyte transendothelial migration	0.515	0.559	0.149	0.415	0.273	VI	NA
AGUIRRE_PANCREAS_CHR12	Genes on chromosome 1 with copy-number-driven expression in pancreatic adenocarcinoma.	0.002	0.613	0.202	0.128	0.158	VII	VI
SHEPARD_NEG_REG_OF_CELL_PROLIFERATION	Negative regulators of cell proliferation in zebra fish	0.432	0.646	0.266	0.149	0.199	VII	NA
CELL_PROLIFERATION	The multiplication or reproduction of cells, resulting in the rapid expansion of a cell population.	0.144	0.834	0.616	0.565	0.923	NA	VI
PROLIFERATION_GENES	Proliferation related genes	0.106	0.673	0.317	0.425	0.581	NA	VI
SHEPARD_CELL_PROLIFERATION	Cell proliferation genes determined in zebra fish	0.144	0.834	0.616	0.565	0.923	NA	VI
AGUIRRE_PANCREAS_CHR17	Genes on chromosome 17 with copy-number-driven expression in pancreatic adenocarcinoma.	0.140	0.671	0.302	0.273	0.456	NA	VI
UVC_HIGH_D5_DN	Progressively down-regulated through 18 hours following treatment of WS1 human skin fibroblasts with UVC at a high dose (50 J/m^2) (cluster d5)	0.144	0.665	0.328	0.561	0.587	NA	VI
CALRES_RHESUS_UP	Upregulated in the vastus lateralis muscle of middle-aged rhesus monkeys subjected to caloric restriction since young adulthood vs. age-matched controls	0.138	0.833	0.621	0.360	0.580	NA	VI
CMV_HCMV_TIMECOURSE_ALL_DN	Down-regulated in fibroblasts following infection with human cytomegalovirus (at least 3-fold, with Affymetrix change call, in at least two consectutive timepoints)	0.101	0.667	0.313	0.358	0.572	NA	VI
RADAEVA_IFNA_UP	Genes up-regulated by interferon-alpha in primary hepatocyte	0.814	0.088	0.820	0.465	0.775	NA	VII
FRASOR_ER_DN	Selective estrogen receptor modulators downregulated signature	0.666	0.087	0.489	0.371	0.589	NA	VII
DER_IFNG_UP	Genes up-regulated by interferon-gamma in HT1080 (fibrosarcoma)	0.661	0.136	0.474	0.360	0.589	NA	VII
LINDSTEDT_DEND_UP	Genes up-regulated in DC stimulated for 8 and 48 h	0.712	0.086	0.568	0.542	0.886	NA	VII
ET743_HELA_UP	Upregulated by Et-743 in HeLa cells	0.661	0.023	0.462	0.370	0.606	NA	VII
ADIP_DIFF_CLUSTER1	Progressively downregulated over 24 hours during differentiation of 3T3-L1 fibroblasts into adipocytes (cluster 1)	0.607	0.105	0.274	0.239	0.354	NA	VII
CARIES_PULP_UP	Up-regulated in pulpal tissue from extracted carious teeth (cavities), compared to tissue from extracted healthy teeth	0.660	0.089	0.426	0.281	0.486	NA	VII

APPENDIX B

Q-VALUES OF ENRICHED PATHWAYS FOR A LUNG CANCER STUDY

Q-values of enriched pathways detected by individual studies and MAPE methods in drug response data (column 3-7: q-value threshold 0.05 and significant q-values marked in red) and categories (column 8-9) that correspond to Figure 6 in the manuscript. “Categories comparing MAPE_P, MAPE_G & MAPE_I” correspond to the categories in Figure 2.16A and 2.16B. “Categories comparing Beer, Bhat & MAPE_I” correspond to the categories in Figure 2.16C and 2.16D.

Pathways	Description Genes that are downregulated in AML NPM1 mutant versus AML NPM1 wild type	Beer	Bhat	MAPE_P	MAPE_G	MAPE_I	CA	CB
LE_MYELIN_DN	Genes downregulated in Egr2Lo/Lo mice (who bear mutations in the transcription factor Egr2 and in which peripheral nerve myelination is disrupted) whose expression is significantly altered after sciatic nerve injury.	0.000	0.039	0.001	0.002	0.001	I	I
ICHIBA_GVHD	Genes whose expression is altered greater than twofold in mouse livers experiencing graft-versus-host disease (GVHD) as a result of allogenic bone marrow transplantation.	0.154	0.035	0.003	0.029	0.007	I	III
GNATENKO_PLATELET	Top expressed genes in human platelet cells.	0.085	0.000	0.000	0.015	0.001	I	III
PASSERINI_TRANSCRIPTION	Genes associated with cellular adhesion that are differentially expressed in endothelial cells of pig aortas from regions of disturbed flow (inner aortic arch) versus regions of undisturbed laminar flow (descending thoracic aorta).	0.158	0.010	0.003	0.007	0.007	I	III
SANA_TNFA_ENDOTHELIAL_DN	Genes down-regulated by TNFA in colon,derm,iliac,aortic,lung endothelial cells	0.021	0.013	0.000	0.010	0.000	I	I
PEART_HISTONE_UP	Cell-proliferation-related genes upregulated by SAHA and depsipeptide (histone deacetylase inhibitors)	0.170	0.086	0.007	0.017	0.014	I	IV
FLECHNER_KIDNEY_TRANSPLANT_REJECTION_UP	Genes upreglated in acute rejection transplanted kidney biopsies relative to well functioning transplanted kidney biopsies from sTable, immunosuppressed, recipients (median FDR < 0.14% per comparison)	0.000	0.002	0.000	0.001	0.000	I	I
CROONQUIST_RAS_STROMA_DN	Genes downregulated in multiple myeloma cells with N-ras-activating mutations versus those co-cultured with bone marrow stromal cells.	0.138	0.002	0.002	0.006	0.004	I	III
JECHLINGER_EMT_DN	Genes downregulated for epithelial plasticity in tumor progression	0.085	0.005	0.000	0.019	0.001	I	III
CORDERO_KRAS_KD_VS_CONTR	Genes upregulated in kras knockdown vs control in a human cell line	0.193	0.044	0.010	0.047	0.020	I	III

OL_UP								
SHIPP_FL_VS_DLBCL_DN	Genes upregulated in diffuse B-cell lymphomas (DLBCL) and downregulated in follicular lymphoma (FL) (fold change of at least 3)	0.147	0.001	0.002	0.038	0.005	I	III
CHIARETTI_T_ALL	Genes overexpressed in leukemia cells.	0.125	0.000	0.001	0.014	0.003	I	III
PROSTAGLANDIN_SYNTHESIS_REGULATION		0.105	0.034	0.001	0.017	0.001	I	III
STRIATED_MUSCLE_CONTRACTION		0.287	0.127	0.034	0.019	0.023	I	IV
CROONQUIST_IL6_STROMA_UP	Genes upregulated in multiple myeloma cells exposed to the pro-proliferative cytokine IL-6 versus those co-cultured with bone marrow stromal cells.	0.048	0.044	0.001	0.002	0.001	I	I
RUIZ_TENASCIN_TARGETS	Tenascin-C target genes	0.043	0.020	0.000	0.003	0.000	I	I
YAO_P4_KO_VS_WT_UP	Genes that have at least a 15 fold increase in expression in the KO compared to WT at 6 hours after P4 injection in ovariectomized mice	0.003	0.000	0.000	0.007	0.000	I	I
BOQUEST_CD31PLUS_VS_CD31MINUS_DN	Genes overexpressed 3-fold or more in freshly isolated CD31- versus freshly isolated CD31+ cells	0.000	0.004	0.000	0.013	0.000	I	I
LEI_MYB_REGULATED_GENES	Myb-regulated genes	0.195	0.000	0.009	0.001	0.001	I	III
CHIARETTI_T_ALL_DIFF	Genes expressed in T-cell acute lymphocytic leukemia	0.085	0.000	0.000	0.007	0.001	I	III
BOQUEST_CD31PLUS_VS_CD31MINUS_UP	Genes overexpressed 3-fold or more in freshly isolated CD31+ versus freshly isolated CD31- cells	0.000	0.000	0.000	0.000	0.000	I	I
RORIE_ES_PNET_UP	The 30 genes showing the greatest increase in expression in NBa Ews/Fli-1 infectants	0.203	0.034	0.011	0.017	0.020	I	III
HOHENKIRK_MONOCYTE_DEND_UP	Up-regulated mRNAs in monocyte-derived DCs	0.022	0.000	0.000	0.000	0.000	I	I
HOHENKIRK_MONOCYTE_DEND_DN	Down-regulated mRNAs in monocyte-derived DCs	0.007	0.037	0.001	0.017	0.001	I	I
GERY_CEBP_TARGETS	Complete list of differentially regulated C/EBP-target genes, sorted by P-value	0.040	0.015	0.000	0.007	0.000	I	I
VERHAAK_AML_NPM1_MUT_VS_WT_UP	Genes that are upregulated in AML NPM1 mutant versus AML NPM1 wild type	0.026	0.003	0.000	0.032	0.000	I	I
IRITANI_ADPROX_LYMPH	LYMPHATIC EC	0.084	0.000	0.000	0.034	0.001	I	III
LI_FETAL_VS_WT_KIDNEY_UP	These are genes identified by simple statistical criteria as differing in their mRNA expresssion between WTs and fetal kidneys LOW	0.087	0.000	0.000	0.013	0.001	I	III
TAKEDA_NUP8_HOXA9_10D_DN	Effect of NUP98-HOXA9 on gene transcription at 10 d after transduction Down	0.014	0.014	0.000	0.003	0.000	I	I
NAKAJIMA_MCSMBP_MAST	Top 50 most-increased mast cell specific transcripts	0.253	0.001	0.025	0.023	0.028	I	III
TAVOR_CEBP_UP	C/EBP up-regulated genes in KCL22 cells	0.045	0.011	0.000	0.014	0.000	I	I
IGLESIAS_E2FMINUS_UP	Genes that increase in the absence of E2F1 and E2F2	0.064	0.000	0.000	0.001	0.000	I	III
GNATENKO_PLATELET_UP	Top 50 human platelet-expressed genes	0.085	0.000	0.000	0.015	0.001	I	III
RUTELLA_HEMATOGFSNDCS_DIFF	The 672 significantly changing genes	0.111	0.000	0.001	0.002	0.001	I	III
KUMAR_HOXA_DIFF	Genes that were significantly different between wild-type, preleukemic, and leukemic mice	0.203	0.046	0.011	0.017	0.020	I	III

P21_ANY_DN	Down-regulated at any timepoint (4-24 hrs) follwing ectopic expression of p21 (CDKN1A) in OvCa cells	0.048	0.015	0.000	0.001	0.000	I	I
HYPOXIA_REVIEW	Genes known to be induced by hypoxia	0.165	0.033	0.005	0.047	0.010	I	III
BLEO_HUMAN_LYMPH_HIGH_24HRS_UP	Up-regulated at 24 hours following treatment of human lymphocytes (TK6) with a high dose of bleomycin	0.109	0.045	0.001	0.034	0.003	I	III
ATRIA_UP	Upregulated in the atria of healthy hearts, compared to venticles	0.004	0.004	0.000	0.002	0.000	I	I
PLATELET_EXPRESSED	Fifty genes most strongly expressed in human platlets from three healthy donors	0.048	0.006	0.000	0.027	0.000	I	I
CMV_HCMV_TIMECOURSE_20HRS_DN	Down-regulated in fibroblasts following infection with human cytomegalovirus (at least 3-fold, with Affymetrix change call, in at least two consectutive timepoints), with maximum change at 20 hours	0.105	0.082	0.006	0.047	0.014	I	IV
LVAD_HEARTFAILURE_UP	Upregulated in the left ventricle myocardium of patients with heart failure following implantation of a left ventricular assist device	0.044	0.000	0.000	0.007	0.000	I	I
AGEING_BRAIN_UP	Age-upregulated in the human frontal cortex	0.042	0.003	0.000	0.004	0.000	I	I
IDX_TSA_UP_CLUSTER3	Strongly up-regulated at 16-24 hours during differentiation of 3T3-L1 fibroblasts into adipocytes with IDX (insulin, dexamethasone and isobutylxanthine), vs. fibroblasts treated with IDX + TSA to prevent differentiation (cluster 3)	0.270	0.006	0.028	0.023	0.028	I	III
IDX_TSA_UP_CLUSTER2	Strongly up-regulated at 8 hours during differentiation of 3T3-L1 fibroblasts into adipocytes with IDX (insulin, dexamethasone and isobutylxanthine), vs. fibroblasts treated with IDX + TSA to prevent differentiation (cluster 2)	0.167	0.116	0.012	0.003	0.002	I	IV
AGED_MOUSE_NEOCORTEX_UP	Upregulated in the neocortex of aged adult mice (30-month) vs. young adult (5-month)	0.020	0.028	0.000	0.017	0.001	I	I
ADIP_DIFF_CLUSTER5	Strongly upregulated at 24 hours during differentiation of 3T3-L1 fibroblasts into adipocytes (cluster 5)	0.193	0.002	0.010	0.038	0.020	I	III
ADIP_DIFF_CLUSTER2	Strongly upregulated at 2 hours during differentiation of 3T3-L1 fibroblasts into adipocytes (cluster 2)	0.172	0.041	0.007	0.034	0.014	I	III
AGED_MOUSE_CEREBELLUM_UP	Upregulated in the cerebellum of aged adult mice (30-month) vs. young adult (5-month)	0.105	0.034	0.001	0.047	0.001	I	III
AGEING_KIDNEY_UP	Up-regulation is associated with increasing age in normal human kidney tissue from 74 patients	0.071	0.046	0.001	0.017	0.003	I	III
SERUM_FIBROBLAST_CELLCYCLE	Cell-cycle dependent genes regulated following exposure to serum in a variety of human fibroblast cell lines	0.163	0.000	0.004	0.004	0.004	I	III
POD1_KO_UP	Up-regulated in glomeruli isolated from Pod1 knockout mice, versus wild-type controls	0.117	0.020	0.001	0.003	0.002	I	III
CALRES_RHESUS_UP	Upregulated in the vastus lateralis muscle of middle-aged rhesus monkeys subjected to caloric restriction since young adulthood vs. age-matched controls	0.045	0.004	0.000	0.047	0.000	I	I
EMT_DN	Down-regulated during the TGFbeta-induced epithelial-to-mesenchymal transition (EMT) of Ras-transformed mouse mammary epithelial (EpH4) cells (EMT is representative of late-stage tumor progression and metastasis)	0.036	0.001	0.000	0.003	0.000	I	I
HEARTFAILURE_VENTRICLE_DN	Downregulated in the ventricles of failing hearts (DCM and ICM) compared to healthy controls	0.162	0.005	0.005	0.014	0.010	I	III
CARIES_PULP_UP	Up-regulated in pulpal tissue from extracted carious teeth (cavities), compared to tissue from extracted healthy teeth	0.004	0.000	0.000	0.000	0.000	I	I
CMV_HCMV_TIMECOURSE_ALL_DN	Down-regulated in fibroblasts following infection with human cytomegalovirus (at least 3-fold, with Affymetrix change call, in at least two consectutive timepoints)	0.021	0.004	0.000	0.004	0.000	I	I
HSA04510_FOCAL_ADHESION	Genes involved in focal adhesion	0.200	0.122	0.013	0.046	0.029	I	IV
HSA04514_CELL_ADHESION_M	Genes involved in cell adhesion molecules (CAMs)	0.212	0.003	0.012	0.003	0.002	I	III

OLECULES								
HSA04670_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	Genes involved in Leukocyte transendothelial migration	0.073	0.014	0.000	0.024	0.001	I	III
SHEPARD_BMYB_MORPHOLINO_DN	Genes upregulated in control vs bmyb morpholino knockdown in zebra fish	0.194	0.153	0.022	0.251	0.045	II	IV
JECHLINGER_EMT_UP	Genes upregulated for epithelial plasticity in tumor progression	0.173	0.028	0.007	0.054	0.015	II	III
PASSERINI_ADHESION	Genes associated with cellular adhesion that are differentially expressed in endothelial cells of pig aortas from regions of disturbed flow (inner aortic arch) versus regions of undisturbed laminar flow (descending thoracic aorta).	0.237	0.007	0.019	0.055	0.039	II	III
TARTE_MATURE_PC	Genes overexpressed in polyclonal plasmablastic cells (PPCs) as compared to mature plasma cells isolated from tonsils (TPCs) and mature plasma cells isolated from bone marrow (BMPCs).	0.161	0.035	0.004	0.169	0.009	II	III
ZELLER_MYC_UP	Genes up-regulated by MYC in >3 papers.	0.235	0.026	0.019	0.181	0.038	II	III
NELSON_ANDROGEN_UP	Genes upregulated by androgen in neoplastic prostate epithelium	0.168	0.092	0.008	0.170	0.016	II	IV
GO_ROS	Reactive oxidative species related genes curated from GO	0.129	0.086	0.007	0.203	0.014	II	IV
SHIPP_FL_VS_DLBCL_UP	Genes upregulated in follicular lymphoma (FL) and downregulated in diffuse B-cell lymphomas (DLBCL) (fold change of at least 3)	0.193	0.059	0.010	0.161	0.021	II	IV
HADDAD_HSC_CD7_UP	Genes upregulated in human hematopoietic stem cells of the line CD45RA(hi) CD7+, which are biased toward developing into T lymphocytes or natural killer cells, versus CD45RA(int) CD7-.	0.149	0.002	0.003	0.054	0.006	II	III
SANSOM_APC_LOSS4_UP	The top 174 genes upregulated following Apc loss at day 4	0.193	0.014	0.010	0.081	0.021	II	III
NAKAJIMA_MCSMBP_EOS	Top 30 increased eosinophil specific transcripts	0.169	0.145	0.019	0.310	0.040	II	IV
STEFFEN_AML_PML_PLZF_TRGT	Target genes shared by AML1-ETO, PML-RAR, and PLZF-RAR	0.161	0.046	0.004	0.133	0.009	II	III
ZHANG_EFT_EWSFLI1_UP	Genes (n = 109) significantly upregulated in RD-EF and also highly expressed in EFT	0.194	0.012	0.010	0.072	0.020	II	III
HADDAD_CD45CD7_PLUS_VS_MINUS_UP	Genes enriched in CD45RAhiCD7hi vs CD45RAintCD7- HPCs	0.149	0.002	0.003	0.054	0.006	II	III
NAKAJIMA_MCS_UP	Most increased transcripts in activated human and mouse MCs	0.168	0.003	0.007	0.122	0.014	II	III
ALCALAY_AML_NPMC_UP	Increased expression in NPMc+ leukemias	0.167	0.046	0.007	0.073	0.014	II	III
KNUDSEN_PMNS_UP	Genes up-regulated in PMNs upon migration to skin lesions	0.168	0.081	0.007	0.189	0.014	II	IV
GAY_YY1_DN	List of YY1 target genes identified in MEFs expressing ~25% of YY1 Down	0.250	0.021	0.024	0.169	0.048	II	III
ALCALAY_AML_NPMC_DN	Decreased expression in NPMc+ leukemias	0.185	0.004	0.009	0.215	0.018	II	III
TAKEDA_NUP8_HOXA9_8D_DN	Effect of NUP98-HOXA9 on gene transcription at 8 d after transduction Down	0.225	0.041	0.016	0.125	0.033	II	III
RADMACHER_AMLNORMALKARYTYPE_SIG	Bullinger Validation Signature (157 Affymetrix probe sets)	0.149	0.085	0.007	0.168	0.014	II	IV
VERHAAK_AML_NPM1_MUT_VS_WT_DN	Description Genes that are downregulated in AML NPM1 mutant versus AML NPM1 wild type	0.194	0.020	0.010	0.066	0.020	II	III
YAGI_AML_PROG_FAB	FAB type-specific probe sets	0.233	0.037	0.018	0.165	0.037	II	III

RAY_P210_DIFF	Functional classification of p210BCR-ABL differentially regulated genes identified by cDNA macroarray	0.228	0.046	0.017	0.161	0.034	II	III
ZHAN_MULTIPLE_MYELOMA_VS_NORMAL_DN	The 50 most significantly down-regulated genes in MM in comparison with normal bone marrow PCs	0.055	0.013	0.000	0.071	0.000	II	III
FALT_BCLL_UP	Genes up-regulated in VH3-21+ B-CLL	0.105	0.036	0.001	0.073	0.001	II	III
LINDSTEDT_DEND_DN	Genes down-regulated in maturing DC	0.144	0.003	0.002	0.185	0.004	II	III
HSC_LTHSC_SHARED	Up-regulated in mouse long-term functional hematopoietic stem cells from both adult bone marrow and fetal liver (Cluster i, LT-HSC Shared)	0.116	0.110	0.010	0.109	0.021	II	IV
POD1_KO_DN	Down-regulated in glomeruli isolated from Pod1 knockout mice, versus wild-type controls	0.149	0.000	0.002	0.157	0.005	II	III
DIAB_NEPH_UP	Upregulated in the glomeruli of cadaver kidneys from patients with diabetic nephropathy, compared to normal controls	0.050	0.007	0.000	0.087	0.000	II	III
HTERT_UP	Upregulated in hTERT-immortalized fibroblasts vs. non-immortalized controls	0.234	0.073	0.019	0.135	0.039	II	IV
PARP_KO_UP	Upregulated in MEF cells from PARP knockout mice	0.147	0.040	0.002	0.072	0.005	II	III
CMV_HCMV_TIMECOURSE_48HRS_DN	Down-regulated in fibroblasts following infection with human cytomegalovirus (at least 3-fold, with Affymetrix change call, in at least two consectutive timepoints), with maximum change at 48 hours	0.192	0.062	0.009	0.161	0.019	II	IV
UVB_NHEK1_UP	Upregulated by UV-B light in normal human epidermal keratinocytes	0.172	0.027	0.006	0.081	0.014	II	III
CARIES_PULP_HIGH_UP	Highly up-regulated (>4-fold) in pulpal tissue from extracted carious teeth (cavities), compared to tissue from extracted healthy teeth	0.171	0.041	0.006	0.068	0.014	II	III
EMT_UP	Up-regulated during the TGFbeta-induced epithelial-to-mesenchymal transition (EMT) of Ras-transformed mouse mammary epithelial (EpH4) cells (EMT is representative of late-stage tumor progression and metastasis)	0.192	0.046	0.009	0.098	0.019	II	III
HSC_LTHSC_FETAL	Up-regulated in mouse long-term functional hematopoietic stem cells from fetal liver (LT-HSC Shared)	0.116	0.110	0.010	0.109	0.021	II	IV
AGEING_KIDNEY_SPECIFIC_UP	Up-regulation is associated with increasing age in normal human kidney tissue from 74 patients, and expression is higher in kidney than in whole blood	0.234	0.149	0.020	0.072	0.042	II	IV
ROS_MOUSE_AORTA_DN	Down-regulated in mouse aorta by chronic treatment with PPARgamma agonist rosiglitazone	0.022	0.035	0.001	0.053	0.001	II	I
ADIP_DIFF_CLUSTER1	Progressively downregulated over 24 hours during differentiation of 3T3-L1 fibroblasts into adipocytes (cluster 1)	0.233	0.036	0.020	0.206	0.040	II	III
E2F3_ONCOGENIC_SIGNATURE	Genes selected in supervised analyses to discriminate cells expressing E2F3 oncogene from control cells expressing GFP.	0.182	0.116	0.012	0.262	0.024	II	IV
HSA00350_TYROSINE_METABOLISM	Genes involved in tyrosine metabolism	0.106	0.087	0.007	0.089	0.014	II	IV
HSA04530_TIGHT_JUNCTION	Genes involved in tight junction	0.233	0.089	0.019	0.076	0.040	II	IV
BASSO_REGULATORY_HUBS	Genes which comprise the top 1% of highly interconnected genes (major hubs) that account for most of the interactions in the reconstructed regulatory networks from expression profiles in human B cells.	0.329	0.000	0.063	0.008	0.009	III	III
LI_FETAL_VS_WT_KIDNEY_DN	These are genes identified by simple statistical criteria as differing in their mRNA expresssion between WTs and fetal kidneys HIGH	0.317	0.007	0.057	0.026	0.033	III	III
HALMOS_CEBP_DN	The list of most highly downregulated genes after conditional expression of C/EBPalpha	0.298	0.216	0.058	0.034	0.044	III	IV
HDACI_COLON_BUT_UP	Upregulated by butyrate at any timepoint up to 48 hrs in SW260 colon carcinoma cells	0.409	0.020	0.139	0.027	0.036	III	III
CMV_24HRS_DN	Downregulated at 24hrs following infection of primary human foreskin fibroblasts with CMV	0.349	0.004	0.079	0.017	0.019	III	III
CMV_ALL_DN	Downregulated at any timepoint following infection of	0.365	0.015	0.094	0.017	0.021	III	III

	primary human foreskin fibroblasts with CMV							
HDACI_COLON_BUT48HRS_UP	Upregulated by butyrate at 48 hrs in SW260 colon carcinoma cells	0.514	0.040	0.259	0.030	0.039	III	III
HDACI_COLON_TSABUT_UP	Upregulated by both butyrate and TSA at any timepoint up to 48 hrs in SW260 colon carcinoma cells	0.339	0.039	0.071	0.016	0.019	III	III
IDX_TSA_DN_CLUSTER3	Strongly down-regulated at 8-96 hours during differentiation of 3T3-L1 fibroblasts into adipocytes with IDX (insulin, dexamethasone and isobutylxanthine), vs. fibroblasts treated with IDX + TSA to prevent differentiation (cluster 3)	0.223	0.240	0.070	0.034	0.044	III	IV
HSA01430_CELL_COMMUNICATION	Genes involved in cell communication	0.371	0.015	0.099	0.016	0.019	III	III
HSA04610_COMPLEMENT_AND_COAGULATION_CASCADES	Genes involved in complement and coagulation cascades	0.039	0.284	0.098	0.013	0.015	III	II
OXSTRESS_RPETHREE_DN	Downregulated by all three of H2O2, HNE and t-BH in retinal pigment epithelium cells (Table 2)	0.101	0.202	0.047	0.047	0.064	V	NA
HTERT_DN	Downregulated in hTERT-immortalized fibroblasts vs. non-immortalized controls	0.157	0.160	0.026	0.048	0.053	V	NA
HDACI_COLON_BUT24HRS_UP	Upregulated by butyrate at 24 hrs in SW260 colon carcinoma cells	0.308	0.159	0.045	0.041	0.055	V	NA
ROSS_CBF	Genes that distinguish pediatric acute myeloid leukemia (AML) core-binding factor (CBF) subtypes.	0.267	0.111	0.028	0.164	0.055	VI	NA
PASSERINI_SIGNAL	Genes associated with cellular adhesion that are differentially expressed in endothelial cells of pig aortas from regions of disturbed flow (inner aortic arch) versus regions of undisturbed laminar flow (descending thoracic aorta).	0.310	0.203	0.047	0.347	0.093	VI	NA
HOGERKORP_ANTI_CD44_UP	Genes differentially expressed in human B cells cultured in vitro in the presence or absence of CD44 ligation, together with anti-immunoglobulin and anti-CD40 antibodies	0.161	0.207	0.050	0.490	0.095	VI	NA
WIELAND_HEPATITIS_B_INDUCED	Genes induced in the liver during hepatitis B viral clearance in chimpanzees.	0.310	0.032	0.046	0.243	0.091	VI	VII
MANALO_HYPOXIA_DN	Genes downregulated in human pulmonary endothelial cells under hypoxic conditions or after exposure to AdCA5, an adenovirus carrying constitutively active hypoxia-inducible factor 1 (HIF-1alpha).	0.253	0.175	0.032	0.239	0.062	VI	NA
BROCKE_IL6	Genes whose expression was modulated at least 1.5-fold in multiple myeloma INA-6 cells on addition of interleukin-6.	0.308	0.004	0.045	0.144	0.087	VI	VII
ROSS_CBF_LEUKEMIA	Genes upregulated in AML samples with the CBF subtype	0.302	0.090	0.042	0.206	0.083	VI	NA
LEE_MYC_E2F1_UP	Genes up-regulated in hepatoma tissue of Myc+E2f1 transgenic mice	0.271	0.054	0.031	0.127	0.061	VI	NA
CELL_ADHESION	The attachment of a cell, either to another cell or to the extracellular matrix, via cell adhesion molecules.	0.257	0.088	0.026	0.251	0.053	VI	NA
CALCIUM_REGULATION_IN_CARDIAC_CELLS		0.306	0.187	0.045	0.489	0.087	VI	NA
SMOOTH_MUSCLE_CONTRACTION		0.235	0.177	0.032	0.215	0.063	VI	NA
TYROSINE_METABOLISM		0.309	0.207	0.049	0.091	0.094	VI	NA
FALT_BCLL_IG_MUTATED_VS_WT_UP	Genes upregulated in Ig-mutated non-VH3-21 B-CLL	0.282	0.036	0.032	0.065	0.064	VI	VII
ZHAN_MMPC_EARLYVS	Early differentiation genes top 50 differentially expressed genes in comparison of CD19-enriched tonsil BCs and CD138-enriched tonsil PCs	0.312	0.197	0.047	0.185	0.093	VI	NA
TAKEDA_NUP8_HOXA9_6H_DN	Effect of NUP98-HOXA9 on gene transcription at 6 h after transfection Down	0.257	0.025	0.026	0.150	0.052	VI	VII
ZHAN_MM_CD138_MF_VS_REST	50 top ranked SAM-defined over-expressed genes in each subgroup_MF	0.196	0.189	0.042	0.079	0.082	VI	NA

YANG_OSTECLASTS_SIG	Relative gene expression for osteoclast-associated genes, chemokines, and chemokine receptors	0.225	0.194	0.044	0.171	0.085	VI	NA
ZHAN_MMPC_SIMAL	LDGs showing similar expression patterns in tonsil PCs and all or subsets of MM	0.160	0.204	0.047	0.160	0.093	VI	NA
KANG_TERT_DN	Expressed gene profile of ATSCs and ATSC-TERT cells and partial list of genes that were downregulated in ATSC-TERT cells	0.105	0.164	0.028	0.135	0.055	VI	NA
IRITANI_ADPROX_DN	BEC-specific suppressed by AdProx-1	0.275	0.060	0.032	0.135	0.063	VI	NA
MENSE_HYPOXIA_TRANSPORTER_GENES	List of Hypoxia-induced/suppressed genes encoding transporters in Astrocytes	0.173	0.163	0.027	0.116	0.055	VI	NA
BASSO_GERMINAL_CENTER_CD40_UP	CD40 up-regulated genes	0.104	0.201	0.046	0.186	0.091	VI	NA
KRETZSCHMAR_IL6_DIFF	Shown are those probe sets that report at least a 15-fold expression change in response to IL-6 addition to INA-6 cells	0.308	0.004	0.045	0.144	0.087	VI	VII
CMV_UV-CMV_COMMON_HCMV_6HRS_DN	Down-regulated in fibroblasts at 6 hours following infection with either human cytomegalovirus (CMV) or UV-inactivated CMV	0.272	0.046	0.031	0.184	0.061	VI	VII
DOX_RESIST_GASTRIC_UP	Upregulated in gastric cancer cell lines reistant to doxorubicin, compared to parent chemosensitive lines	0.279	0.002	0.032	0.163	0.064	VI	VII
HSC_LTHSC_ADULT	Up-regulated in mouse long-term functional hematopoietic stem cells from adult bone marrow (LT-HSC Shared + Adult)	0.172	0.185	0.035	0.172	0.070	VI	NA
ADIPOGENESIS_HMSC_CLASS8_DN	Down-regulated 1-14 days following the differentiation of human bone marrow mesenchymal stem cells (hMSC) into adipocytes, versus untreated hMSC cells (Class VIII)	0.272	0.207	0.050	0.053	0.079	VI	NA
CARIES_PULP_DN	Down-regulated in pulpal tissue from extracted carious teeth (cavities), compared to tissue from extracted healthy teeth	0.148	0.178	0.032	0.377	0.064	VI	NA
ADIP_VS_FIBRO_UP	Upregulated following 7-day differentiation of murine 3T3-L1 fibroblasts into adipocytes	0.233	0.189	0.042	0.282	0.082	VI	NA
HSA00251_GLUTAMATE_METABOLISM	Genes involved in glutamate metabolism	0.232	0.184	0.034	0.054	0.069	VI	NA
HSA04340_HEDGEHOG_SIGNALING_PATHWAY	Genes involved in Hedgehog signaling pathway	0.178	0.192	0.043	0.171	0.084	VI	NA
HOFFMANN_BIVSBII_LGBII	Genes with at least five fold change in expression between large and small Pre-BII cells	0.338	0.000	0.071	0.042	0.056	VII	VII
ZHAN_MMPC_SIM	LDGs showing similar expression patterns in bone marrow PC and subsets of MM	0.443	0.012	0.185	0.041	0.055	VII	VII
GREENBAUM_E2A_UP	Table includes transcripts up-regulated 3-fold or greater in the E2A-deficient cell lines	0.317	0.001	0.057	0.046	0.061	VII	VII
HDACI_COLON_BUT16HRS_UP	Upregulated by butyrate at 16 hrs in SW260 colon carcinoma cells	0.435	0.034	0.179	0.045	0.061	VII	VII
HDACI_COLON_BUT2HRS_UP	Upregulated by butyrate at 2 hrs in SW260 colon carcinoma cells	0.329	0.046	0.064	0.039	0.053	VII	VII
NI2_MOUSE_DN	Downregulated by nickel(II) in sensitive A/J mouse lung tissue	0.194	0.375	0.191	0.045	0.060	VII	NA
H2O2_CSBRESCUED_C1_UP	Upregulated by H2O2 in CSB-rescued fibroblasts (Table 1, cluster 1)	0.364	0.035	0.092	0.047	0.065	VII	VII
KLEIN_PEL_UP	Genes downregulated in AIDS-related primary effusion lymphoma (PEL) cells compared to normal B cells and other tumor subtypes.	0.047	0.304	0.114	0.245	0.231	NA	VI
TAKEDA_NUP8_HOXA9_16D_DN	Effect of NUP98-HOXA9 on gene transcription at 16 d after transduction Down	0.037	0.550	0.452	0.195	0.318	NA	VI
HSA04210_APOPTOSIS	Genes involved in apoptosis	0.020	0.515	0.391	0.071	0.102	NA	VI
TARTE_PC	Genes overexpressed in polyclonal plasmablastic cells (PPCs), mature plasma cells isolated from tonsils (TPCs), and mature	0.519	0.014	0.267	0.267	0.429	NA	VII

	plasma cells isolated from bone marrow (BMPCs), as compared to B cells purified from peripheral blood (PBBs) and tonsils (TBCs).							
GLUCONEOGENESIS		0.536	0.039	0.291	0.156	0.232	NA	VII
DAVIES_MGUS_MM	Genes differentially expressed in monoclonal gammopathy of uncertain significance (MGUS, a precursor state for multiple myeloma) versus multiple myeloma (MM) plasma cells. Fold Change uses MGUS as the baseline.	0.453	0.046	0.194	0.398	0.373	NA	VII
PENG_LEUCINE_DN	Genes downregulated in response to leucine starvation	0.714	0.003	0.614	0.542	0.768	NA	VII
CROONQUIST_IL6_STARVE_UP	Genes upregulated in multiple myeloma cells exposed to the pro-proliferative cytokine IL-6 versus those that were IL-6-starved.	0.329	0.001	0.064	0.066	0.095	NA	VII
ZUCCHI_EPITHELIAL_UP	The 50 most upregulated genes in primary invasive breast dutcal carcinoma or metastatic breast carcinoma isolated from lymph nodes, as compared to normal mammary epithelium.	0.615	0.014	0.406	0.507	0.698	NA	VII
GLYCOLYSIS		0.536	0.039	0.291	0.156	0.232	NA	VII
HCC_SURVIVAL_GOOD_VS_POOR_DN	Genes highly expressed in hepatocellular carcinoma with poor survival.	0.695	0.018	0.570	0.292	0.470	NA	VII
TARTE_PLASMA_BLASTIC	Genes overexpressed in mature plasma cells isolated from tonsils (TPCs) and mature plasma cells isolated from bone marrow (BMPCs) as compared to polyclonal plasmablastic cells (PPCs).	0.675	0.007	0.528	0.346	0.563	NA	VII
LE_MYELIN_UP	Genes upregulated in Egr2Lo/Lo mice (who bear mutations in the transcription factor Egr2 and in which peripheral nerve myelination is disrupted) whose expression is significantly altered after sciatic nerve injury.	0.497	0.046	0.241	0.108	0.164	NA	VII
SCHUMACHER_MYC_UP	Genes up-regulated by MYC in P493-6 (B-cell)	0.434	0.014	0.177	0.053	0.073	NA	VII
CHANG_SERUM_RESPONSE_DN	CSR Stanford signature for quiscent genes	0.372	0.016	0.104	0.127	0.191	NA	VII
PENG_RAPAMYCIN_DN	Genes downregulated in response to rapamycin starvation	0.694	0.037	0.564	0.610	0.786	NA	VII
PENG_RAPAMYCIN_UP	Genes upregulated in response to rapamycin starvation	0.448	0.026	0.191	0.136	0.203	NA	VII
BHATTACHARYA_ESC_UP	Genes upregulated in undifferentiated human embryonic stem cells.	0.691	0.005	0.557	0.522	0.761	NA	VII
LEE_TCELLS10_UP	Transcripts showing more than 2 fold higher expression in CB4 than in AB4	0.370	0.026	0.100	0.121	0.184	NA	VII
NADLER_OBESITY_UP	Genes with increased expression with obesity	0.377	0.020	0.110	0.126	0.188	NA	VII
LEE_TCELLS8_UP	Transcripts enriched in na???ve CD4 T cells (CB4, and AB4) more than 3-fold, with average signal value differences of at least 100 between thymocytes (ITTP, DP, SP4) and naive-phenotype CD4 T (CB4, and AB4) cells	0.370	0.026	0.100	0.121	0.184	NA	VII
FERRANDO_MLL_T_ALL_UP	Top 100 nearest neighbor genes positively associated with MLL T-ALL cases	0.351	0.019	0.082	0.122	0.167	NA	VII
MATSUDA_VALPHAINKT_DIFF	Differential gene expression between developmental stages of Va14i NKT cells	0.542	0.026	0.296	0.256	0.411	NA	VII
FERRANDO_MLL_T_ALL_DN	Top 100 nearest neighbor genes negatively associated with MLL T-ALL cases	0.715	0.022	0.613	0.530	0.763	NA	VII
IRITANI_ADPROX_VASC	BLOOD VASCULAR EC	0.378	0.046	0.110	0.198	0.220	NA	VII
BASSO_HCL_DIFF	Identification of HCL-specific genes, The analysis identified 89 genes that are differentially expressed in HCL versus all the other samples	0.464	0.001	0.204	0.171	0.271	NA	VII
HOFFMANN_BIVSBII_BI_TABLE2	Genes with at least five fold change in expression between Pre-BI and Large Pre-BII cells	0.553	0.020	0.307	0.358	0.560	NA	VII

LINDSTEDT_DEND_8H_VS_48H_DN	Genes up-regulated in DC stimulated for 48 h as compared to DC stimulated for 8 h	0.581	0.019	0.345	0.346	0.557	NA	VII
LEE_TCELLS3_UP	Transcripts enriched in both ITTP and DP more than 3-fold, with average signal value differences of at least 100 between less mature (ITTP, DP) and more mature (SP4, CB4, and AB4) cells	0.436	0.034	0.177	0.485	0.335	NA	VII
YAGI_AML_PROG_ASSOC	Prognosis-associated probe sets	0.631	0.038	0.434	0.434	0.683	NA	VII
LEE_TCELLS1_UP	Transcripts enriched in more mature cells (SP4, CB4, and AB4) more than 3-fold, with average signal value differences of at least 100 between less mature (ITTP, DP) and more mature (SP4, CB4, and AB4) cells	0.370	0.026	0.100	0.121	0.184	NA	VII
AD12_ANY_DN	Down-regulated 2-fold in HeLa cells by Adenovirus type 12 (Ad12) at any timepoint to 48 hrs hours post-infection	0.494	0.019	0.236	0.328	0.452	NA	VII
BRCA1_OVEREXP_PROSTATE_UP	Up-regulated with sTable, ectopic overexpression of BRCA1 in DU-145 human prostate cancer cell lines, compared to neo-only controls	0.603	0.027	0.381	0.109	0.164	NA	VII
BREASTCA_THREE_CLASSES	Gene set that can be used to differentiate BRCA1-linked, BRCA2-linked, and sporadic primary breast cancers	0.438	0.041	0.176	0.210	0.330	NA	VII
CANCER_NEOPLASTIC_META_UP	Sixty-seven genes commonly upregulated in cancer relative to normal tissue, from a meta-analysis of the OncoMine gene expression database	0.474	0.000	0.214	0.172	0.269	NA	VII
MAMMARY_DEV_UP	Up-regulated in the intact developing mouse mammary gland; higher expression in 5/6 week pubertal glands than in 3 week, mid-pregnant, lactating, involuting or resuckled glands	0.333	0.047	0.067	0.066	0.093	NA	VII
NI2_MOUSE_UP	Upregulated by nickel(II) in sensitive A/J mouse lung tissue	0.568	0.008	0.324	0.300	0.488	NA	VII
ADIP_VS_PREADIP_DN	Downregulated in mature murine adipocytes (7 day differentiation) vs. preadipocytes (6 hr differentiation)	0.323	0.037	0.060	0.065	0.092	NA	VII
TSA_HEPATOMA_UP	Up-regulated in more than one of several human hepatoma cell lines by 24-hour treatment with trichostatin A	0.363	0.045	0.094	0.185	0.188	NA	VII
CANCER_UNDIFFERENTIATED_META_UP	Sixty-nine genes commonly upregulated in undifferentiated cancer relative to well-differentiated cancer, from a meta-analysis of the OncoMine gene expression database	0.501	0.006	0.246	0.203	0.330	NA	VII
CMV_IE86_UP	Upregulated by expression of cytomegalovirus IE86 protein in primary human fibroblasts	0.637	0.012	0.444	0.204	0.334	NA	VII
RCC_NL_UP	Upregulated in VHL-rescued renal carcinoma vs. normal renal cells (Fig. 2d+e)	0.705	0.027	0.588	0.418	0.662	NA	VII
CAMPTOTHECIN_PROBCELL_DN	Down-regulated in pro-B cells (FL5.12) following treatment with camptothecin	0.367	0.026	0.094	0.181	0.187	NA	VII
STRESS_ARSENIC_SPECIFIC_DN	Genes down-regulated 4 hours following arsenic treatment that discriminate arsenic from other stress agents	0.452	0.022	0.193	0.244	0.371	NA	VII
UVB_SCC_UP	Upregulated by UV-B light in squamous cell carcinoma cells	0.696	0.032	0.569	0.486	0.733	NA	VII
HDACI_COLON_BUT12HRS_DN	Downregulated by butyrate at 12 hrs in SW260 colon carcinoma cells	0.673	0.014	0.517	0.486	0.733	NA	VII
BCRABL_HL60_CDNA_DN	Down-regulated by expression of p210(BCR-ABL) in human leukemia (HL-60) cells; detected by spotted cDNA arrays	0.469	0.040	0.209	0.262	0.402	NA	VII
CMV_HCMV_6HRS_DN	Down-regulated in fibroblasts at 6 hours following infection with human cytomegalovirus (CMV)	0.434	0.004	0.179	0.184	0.291	NA	VII
H2O2_CSBRESCUED_UP	Upregulated by H2O2 in CSB-rescued fibroblasts (Table 1)	0.365	0.041	0.093	0.083	0.129	NA	VII
UVB_NHEK1_C2	Upregulated by UV-B light in normal human epidermal keratinocytes, cluster 2	0.341	0.022	0.075	0.104	0.153	NA	VII
ET743_RESIST_DN	Down-regulated in two Et-743-resistant cell lines (chondrosarcoma and ovarian carcinoma) compared to sensitive parental lines	0.518	0.019	0.266	0.187	0.307	NA	VII
HSA00010_GLYCOLYSIS_AND_GLUCONEOGENESIS	Genes involved in glycolysis and gluconeogenesis	0.625	0.043	0.424	0.167	0.257	NA	VII

APPENDIX C

Q-VALUES OF ENRICHED PATHWAYS FOR A PROSTATE CANCER STUDY

Q-values of enriched pathways detected by individual studies and MAPE methods in drug response data (column 3-7: q-value threshold 0.05 and significant q-values marked in red) and categories (column 8-9) that correspond to Figure 6 in the manuscript. “Categories comparing MAPE_P, MAPE_G & MAPE_I” correspond to the categories in Figure 2.17A and 2.17B. “Categories comparing Welsh, Singh & MAPE_I” correspond to the categories in Figure 2.17C and 2.17D.

Pathways	Description	Welsh	Singh	MAPE_P	MAPE_G	MAPE_I	CA	CB
TARTE_PC	Genes overexpressed in polyclonal plasmablastic cells (PPCs), mature plasma cells isolated from tonsils (TPCs), and mature plasma cells isolated from bone marrow (BMPCs), as compared to B cells purified from peripheral blood (PBBs) and tonsils (TBCs).	0.000	0.001	0.000	0.000	0.000	I	I
RIBOSOMAL_PROTEINS		0.004	0.000	0.000	0.000	0.000	I	I
SCHUMACHER_MYC_UP	Genes up-regulated by MYC in P493-6 (B-cell)	0.002	0.001	0.000	0.019	0.000	I	I
BHATTACHARYA_ESC_UP	Genes upregulated in undifferentiated human embryonic stem cells.	0.082	0.002	0.002	0.021	0.004	I	III
LI_FETAL_VS_WT_KIDNEY_UP	These are genes identified by simple statistical criteria as differing in their mRNA expresssion between WTs and fetal kidneys LOW	0.000	0.126	0.003	0.005	0.004	I	II
UVB_NHEK2_UP	Upregulated by UV-B light in normal human epidermal keratinocytes	0.007	0.000	0.000	0.000	0.000	I	I
CANCER_NEOPLASTIC_META_UP	Sixty-seven genes commonly upregulated in cancer relative to normal tissue, from a meta-analysis of the OncoMine gene expression database	0.000	0.000	0.000	0.000	0.000	I	I
ET743_SARCOMA_72HRS_UP	Up-regulated at48 hours following treatment with Et-743 in at least 6 of 8 sarcoma cell lines	0.003	0.110	0.002	0.007	0.004	I	II
HDACI_COLON_CUR24HRS_UP	Upregulated by curcumin at 24 hrs in SW260 colon carcinoma cells	0.099	0.005	0.003	0.038	0.006	I	III
HSA03010_RIBOSOME	Genes involved in ribosome	0.075	0.000	0.001	0.000	0.000	I	III
HUMAN_MITODB_6_2002	Mitochondrial genes	0.195	0.038	0.028	0.065	0.049	II	III
BASSO_REGULATORY_HUBS	Genes which comprise the top 1% of highly interconnected genes (major hubs) that account for most	0.141	0.074	0.012	0.353	0.022	II	IV

	of the interactions in the reconstructed regulatory networks from expression profiles in human B cells.							
MENSSEN_MYC_UP	Genes up-regulated by MYC in HUVEC (umbilical vein endothelial cell)	0.123	0.004	0.007	0.089	0.014	II	III
PENG_LEUCINE_DN	Genes downregulated in response to leucine starvation	0.063	0.002	0.001	0.280	0.001	II	III
CHANG_SERUM_RESPONSE_UP	CSR (Serum Response) signature for activated genes (Stanford)	0.082	0.074	0.002	0.732	0.004	II	IV
ZUCCHI_EPITHELIAL_UP	The 50 most upregulated genes in primary invasive breast dutcal carcinoma or metastatic breast carcinoma isolated from lymph nodes, as compared to normal mammary epithelium.	0.164	0.016	0.018	0.294	0.034	II	III
NELSON_ANDROGEN_UP	Genes upregulated by androgen in neoplastic prostate epithelium	0.093	0.233	0.013	0.294	0.025	II	IV
ROME_INSULIN_2F_UP	Genes 2fold upregulated by insulin	0.056	0.002	0.000	0.295	0.001	II	III
HCC_SURVIVAL_GOOD_VS_POOR_DN	Genes highly expressed in hepatocellular carcinoma with poor survival.	0.003	0.003	0.000	0.086	0.000	II	I
TARTE_PLASMA_BLASTIC	Genes overexpressed in mature plasma cells isolated from tonsils (TPCs) and mature plasma cells isolated from bone marrow (BMPCs) as compared to polyclonal plasmablastic cells (PPCs).	0.094	0.027	0.003	0.144	0.005	II	III
MITOCHONDRIA	Mitochondrial genes	0.103	0.020	0.004	0.186	0.007	II	III
SHIPP_FL_VS_DLBCL_DN	Genes upregulated in diffuse B-cell lymphomas (DLBCL) and downregulated in follicular lymphoma (FL) (fold change of at least 3)	0.054	0.236	0.014	0.101	0.026	II	IV
NING_COPD_UP	Upregulated genes in lung tissue of smokers with chronic obstructive pulmonary disease (COPD) vs smokers without disease (GOLD-2 vs GOLD-0)	0.151	0.237	0.014	0.814	0.026	II	IV
PENG_RAPAMYCIN_DN	Genes downregulated in response to rapamycin starvation	0.084	0.014	0.002	0.125	0.003	II	III
PENG_GLUTAMINE_DN	Genes downregulated in response to glutamine starvation	0.046	0.001	0.000	0.293	0.001	II	I
BOQUEST_CD31PLUS_VS_CD31MINUS_DN	Genes overexpressed 3-fold or more in freshly isolated CD31- versus freshly isolated CD31+ cells	0.000	0.015	0.000	0.203	0.000	II	I
NADLER_OBESITY_DN	Genes with decreased expression with obesity	0.064	0.277	0.021	0.177	0.038	II	IV
BOQUEST_CD31PLUS_VS_CD31MINUS_UP	Genes overexpressed 3-fold or more in freshly isolated CD31+ versus freshly isolated CD31- cells	0.159	0.102	0.016	0.895	0.029	II	IV
JISON_SICKLECELL_DIFF	Significantly differentially expressed genes in sickle cell patients	0.086	0.002	0.002	0.266	0.004	II	III
HEARTFAILURE_ATRIA_DN	Downregulated in the atria of failing hearts (DCM and ICM) compared to healthy controls	0.053	0.038	0.000	0.434	0.001	II	III
BRCA1_OVEREXP_PROSTATE_UP	Up-regulated with sTable, ectopic overexpression of BRCA1 in DU-145 human prostate cancer cell lines, compared to neo-only controls	0.054	0.044	0.000	0.199	0.001	II	III
PRMT5_KD_UP	Up-regulated by sTable RNAi knock-down of PRMT5 in NIH 3T3 cells	0.073	0.174	0.007	0.138	0.013	II	IV
HYPOPHYSECTOMY_RAT_UP	Up-regulated in liver, heart or kidney tissue from hypophysectomized rats (lacking growth hormone), compared to normal controls	0.135	0.004	0.009	0.056	0.017	II	III
IDX_TSA_UP_CLUSTER5	Up-regulated at 48-96 hours during differentiation of 3T3-L1 fibroblasts into adipocytes with IDX (insulin, dexamethasone and isobutylxanthine), vs. fibroblasts treated with IDX + TSA to prevent differentiation (cluster 5)	0.051	0.001	0.000	0.083	0.001	II	III
ELONGINA_KO_DN	Downregulated in MES cells from elongin-A knockout mice	0.053	0.045	0.000	0.096	0.001	II	III
AGEING_BRAIN_U	Age-upregulated in the human frontal cortex	0.000	0.171	0.007	0.295	0.013	II	II

P								
BRCA_ER_POS	Genes whose expression is consistently positively correlated with estrogen receptor status in breast cancer - higher expression is associated with ER-positive tumors	0.001	0.003	0.000	0.068	0.000	II	I
ALZHEIMERS_INCIPIENT_DN	Downregulated in correlation with incipient Alzheimer's Disease, in the CA1 region of the hippocampus	0.054	0.224	0.012	0.358	0.023	II	IV
SERUM_FIBROBLAST_CORE_UP	Core group of genes consistently up-regulated following exposure to serum in a variety of human fibroblast cell lines (higher expression in activated cells, not cell-cycle dependent)	0.010	0.073	0.002	0.189	0.003	II	II
DIAB_NEPH_DN	Downregulated in the glomeruli of cadaver kidneys from patients with diabetic nephropathy, compared to normal controls	0.002	0.164	0.005	0.456	0.010	II	II
OLD_FIBRO_DN	Downregulated in fibroblasts from old individuals, compared to young	0.192	0.085	0.028	0.054	0.038	II	IV
HDACI_COLON_CUR_UP	Upregulated by curcumin at any timepoint up to 48 hrs in SW260 colon carcinoma cells	0.194	0.017	0.028	0.482	0.049	II	III
AGED_MOUSE_HYPOTH_UP	Up-regulated in the hypothalamus of aged (22 months) BALB/c mice, compared to young (2 months) controls	0.147	0.041	0.012	0.590	0.023	II	III
HSA00051_FRUCTOSE_AND_MANNOSE_METABOLISM	Genes involved in fructose and mannose metabolism	0.096	0.067	0.003	0.531	0.006	II	IV
HSA00190_OXIDATIVE_PHOSPHORYLATION	Genes involved in oxidative phosphorylation	0.282	0.000	0.075	0.024	0.017	III	III
ELECTRON_TRANSPORT_CHAIN	Genes involved in electron transport	0.382	0.000	0.159	0.064	0.046	IV	III
IDX_TSA_UP_CLUSTER6	Strongly up-regulated at 96 hours during differentiation of 3T3-L1 fibroblasts into adipocytes with IDX (insulin, dexamethasone and isobutylxanthine), vs. fibroblasts treated with IDX + TSA to prevent differentiation (cluster 6)	0.576	0.016	0.475	0.064	0.050	IV	III
NING_COPD_DN	Downregulated genes in lung tissue of smokers with chronic obstructive pulmonary disease (COPD) vs smokers without disease (GOLD-2 vs GOLD-0)	0.087	0.315	0.029	0.658	0.053	VI	NA
PENG_GLUCOSE_DN	Genes downregulated in response to glucose starvation	0.195	0.292	0.028	0.505	0.050	VI	NA
MUNSHI_MM_UP	Genes upregulated in multiple myeloma (MM) cells versus the normal plasma cells of patients' identical twins.	0.228	0.001	0.043	0.138	0.078	VI	VII
FLECHNER_KIDNEY_TRANSPLANT_WELL_UP	Genes upreglated in well functioning transplanted kidney biopsies from sTable, immunosuppressed recipients relative to normal healthy donor kidney biopsies (median FDR < 0.16% per comparison)	0.000	0.368	0.042	0.286	0.077	VI	VI
MOREAUX_TACI_HI_IN_PPC_UP	PPC genes overexpressed in TACI low patients	0.227	0.084	0.042	0.292	0.076	VI	NA
HSIAO_LIVER_SPECIFIC_GENES	Liver selective genes	0.193	0.316	0.031	0.590	0.056	VI	NA
ET743_SARCOMA_UP	Up-regulated following treatment with Et-743 at any timepoint in at least 8 of 11 sarcoma cell lines	0.211	0.038	0.034	0.142	0.061	VI	VII
HTERT_DN	Downregulated in hTERT-immortalized fibroblasts vs. non-immortalized controls	0.192	0.369	0.043	0.435	0.078	VI	NA
LVAD_HEARTFAILURE_UP	Upregulated in the left ventricle myocardium of patients with heart failure following implantation of a left ventricular assist device	0.239	0.079	0.049	0.835	0.089	VI	NA
BLEO_MOUSE_LYMPH_LOW_24HRS_DN	Down-regulated at 24 hours following treatment of mouse lymphocytes (TK 3.7.2C) with a low dose of bleomycin	0.094	0.369	0.043	0.765	0.077	VI	NA
NAB_LUNG_UP	Up-regulated in human non-small cell lung carcinoma cell line H460 following 24-hour treatment with sodium butyrate	0.224	0.230	0.037	0.418	0.067	VI	NA

CTNNB1_oncogenic_signature	Genes selected in supervised analyses to discriminate cells expressing activated beta-catenin (CTNNB1) oncogene from control cells expressing GFP.	0.087	0.331	0.035	0.288	0.063	VI	NA
HSA00230_PURINE_METABOLISM	Genes involved in purine metabolism	0.223	0.382	0.048	0.368	0.088	VI	NA
CORDERO_KRAS_KD_VS_CONTROL_UP	Genes upregulated in kras knockdown vs control in a human cell line	0.039	0.961	0.769	1.000	1.000	NA	VI
FLECHNER_KIDNEY_TRANSPLANT_REJECTION_DN	Genes downreglated in acute rejection transplanted kidney biopsies relative to well functioning transplanted kidney biopsies from sTable, immunosuppressed recipients. (median FDR < 0.14% per comparison).	0.001	0.501	0.113	0.210	0.203	NA	VI
BASSO_GERMINAL_CENTER_CD40_DN	CD40 down-regulated genes	0.003	0.951	0.732	0.482	0.609	NA	VI
IRITANI_ADPROX_VASC	BLOOD VASCULAR EC	0.022	0.964	0.972	1.000	1.000	NA	VI
IGLESIAS_E2FMINUS_UP	Genes that increase in the absence of E2F1 and E2F2	0.031	0.983	0.971	0.994	1.000	NA	VI
ATRIA_UP	Upregulated in the atria of healthy hearts, compared to venticles	0.006	0.532	0.149	0.195	0.169	NA	VI
TGFBETA_ALL_UP	Upregulated by TGF-beta treatment of skin fibroblasts, at any timepoint	0.046	0.962	0.966	1.000	1.000	NA	VI
ELONGINA_KO_UP	Upregulated in MES cells from elongin-A knockout mice	0.006	0.530	0.137	0.447	0.250	NA	VI
CMV_24HRS_DN	Downregulated at 24hrs following infection of primary human foreskin fibroblasts with CMV	0.006	0.953	0.736	1.000	1.000	NA	VI
AGEING_KIDNEY_SPECIFIC_UP	Up-regulation is associated with increasing age in normal human kidney tissue from 74 patients, and expression is higher in kidney than in whole blood	0.003	0.605	0.219	0.487	0.382	NA	VI
CMV_ALL_DN	Downregulated at any timepoint following infection of primary human foreskin fibroblasts with CMV	0.018	1.000	0.912	1.000	1.000	NA	VI
BAF57_BT549_UP	Up-regulated following sTable re-expression of BAF57 in Bt549 breast cancer cells that lack functional BAF57	0.000	0.695	0.300	0.769	0.507	NA	VI
HSA04512_ECM_RECEPTOR_INTERACTION	Genes involved in ECM-receptor interaction	0.025	0.993	0.949	1.000	1.000	NA	VI
ZELLER_MYC_UP	Genes up-regulated by MYC in >3 papers.	0.488	0.010	0.295	0.591	0.499	NA	VII
POMEROY_DESMOPLASIC_VS_CLASSIC_MD_UP	Genes expressed in desmoplastic medulloblastomas. (p < 0.01)	0.364	0.005	0.144	0.645	0.257	NA	VII
PROTEASOME_DEGRADATION	Genes involved in proteasome degradation	0.680	0.004	0.696	0.822	1.000	NA	VII
MOOTHA_VOXPHOS	Oxidative Phosphorylation	0.521	0.000	0.374	0.291	0.289	NA	VII
OXIDATIVE_PHOSPHORYLATION		0.493	0.007	0.310	0.360	0.427	NA	VII
POMEROY_MD_TREATMENT_GOOD_VS_POOR_DN	Genes highly associated with medulloblastoma treatment failure	0.494	0.005	0.316	0.276	0.317	NA	VII
FLOTHO_CASP8AP2_MRD_DIFF	Genes significantly associated with MRD on day 46	0.670	0.021	0.677	0.356	0.425	NA	VII
MOREAUX_TACI_HI_VS_LOW_DN	Genes overexpressed in TACI low patients	0.434	0.033	0.223	0.332	0.381	NA	VII
MUNSHI_MM_VS_PCS_UP	Selected up-regulated genes in patient MM cells versus normal twin PCs	0.398	0.001	0.181	0.139	0.116	NA	VII
BLEO_MOUSE_LYMPH_HIGH_24HRS_DN	Down-regulated at 24 hours following treatment of mouse lymphocytes (TK 3.7.2C) with a high dose of bleomycin	0.494	0.017	0.320	0.450	0.545	NA	VII
HDACI_COLON_CUR48HRS_UP	Upregulated by curcumin at 48 hrs in SW260 colon carcinoma cells	0.419	0.040	0.197	0.661	0.338	NA	VII
GENOTOXINS_24H	Group of genes whose regulation pattern significantly	0.697	0.011	0.729	0.843	1.000	NA	VII

RS_DISCR	discriminates between direct (cisplatin, methyl methanesulfonate, mitomycin C) and indirect (taxol, hydroxyurea, etoposide) genotoxins, 24 hours following treatment of mouse lymphocytes (TK 3.7.2C)							
IFN_BETA_GLIOMA_DN	Down-regulated in human glioma cells (T98) at 48 hours following treatment with interferon-beta	0.382	0.039	0.160	0.354	0.288	NA	VII
UVB_NHEK3_ALL	Regulated by UV-B light in normal human epidermal keratinocytes	0.292	0.021	0.082	0.206	0.152	NA	VII
UVB_NHEK1_UP	Upregulated by UV-B light in normal human epidermal keratinocytes	0.514	0.011	0.360	0.945	0.607	NA	VII
CANTHARIDIN_DN	Downregulated in HL-60 promyeloid leukemic cells after treatment with the cytotoxic drug cantharidin	0.311	0.011	0.094	0.195	0.168	NA	VII
HIPPOCAMPUS_DEVELOPMENT_PRENATAL	Highly expressed in prenatal mouse hippocampus (cluster 1)	0.609	0.000	0.553	0.290	0.336	NA	VII
UVB_NHEK1_C1	Upregulated by UV-B light in normal human epidermal keratinocytes, cluster 1	0.762	0.001	0.899	0.643	0.860	NA	VII
BRCA1_OVEREXP_DN	Downregulated by induction of exogenous BRCA1 in EcR-293 cells	0.705	0.017	0.758	0.344	0.392	NA	VII
HSA03050_PROTEASOME	Genes involved in proteasome	0.607	0.000	0.557	0.591	0.760	NA	VII

BIBLIOGRAPHY

Biocarta Pathway Collections, *http://www.biocarta.com/genes/allPathways.asp*.

Ackermann, M. and Strimmer, K. (2009) A general modular framework for gene set enrichment analysis, *BMC Bioinformatics*, **10**, 47.

Annie J. Sasco, Albert B. Lowenfels and Pieternel Pasker-De Jong (1993) Review article: Epidemiology of male breast cancer. A meta-analysis of published case-control studies and discussion of selected aetiological factors, *International Journal of Cancer*, **53**, 538-549.

Auer, H., Newsom, D.L. and Kornacker, K. (2009) Expression Profiling Using Affymetrix GeneChip Microarrays, *Methods Mol Biol*, **509**, 35-46.

Bair, E., *et al.* (2006) Prediction by Supervised Principal Components, *Journal of the American Statistical Association*, **101**, 119-137.

Bair, E. and Tibshirani, R. (2004) Semi-supervised methods to predict patient survival from gene expression data, *PLoS Biol*, **2**, E108.

Barrett, T., *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data, *Nucleic Acids Res*, **37**, D885-890.

Beer, D., *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nature Medicine*, **9**, 816 - 824.

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 Series B*, **57**, 289–300.

Berriz, G.F., *et al.* (2003) Characterizing gene sets with FuncAssociate, *Bioinformatics*, **19**, 2502-2504.

Bertheau, P., *et al.* (2007) Exquisite sensitivity of TP53 mutant and basal breast cancers to a dose-dense epirubicin-cyclophosphamide regimen, *PLoS Med*, **4**, e90.

Bhattacharjee, A., *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 13790-13795.

Birnbaum, A. (1954) Combining independent tests of significance, *Journal of the American Statistical Association*, **49**, 559-574.

Bolstad, B.M., *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics*, **19**, 185-193.

Borovecki, F., *et al.* (2005) Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 11023-11028.

Brazma, A., *et al.* (2001) Minimum information about a microarray experiment (MIAME)[mdash]toward standards for microarray data, *Nat Genet*, **29**, 365-371.

Bruford, E.A., *et al.* (2008) The HGNC Database in 2008: a resource for the human genome, *Nucleic Acids Res*, **36**, D445-448.

Buzdar, A.U., *et al.* (2005) Significantly Higher Pathologic Complete Remission Rate After Neoadjuvant Therapy With Trastuzumab, Paclitaxel, and Epirubicin Chemotherapy: Results of a Randomized Trial in Human Epidermal Growth Factor Receptor 2-Positive Operable Breast Cancer, *J Clin Oncol*, **23**, 3676-3685.

Cardoso, J., *et al.* (2007) Expression and genomic profiling of colorectal cancer, *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, **1775**, 103-137.

Chang, J.C., *et al.* (2003) Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer, *Lancet*, **362**, 362-369.

Choi, H., *et al.* (2007) A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments, *BMC Bioinformatics*, **8**, 364.

Choi, J.K., *et al.* (2003) Combining multiple microarray studies and modeling interstudy variation, *Bioinformatics*, **19**, i84-90.

Conlon, E.M., Song, J.J. and Liu, A. (2007) Bayesian meta-analysis models for microarray data: a comparative study, *BMC Bioinformatics*, **8**, 80.

Dahlquist, K.D., *et al.* (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways, *Nat Genet*, **31**, 19-20.

Dan, S., *et al.* (2002) An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines, *Cancer Res*, **62**, 1139-1147.

DeRisi, J., *et al.* (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer, *Nat Genet*, **14**, 457-460.

Dorum, G., *et al.* (2009) Rotation testing in gene set enrichment analysis for small direct comparison experiments, *Stat Appl Genet Mol Biol*, **8**, Article34.

Draghici, S., *et al.* (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate, *Nucl. Acids Res.*, **31**, 3775-3781.

Dudoit, S., Gentleman, R.C. and Quackenbush, J. (2003) Open source software for the analysis of microarray data, *BioTechniques*, **Suppl**, 45-51.

Edgar, R. and Barrett, T. (2006) NCBI GEO standards and services for microarray data, *Nat Biotechnol*, **24**, 1471-1472.

Efron, B. and Tibshirani, R. (2007) On testing the significance of sets of genes, *Annals of Applied Statistics*, **1**, 107-129.

Ein-Dor, L., *et al.* (2005) Outcome signature genes in breast cancer: is there a unique set?, *Bioinformatics*, **21**, 171-178.

Falcon, S. and Gentleman, R. (2007) Using GOstats to test gene lists for GO term association, *Bioinformatics*, **23**, 257-258.

Fan, J.B., *et al.* (2006) Illumina universal bead arrays, *Methods Enzymol*, **410**, 57-73.

Fields Development Team (2006) Fields: Tools for Spatial Data. National Center for Atmospheric Research, Boulder, CO.

Garman, K.S., Nevins, J.R. and Potti, A. (2007) Genomic strategies for personalized cancer therapy, *Hum. Mol. Genet.*, **16**, R226-232.

Geller, S.C., *et al.* (2003) Transformation and normalization of oligonucleotide microarray data, *Bioinformatics*, **19**, 1817-1823.

Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006, *Nucl. Acids Res.*, **34**, D322-326.

Gentleman, R., *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics, *Genome Biology*, **5**, R80.

Gentleman, R.C., *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics, *Genome Biol*, **5**, R80.

Ghosh, D., *et al.* (2003) Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer *Functional and Integrative Genomics*, **3**, 180-188.

Gianni, L., *et al.* (2005) Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer, *J Clin Oncol*, **23**, 7265-7277.

Goeman, J.J. and Buhlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues, *Bioinformatics*, **23**, 980-987.

Gonzalez-Angulo, A.M., Morales-Vasquez, F. and Hortobagyi, G.N. (2007) Overview of resistance to systemic therapy in patients with breast cancer, *Adv Exp Med Biol*, **608**, 1-22.

Goods, I.J. (1955) On the Weighted Combination of Significance Tests, *Journal of the Royal Statistical Society: Series B*, **17**, 264-265

Haigh, P.I., *et al.* (2000) Biopsy method and excision volume do not affect success rate of subsequent sentinel lymph node dissection in breast cancer, *Ann Surg Oncol*, **7**, 21-27.

Hedges, L.V. (1992) Meta-analysis, *Journal of Educational Statistic*, **17**, 279-296.

Hess, K.R., *et al.* (2006) Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer, *J Clin Oncol*, **24**, 4236-4244.

Hettema, J.M., Neale, M.C. and Kendler, K.S. (2001) A Review and Meta-Analysis of the Genetic Epidemiology of Anxiety Disorders, *Am J Psychiatry*, **158**, 1568-1578.

Hoeflich, K.P., *et al.* (2009) In vivo Antitumor Activity of MEK and Phosphatidylinositol 3-Kinase Inhibitors in Basal-Like Breast Cancer Models, *Clinical Cancer Research*, **15**, 4649-4664.

Hsu, D.S., *et al.* (2007) Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer, *J Clin Oncol*, **25**, 4350-4357.

Hu, P., Greenwood, C. and Beyene, J. (2005) Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models, *BMC Bioinformatics*, **6**, 128.

Huang, F., *et al.* (2007) Identification of Candidate Molecular Markers Predicting Sensitivity in Solid Tumors to Dasatinib: Rationale for Patient Selection, *Cancer Res*, **67**, 2226-2238.

Irizarry, R.A., *et al.* (2003) Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Res*, **31**, e15.

Irizarry, R.A., *et al.* (2003) Summaries of Affymetrix GeneChip probe level data, *Nucl. Acids Res.*, **31**, e15-.

Irizarry, R.A., *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, **4**, 249-264.

Iwao-Koizumi, K., *et al.* (2005) Prediction of docetaxel response in human breast cancer by gene expression profiling, *J Clin Oncol*, **23**, 422-431.

Jemal, A., *et al.* (2008) Cancer statistics, 2008, *CA Cancer J Clin*, **58**, 71-96.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucl. Acids Res.*, **28**, 27-30.

Kang, H.C., *et al.* (2004) Identification of genes with differential expression in acquired drug-resistant gastric cancer cells using high-density oligonucleotide microarrays, *Clin Cancer Res*, **10**, 272-284.

Kauffmann, A., *et al.* (2009) Importing ArrayExpress datasets into R/Bioconductor, *Bioinformatics*.

Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems, *Bioinformatics*, **21**, 3587-3595.

Kim, C., Taniyama, Y. and Paik, S. (2009) Gene expression-based prognostic and predictive markers for breast cancer: a primer for practicing pathologists, *Arch Pathol Lab Med*, **133**, 855-859.

Kornblith, P., *et al.* (2004) Differential in vitro effects of chemotherapeutic agents on primary cultures of human ovarian carcinoma, *Int J Gynecol Cancer*, **14**, 607-615.

Kornblith, P., *et al.* (2003) In vitro responses of ovarian cancers to platinums and taxanes, *Anticancer Res*, **23**, 543-548.

Koziol, A.J. and Perlman, D.M. (1978) Combining Independent Chi-Squared Tests, *Journal of the American Statistical Association*, **73**, 753-763

Kulesh, D.A., *et al.* (1987) Identification of interferon-modulated proliferation-related cDNA sequences, *Proc Natl Acad Sci U S A*, **84**, 8453-8457.

Kuo, W., *et al.* (2002) Analysis of matched mRNA measurements from two different microarray technologies, *Bioinformatics*, **18**, 405 - 412.

Lapointe, J., *et al.* (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer, *Proc Natl Acad Sci U S A*, **101**, 811-816.

Lashkari, D.A., *et al.* (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis, *Proc Natl Acad Sci U S A*, **94**, 13057-13062.

Lee, J.K., *et al.* (2010) Prospective comparison of clinical and genomic multivariate predictors of response to neoadjuvant chemotherapy in breast cancer, *Clin Cancer Res*, **16**, 711-718.

Lee, J.K., *et al.* (2007) A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery, *Proc Natl Acad Sci U S A*, **104**, 13086-13091.

Li , J. (2008) Statistical issues in meta-analysis for identifying signature genes in the integration of multiple genomic studies, *PhD dissertation, University of Pittsburgh*, 18-36.

Liedtke, C., *et al.* (2009) Genomic grade index is associated with response to chemotherapy in patients with breast cancer, *J Clin Oncol*, **27**, 3185-3191.

Liedtke, C., *et al.* (2009) Clinical evaluation of chemotherapy response predictors developed from breast cancer cell lines *Breast Cancer Research and Treatment*.

Liedtke, C., *et al.* (2009) Clinical evaluation of chemotherapy response predictors developed from breast cancer cell lines, *Breast Cancer Res Treat*.

Loughin, T.M. (2004) A systematic comparison of methods for combining p-values from independent tests, *Computational Statistics & Data Analysis*, **47**, 467-485.

Maglott, D., *et al.* (2005) Entrez Gene: gene-centered information at NCBI, *Nucleic Acids Res*, **33**, D54-58.

Manoli, T., *et al.* (2006) Group testing for pathway analysis improves comparability of different microarray datasets, *Bioinformatics*, **22**, 2500-2506.

Marchionni, L., *et al.* (2008) Systematic review: gene expression profiling assays in early-stage breast cancer, *Ann Intern Med*, **148**, 358-369.

Mariadason, J.M., *et al.* (2003) Gene expression profiling-based prediction of response of colon carcinoma cells to 5-fluorouracil and camptothecin, *Cancer Res*, **63**, 8791-8812.

Marsaglia, G., Tsang, W.W. and Wang, J. (2003) Evaluating Kolmogorov's distribution, *Journal of Statistical Software*, **8**.

Mehta, C.R., Patel, N.R. and Tsiatis, A.A. (1984) Exact significance testing to establish treatment equivalence with ordered categorical data, *Biometrics*, **40**, 819-825.

Mosteller, F. and Fisher, R.A. (1948) Questions and Answers, *The American Statistician*, **2**, 30-31.

Nam, D. and Kim, S.-Y. (2008) Gene-set approach for expression pattern analysis, *Brief Bioinform*, bbn001.

Neve, R.M., *et al.* (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes *Cancer Cell*, **10**, 515-527.

Newton, M., *et al.* (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis, *Ann. Appl. Stat.*, **1**, 85-106.

Nie, H., *et al.* (2009) Microarray data mining using Bioconductor packages, *BMC Proc*, **3 Suppl 4**, S9.

Paik, S., *et al.* (2006) Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer, *J Clin Oncol*, **24**, 3726-3734.

Pearson, K. (1904) Report on certain enteric fever inoculation statistics, *British Medical Journal*, **3**, 1243-1246.

Pirooznia, M., Nagarajan, V. and Deng, Y. (2007) GeneVenn - A web application for comparing gene lists using Venn diagrams, *Bioinformation*, **1**, 420-422.

Potti, A., *et al.* (2006) Genomic signatures to guide the use of chemotherapeutics, *Nat Med*, **12**, 1294-1300.

Potti, A. and Nevins, J.R. (2008) Utilization of genomic signatures to direct use of primary chemotherapy, *Curr Opin Genet Dev*, **18**, 62-67.

Pusztai, L., Anderson, K. and Hess, K.R. (2007) Pharmacogenomic predictor discovery in phase II clinical trials for breast cancer, *Clin Cancer Res*, **13**, 6080-6086.

Quackenbush, J. (2002) Microarray data normalization and transformation, *Nat Genet*, **32 Suppl**, 496-501.

R Development Core Team (2005) R: A language and environment for statistical computing, *R Foundation for Statistical Computing*.

Rhodes, D.R., *et al.* (2002) Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer, *Cancer Res*, **62**, 4427-4433.

Rustici, G., *et al.* (2008) Data storage and analysis in ArrayExpress and Expression Profiler, *Curr Protoc Bioinformatics*, **Chapter 7**, Unit 7 13.

Salter, K.H., *et al.* (2008) An integrated approach to the prediction of chemotherapeutic response in patients with breast cancer, *PLoS One*, **3**, e1908.

Schadt, E.E., *et al.* (2001) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data, *J Cell Biochem Suppl*, **Suppl 37**, 120-125.

Schena, M., *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, **270**, 467-470.

Schmid, J.E., Koch, G.G. and LaVange, L.M. (1991) An overview of statistical issues and methods of meta-analysis, *J Biopharm Stat*, **1**, 103-120.

Segal, E., *et al.* (2004) A module map showing conditional activity of expression modules in cancer, *Nat. Genet.*, **36**, 1090-1098.

Shen, R., Ghosh, D. and Chinnaiyan, A. (2004) Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data, *BMC Genomics*, **5**, 94.

Singh, D., *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, **1**, 203-209.

Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Stat Appl Genet Mol Biol*, **3**, Article3.

Sotiriou, C. and Pusztai, L. (2009) Gene-expression signatures in breast cancer, *N Engl J Med*, **360**, 790-800.

Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis, *J Mol Biol*, **98**, 503-517.

Stalteri, M.A. and Harrison, A.P. (2007) Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips, *BMC Bioinformatics*, **8**, 13.

Staunton, J.E., *et al.* (2001) Chemosensitivity prediction by transcriptional profiling, *Proc Natl Acad Sci U S A*, **98**, 10787-10792.

Steinfath, M., *et al.* (2001) Automated image analysis for array hybridization experiments, *Bioinformatics*, **17**, 634-641.

Storey, D.J. (2002) A direct approach to false discovery rates, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **46**, 479-498.

Stouffer, S.A., *et al.* (1949) The American Soldier, volumn I: Adjustment during Army Life, *Princeton University Press*.

Stroup, D.F., *et al.* (2000) Meta-analysis of Observational Studies in Epidemiology: A Proposal for Reporting, *JAMA*, **283**, 2008-2012.

Stuart, R.O., *et al.* (2004) In silico dissection of cell-type-associated patterns of gene expression in prostate cancer, *Proc Natl Acad Sci U S A*, **101**, 615-620.

Subramanian, A., *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545-15550.

Tian, L., *et al.* (2005) Discovering statistically significant pathways in expression profiling studies, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 13544-13549.

Tippett, L.H.C. (1931) The Methods in Statistics, *Williams and Norgate,Ltd.*

Tomfohr, J., Lu, J. and Kepler, T.B. (2005) Pathway level analysis of gene expression using singular value decomposition, *BMC Bioinformatics*, **6**, 225.

Tordai, A., *et al.* (2008) Evaluation of biological pathways involved in chemotherapy response in breast cancer, *Breast Cancer Research*, **10**, R37.

Tseng, G.C., *et al.* (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects, *Nucl. Acids Res.*, **29**, 2549-2557.

Tu, Y., Stolovitzky, G. and Klein, U. (2002) Quantitative noise analysis for gene expression microarray experiments, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 14031-14036.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116-5121.

van de Vijver, M.J., *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer, *N Engl J Med*, **347**, 1999-2009.

Varambally, S., *et al.* (2005) Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression, *Cancer Cell*, **8**, 393-406.

Wang, Y., *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, *Lancet*, **365**, 671-679.

Welsh, J.B., *et al.* (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer, *Cancer Res*, **61**, 5974-5978.

Wheeler, D.L., *et al.* (2003) Database resources of the National Center for Biotechnology, *Nucleic Acids Res*, **31**, 28-33.

Wilkinson, B. (1951) A statistical consideration in psychological research, *Psychological Bulletin*, **48**, 156-158.

Yu, Y.P., *et al.* (2004) Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy, *J Clin Oncol*, **22**, 2790-2799.

Zeeberg, B.R., *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data, *Genome Biol*, **4**, R28.

Zhong, S., Li, C. and Wong, W.H. (2003) ChipInfo: software for extracting gene annotation and gene ontology information for microarray analysis, *Nucl. Acids Res.*, **31**, 3483-3486.