Stony Brook University

Statistical Modeling for Multiplex RNAi Screen Data Analysis

A Dissertation Presented

by

Jianping Zhang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

December 2010

Stony Brook University

The Graduate School

Jianping Zhang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Scott Powers - Dissertation Advisor
Professor, Department of Applied Mathematics and Statistics

Nancy Mendell - Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Wei Zhu - Dissertation Co-Advisor
Professor, Department of Applied Mathematics and Statistics

Ellen Li
Professor, Department of Medicine, Stony Brook University

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Statistical Modeling for Multiplex RNAi Screen Data Analysis

by

Jianping Zhang

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2010

Multiplex RNAi screen is an emerging tool for functional genomics. Most analysis methods presently available for Multiplex RNAi screen are based on single hairpin data. These approaches have serious limitations. They do not account for the redundancies in genome-scale libraries. Thus it is difficult to detect genes with modest but consistent effect. In addition, contradictory conclusions might be reached based on enriched and depleted hairpins for the same gene. Therefore, we propose the RNAi Set Enrichment Analysis (RSEA) framework based on the gene set enrichment analysis framework that will take multiple hairpins into consideration in accessing the gene effect on drug response. The gene set enrichment analysis has been widely used in gene expression microarray study to test whether a certain biological pathway is activated under some treatment. However this method is rarely used in RNAi screen studies. With the RSEA method, we evaluate and compare the performance of different RNAi level statistics, RNAi set statistics and significance assessment choices. Besides these, to model the silencing efficiency and off target effect of RNAi knockdown, we propose Structural Equation Modeling (SEM) with latent variables for RNAi screen data analysis. SEM is intuitive for biological researchers with its path diagrams. In addition, the latent SEM contains the repeated measures ANOVA, both the univariate and the multivariate approaches, as special cases. Our simulation studies revealed that the latent SEM has comparable statistical power to RSEA method when the hairpin off target effect is modest. While the adoption of the SEM to existing experimental data is hampered by the modest sample size, we are able to verify the RSEA method by applying them towards real data generated from our experiments. The result shows that RSEA can successfully identify positive genes whose effects have been validated by the follow-up confirmatory experiments.

Table of Contents

List of Figures

List of Tables

Chapter 1 . Introduction

Cancer is a genetic disease characterized by multiple mutations in the cancer genome, changes in genome copy number and alterations in patterns of epigenetic modification [1]. These multiple alterations are necessary for the development and maintenance of the tumor phenotype and there is redundancy in the pathways of proteins that are deregulated. For most epithelial tumors, it is unlikely that inhibiting one target will be sufficient to inhibit the proliferation of the tumor cell and ultimately kill the tumor cell. Understanding the combination of targets that need to be inhibited is critical for the successful development of novel targeting agents.

The multiplex RNAi screen is an emerging tool for functional genomics [2, 3]. It offers an approach to rapidly screen multiple proteins to identify rational targets to inhibit in combination with the novel targeting agents.

Most presently available analysis methods for large-scale RNAi screen rely on ranking screen data and are based on single RNAi activity or significance value [4-6]. These analyses focus on the identification of highly active RNAis and ignore much of the remainder. These analyses have the following major limitations:

1. After correcting for multiple hypotheses testing, no individual RNAi may meet the threshold for statistical significance, because the relevant biological differences are modest relative to the noise inherent to the microarray technology [7].

2. Statistically significant RNAis with positive and negative activity may both be obtained for the same gene. This will render biological interpretations difficult.

3. These strategies do not exploit redundancies in genome-scale libraries, which typically contain 2~4 RNAis per gene. Thus, it is difficult to systematically identify genes for which multiple RNAis are moderately but consistently active across a screen, which do not fall within an upper threshold [8].

Facing these challenges, Konig et al. developed a statistical score that models the probability of a gene ‘hit’ based on the collective activities of multiple RNAis per gene [8]. In their redundant siRNA activity (RSA) analysis, all RNAis are initially ranked according to their signals. Then, the rank distribution of all RNAis targeting the same gene is examined and a P-value is assigned. Thus, P-value indicates the statistical significance of all RNAis targeting a single gene being unusually distributed toward the top ranking slots, calculated based on an iterative hypergeometric distribution formula. This method is an improvement over methods based on single RNAi activity. However, they didn’t thoroughly investigate its performance or efficiency. Luo et al. proposed RSecond, in multiplex RNAi screen, the sample size is usually small (3~4 samples in each class in our case) and the RNAi redundancy is also small (3 RNAis on average per gene for our case), the statistical power for this method is relatively low compared to other methods, which is shown in our simulation study. Luo et al. proposed RNAi gene enrichment ranking (RIGER) method to

analyze their RNAi screen data [9]. RIGER is actually an adaptation from Gene Set Enrichment Analysis method [7] which is widely applied in gene expression microarray studies. Again, they didn't thoroughly investigate the performance and efficiency of their method.

In this thesis, we propose a general RNAi Set Enrichment Analysis framework based on the traditional gene set enrichment analysis methods applied in gene expression data. RNAi Set Enrichment Analysis uses RNAi level statistic to access individual RNAi activity. It then uses RNAi set statistic to access the gene effect by combining the activity of all the RNAis targeting this gene. Finally, it uses sample permutation or RNAi resampling to access the statistical significance of each gene. For RNAi level statistic, we have multiple choices, such as student's t statistic, regularized t statistic and the SAM ('significance analysis of microarray') t statistic. For RNAi set statistic, we also have multiple choices, such as Wilcoxon rank statistic, max-mean statistic, mean and Kolmogorov-Smirnov statistic. In our study, we investigate the performance of different statistic choices for different sample sizes and numbers of RNAis targeting the same gene based on simulated data

In addition to RSEA, we apply Structural Equation Modeling (SEM) to analyze multiplex RNAi screen data. SEM takes RNAi silence efficiency and off target effect into consideration and is intuitive to biological researchers. We propose three different models based on SEM by different assumption on the covariance structure. Based on the simulated data, we evaluate and compare the performances of the three models.

Besides the simulation studies, we apply RSEA on our multiplex RNAi screen real data. RSEA analysis results show that gene RARA is PLKi drug resistant in four out of five cell lines. And the validation experiments confirm that silencing RAR confers PLKi resistance and RARA activation confers PLKi sensitivity in H460 cancer cell line.

The remaining of this document is divided as follows: Chapter 2 introduces the purpose and design of multiplex RNAi screen. Chapter 3 summarizes presently available methods for RNAi screen data analysis. Chapter 4 presents the details of pre-processing of RNAi screen data. Chapter 5 proposes the general framework of RNAi Set Enrichment Analysis. Chapter 6 applies the Structural Equation Modeling for RNAi Screen. Chapter 7 presents the simulation study result for RNAi Set Enrichment Analysis and SEM. Chapter 8 is about the application of RNAi Set Enrichment Analysis on our multiplex RNAi screen data and the validation experiment results. Chapter 9 discusses choices between RSEA and SEM, and the future work.

Chapter 2 . Multiplex RNAi Screen Experiment

Scholl et al. gave a very good introduction on the application of RNAi screen [10], we summarize their introduction in the paragraph below.

The identification of genes that are causally implicated in human cancer has resulted in novel, pathogenesis-oriented treatment strategies [11]. However, many known oncogenes, are challenging therapeutic targets. For example, researchers have discovered RAS gene family members are mutated in approximately 30% of human tumors and cancer cells are dependent on mutant RAS for their viability and proliferation for a long time, however, efforts in developing drugs to inhibit oncogenic RAS proteins have been largely unsuccessful [12, 13]. One important reason for this challenge is that cancer cells may also develop secondary dependencies on genes that are not oncogenes. Perturbation of these genes may result in oncogene-specific "synthetic lethal" interactions that could provide new therapeutic opportunities [14, 15]. Synthetic lethality occurs when alteration of a gene results in cell death only in the presence of another nonlethal genetic alteration, such as a cancer-associated mutation [10]. Synthetic lethal interactions were first described in model organisms [16, 17], but recent studies indicate that it can be extended to mammalian cells [18, 19]. One way to identify such synthetic lethal genes in human cancer is to systematically determine the functional consequences of gene suppression in cancer cell lines using RNA interference (RNAi) technology [20-22]. For example, In 2006, Ngo et al. identified CARD11 as a regulator of constitutive NFKB signaling in the activated B cell-like DLBCL subtype through RNAi screen [23]. Similarly, functional genetic screens have identified genes whose suppression sensitizes cancer cell lines [24, 25] or untransformed cells engineered to ectopically express a specific oncogene [26] to the effects of defined environmental conditions, such as the presence of a therapeutic agent.

In our study, we use multiplex RNAi screen to identify synthetic lethal genetic interactions. One of the projects is to identify the genes which are able to enhance or weaken the treatment effect of Polo-like kinase 1 inhibitor (PLKi) drug for non-small cell lung cancer, in collaboration with GlaxoSmithKline plc. In this thesis, we focus on the data analysis of the PLKi RNAi screen project. The experiment procedure and design are present below.

2.1 Experiment Procedure

The Multiplex RNAi screen experiment procedure is illustrated in Figure 2.1.

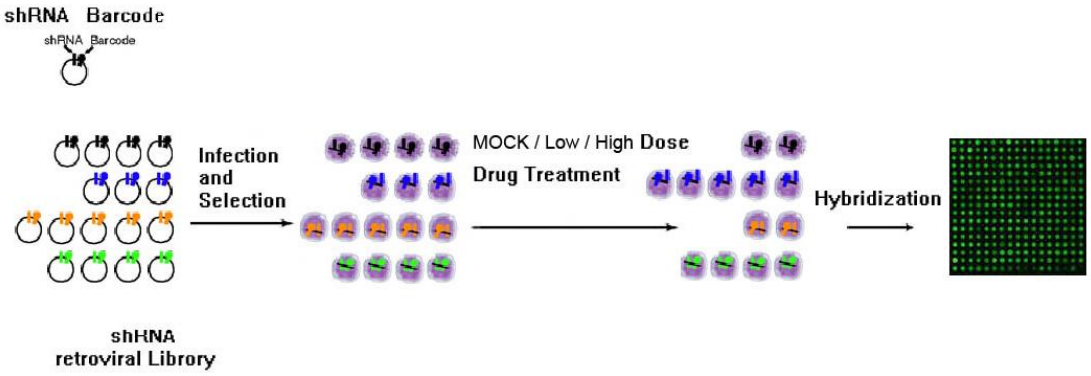

Figure 2.1 Illustration of the Multiplex RNAi screen experimental procedure

Step1 - Infection and Selection: An RNAi library is used to infect the target genetically identical human cancer cells derived from a specific cell line (5 different cell lines are tested separately in our experiments). An RNAi library contains around 4500 RNAis, each with around 1500 copies. Each individual cell receives a single copy of RNAi which silences down its target gene. One average, each gene has 3 RNAis targeting different positions. Uninfected cells are removed via a puromycin selection marker. After RNAi infection, cells are cultured for several days.

Step2 - Drug Treatment: After step 2, cells are mock / low dose drug / high dose drug treated. Upon completion of treatment, the cells are allowed to recover for a period of 7 days. The mock treatment doesn't contain any drug and is used as negative control.

Step3 - Hybridization: Genomic DNA is extracted from the human cancer cells, labeled with Cy3 dye and hybridized to microarray chips.

2.2 Experimental Design

Figure 2.2 illustrates the experiment design from another perspective. Cancer cells genetically identical are divided into 3 groups (DMSO: Mock treatment group; LOW: low dose drug treatment group; HIGH: high dose drug treatment group), 3 replicate plates for each group. In the PLKi RNAi screen project, we actually have an additional group NONE, which is used for background estimation. In the NONE group, similar to the DMSO group, there is no drug treatment. The difference between NONE and DMSO is that: in NONE group, the RNAis haven't taken effect yet in the cancer cells infected with the RNAi library, which means these RNAi haven't silenced down their target gene expression yet.

In the drug treatment step, the low dose drug kills 20% of the total cells in each of the 3 plates, while the high dose drug kills 80% of the total cells in each of the 3 plates.

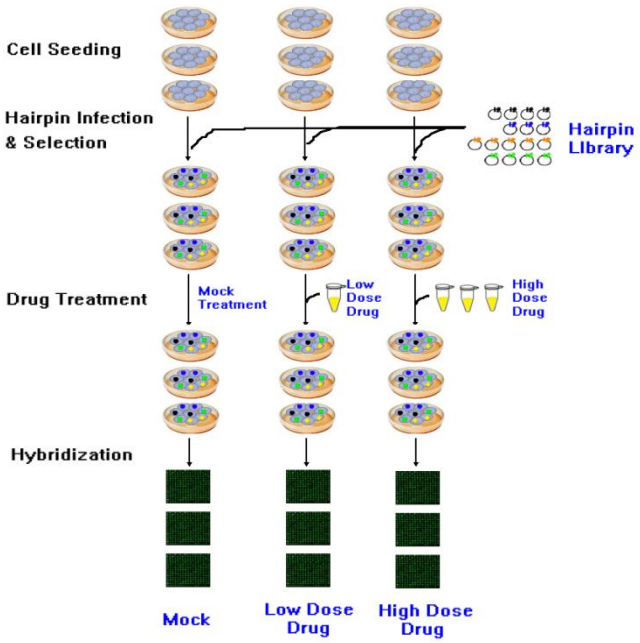

Figure 2.2 RNAi screen experimental design

In the PLKi multiplex RNAi screen experiment, we screened five different cell lines: A549, H522, H322, J42-L83 and H460. Each cell line has 4 groups (NONE, DMSO, LOW and HIGH, see descriptions above). For cell lines A549, H522, H322 and J42-L83, there are three technical replicate plates in each group. For cell line H460, there are four technical replicates for each group. These cell lines have different drug sensitivity. All five cell lines are responsive to PLK inhibitor during the regular 3 day treatment. However, upon washing away the drug to allow the cells to recover for a 3-day period, the difference begins to show: H460, J42-L83 and A549 cells are able to grow back and hence drug resistant. H522 and H322 in contrast, are unable to grow back and hence drug sensitive. Overall, the five cell lines do not differ much in resistance or sensitivity to PLK inhibitor. So far for all the five cell lines tested, there is no cell line that is completely un-responsive to the PLK inhibitor.

In the data analysis, we mainly focus on the comparison between the DMSO and HIGH groups.

Chapter 3 . Existing Data Analysis Methods

Currently, most of the data analysis for RNAi screen is based on single RNAi analysis, which treats the each individual RNAi separately, rather than considering the redundancy in the RNAi library design. To take multiple RNAis into consideration when evaluating the gene effect on drug resistance, Konig et al. [8] proposes the Redundant siRNA Activity (RSA) method. The RSA method analyzes the collective behavior of all siRNAs targeting a gene. Konig et al. demonstrate that RSA outperforms single RNAi analysis in the identification of confirmable activities, even though it's limited by false positive activities. Luo et al. developed RNAi gene enrichment ranking (RIGER) method to analyze their RNAi screen data [9]. In their RNAi library design, each human gene has 5 independent shRNAs. They used RIGER to rank genes based on these multiple shRNAs. Actually their RIGER is an adaption from the Gene Set Enrichment Analysis (GSEA) method utilized in gene expression study [7]. They added Second Best Rank and Weighted Sum as options for the RNAi level statistic, besides the Kolmogorov-Smirnov statistic used in GSEA. The details about these methods are presented below.

3.1 Single RNAi Analysis

Zhang et al. gave a very good review for the single hairpin analysis [27]. We summarize their review in this paragraph. For single RNAi analysis, there are two major types of approaches: one is the use of analytic metrics to assess and rank the size of RNAi effects and the other is the use of hypothesis testing to control false positive and false negative rates [27]. In the first approach, fold change, mean difference, percent activity, percent viability, percent inhibition and strictly standardized mean difference have already been proposed and explored [28-32]. In the second approach, the most popular methods are the use of z-score or t-test for testing the null hypothesis that no difference exists between the means, i.e. mean $\pm k$SD or median $\pm k$MAD (median absolute deviation) [33-39]. These methods usually control for the false positive and false negative rates based on a single test. Given that a large number of RNAis are tested in an assay, the false positive rate will be inflated. One issue for these methods is the adjustment of error rates in multiple hypothesis testing [40]. The other issue is whether to perform plate-wise or experiment-wise analysis. The plate-wise analysis can adjust for different systematic errors within each plate. However, it may produce misleading results if a cluster of active siRNAs is located within a single plate. An experiment-wise analysis is not affected by the distribution of active siRNAs between plates; however, it cannot adjust for systematic errors within each plate [27]. Finally, all the above methods of hit selection utilize information from only a negative reference. It remains unresolved whether a negative control or the majority of sample wells should be used as the negative reference to capture information on the variability [41].

3.2 Redundant siRNA Activity (RSA) Analysis

In redundant siRNA activity analysis [8], all RNAis in an assay are initially ranked according to their signals. Then, the rank distribution of all RNAis targeting the same gene is examined and a P-value is assigned. Thus, P-value indicates the statistical significance of all wells targeting a

single gene being unusually distributed toward the top ranking slots, calculated based on an iterative hypergeometric distribution formula. Subsequently, all RNAis are ranked first based on this score, then by their individual activities. Therefore, RNAis clustered toward the top ranks are labeled as active, and the remaining ones are considered negative. The algorithm of RSA is illustrated in Figure 3.1.

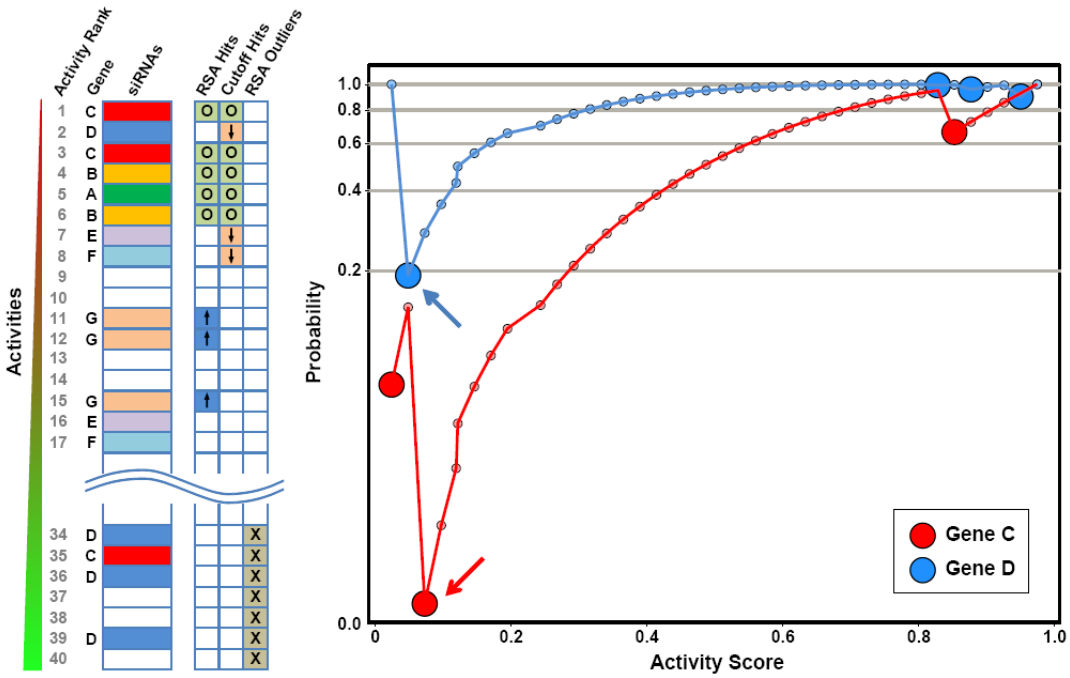

Figure 3.1 Illustration of RSA algorithm [8]

In Figure 3.1, forty RNAis are ranked according to their activities (potent on top) and colored according to their target gene identities. The top eight hits by both RSA and Cutoff algorithms are highlighted, with five common hits marked as "O", RSA-only hits as "↑" and Cutoff-only hits as "↓". RNAis identified as outliers by RSA are marked as "X". After that, interactive RSA P-value is calculated as illustrated by Gene C (3 RNAis) and Gene D (4 RNAis). For a given gene, accumulative hypergeometric P-values are calculated for each RNAi, the curve dips at each RNAi targeting the gene itself (big filled circle). The global minimum is then identified (indicated by arrow) and separate RNAis into two groups: hits and outliers. One and three least potent RNAis are identified as outliers for Gene C and D, respectively. Gene C achieves a global minimum of 0.01, much lower than the 0.2 for Gene D, therefore, the activity distribution of Gene C is much less likely to occur by chance, and therefore the gene is more likely to be confirmed.

3.3 RNAi gene enrichment ranking (RIGER)

RIGER ranks shRNAs according to their differential effects between two classes of samples, and then identifies the genes targeted by the shRNAs at the top of the list. In this way, RIGER identifies genes essential to the difference between the classes. There are three options for to summarize the shRNAs activities targeting the same gene [9].

3.3.1 Second Best Rank

A method based on ranking genes by the rank of the second best scoring hairpin for that gene. This is currently the preferred method for RNAi screen analysis in the RNAi Platform.

3.3.2 Kolmogorov-Smirnov

The empirical distribution function F_n for n iid observation X_i is defined as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x} \quad (3.1)$$

Where $I_{X_i \leq x}$ is the indicator function, equal to 1 if $X_i \leq x$ and equal to 0 otherwise. The Kolmogorov-Smirnov statistic for a given cumulative distribution function $F(x)$ is

$$D_n = \sup_x |F_n(x) - F(x)| \quad (3.2)$$

Where sup x is the supremum of the set of distances. By the Glivenko-Cantelli theorem [42, 43], if the sample comes from distribution F(x), then D_n converges to 0 almost surely. Kolmogorov strengthened this result, by effectively providing the rate of this convergence. The Donsker [44] provides yet a stronger result.

3.3.3 Weighted Sum

This method is a modification of the Second Best Rank in that it takes the combined sum of the first and second best ranks for hairpins for a given gene (http://www.broadinstitute.org/cancer/software/GENE-E/). The best ranking hairpin is given a weight of 0.25 and the second best ranking hairpin is given a weight of 0.75. The sum of these weighted ranks is used to compute a new score, and genes are ranked by this new score.

Regardless of which method to use, users can use either Signal to Noise or Log Fold Change to generate a single hairpin level score from the set of replicates for each hairpin in the RNAi screen. The T-Test option is still experimental and has not been fully tested. Figure 3.2 shows the navigator window for RIGER.

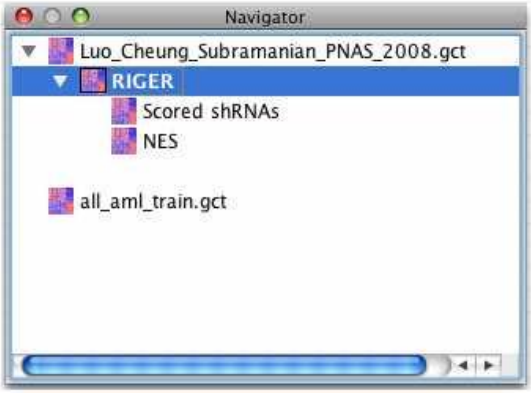

Figure 3.2 RIGER navigator window

3.4 Existing Gene Set Enrichment Analysis Methods

Our RNAi Set Enrichment Analysis framework is based on the gene set enrichment analysis framework. In this section, we briefly review various existing gene set enrichment analysis methods.

In gene expression microarray study, gene set enrichment analysis has been widely applied. Focusing on sets of genes rather than on individual genes has several benefits. From a statistical point of view the analysis of groups instead of individual genes is advantageous as this typically increases power and reduces the dimensionality of the underlying statistical problem [45]. From the biological perspective gene set enrichment analysis allows one to ask and answer questions that are of direct interest to the understanding of the functional mechanism in a cell. Such as "is the pathway more active than other pathways"? These questions directly relate to various null models for gene sets.

A number of statistical procedures to test gene set enrichment have been proposed in the last few years.

Overrepresentation analysis methods were proposed at beginning. Draghici et al. developed Onto-Express (OE) tool to translate the lists of differentially regulated genes into functional profiles characterizing the impact of the condition studied [46]. OE constructs functional profiles (using Gene Ontology terms) for the following categories: biochemical function, biological process, cellular role, cellular component, molecular function, and chromosome location. Statistical significance values are calculated for each category. And many different tools very similar to OE or with minor variation can be found in the published studies [47-50].

However, a drawback to overrepresentation analysis methods is that they rely on the initial gene list in a fundamental way and are sensitive to the choice of both significance criteria and error-control procedure. To address that, average of single gene statistics have been proposed and applied [51-55]. These studies take the average of correlations, p values or other statistics of the genes within each gene set as the get set score. The significance of the gene set score is then accessed through gene resampling or sample lab permutation.

Rather than taking the average of single gene statistics, Subramanian et al. proposed the well-known Gene Set Enrichment Analysis (GSEA) [7, 56]. GSEA tests whether the ranks of the genes in certain gene set differ from a uniform distribution, using a weighted Kolmogorov-Smirnov test. Barry et al. proposed significance analysis of function and expression (SAFE) method. SAFE is very similar to GSEA, except that it adds Wilcoxon rank sum [57] as an option to evaluate the how much the ranks of the genes in certain gene set differ from a uniform distribution. Zahn et al. developed a variant of GSEA [58]. The original GSEA paradigm was intended for datasets with two categories of sample. Zahn et al. replaced the two-sample test statistic in GSEA with an estimated regression slope for age by fitting regression models to continuously varying independent and dependent variables. They also replaced the Kolmogorov-Smirnov statistic with a van der Waerden statistic to reserve the type of dependence that the van der Waerden statistic captures. Finally, they replaced the permutation strategy with a bootstrap in order to better handle covariates. Efron and Tibshirani [59] proposed Gene Set Analysis (GSA) method. 2007).This method has been adopted by the Significance Analysis of

Microarray (SAM) platform (http://www-stat.stanford.edu/~tibs/SAM/). It uses the max-mean statistic to summarize gene-sets, which is the mean of the positive or the negative part of gene scores in the gene set, whichever is larger in absolute value. The genes are re-standardized before the permutation. Based on simulation study, they claim that the resulting test statistic is more powerful than the weighted Kolmogorov-Smirnov statistic used in GSEA. The GSA has been extended to account for a versatile array of data including multi-class, survival, and quantitative outcomes [59]. Keller et al. [60] developed a dynamic programming algorithm for calculating exact significance values of un-weighted GSEA. This algorithm is declared to be able to avoid typical problems of nonparametric permutation tests, as varying findings in different runs caused by the random sampling procedure.

Compared to the non-parametric Kolmogorov-Smirnov test GSEA uses, Kim et al. proposed a modified gene set enrichment analysis strategy PAGE (parametric analysis of gene set enrichment) based on a parametric statistical analysis model. PAGE employs fold change between experimental groups or other parametric data to calculate Z scores of predefined gene sets and use normal distribution to infer statistical significance of gene sets. They declared that PAGE improves analysis of minimally changed gene expression profiles and is statistically more sensitive and required much less computational effort than GSEA. Dinu et al. claim that GSEA has important limitations as a gene-set analysis approach for microarray experiments for identifying biological pathways associated with a binary phenotype [61]. They propose SAM-GS. SAM-GS calculate the statistic $d_i = \frac{\overline{X}_1(i) - \overline{X}_2(i)}{S(i) + S_0}$ for an individual gene analysis, which is first proposed by Tusher et al. [62]. Then the gene set statistic is defined as $\sum_{i=1}^{|S|} d_i^2$. The gene set significance assessment procedure is the same as GSEA.

To address the inefficiency of permutation method adopted by GSEA when few microarrays enter the permutation, and the concerns that the null hypothesis in GSEA permutation refers to the complete absence of differential expression rather than to the absence of enrichment, Newton et al. proposed the random-set method for measuring enrichment [63]. Random-set method adopts category-level statistic like in GSEA, but calibrates them in the same way that Fisher's exact test calibrates the intersection of a functional category and a selected list [63]. It calibrates them conditionally on results of the differential expression analysis by considering values of the category level statistic that would be achieved by a random set of genes.

Different from the above testing structures, Goeman et al. proposed a global test [64]. They use generalized linear models [65] to analyze the dependency of a biological phenotype Y on a measured gene expression X. The testing problem is interpreted in the framework of a random effect model. A score test procedure proposed in [66] is used to derive the test statistic and its asymptotic distribution. Mansmann et al. proposed ANCOVA global test [67]. It is equivalent to Goeman's global test in a setting of independent genes. In the situations of correlated genes, simulation studies show ANCOVA has better performance compared to Goeman's test, especially in cases where the asymptotic distribution cannot be used.

Besides these methods introduced above, there are a lot of other methods proposed [61, 68-74]. For a thorough review on these methods, please refer to [75-80].

Chapter 4 . Data Preprocessing and Quality Check

Our RNAi screen experiment is done with Agilent microarray products. After microarray hybridization, we obtain text files containing the raw probe intensity data. Before we can run the RNAi Set Enrichment Analysis or Structural Equation Modeling analysis which will be discussed in Chapter 5 and 6, we must preprocess the raw data. Originally we did dual color microarray hybridization for RNAi screen, and later we switched to single color hybridization because we found that single color hybridization can save time and cost while having the same or even higher level of signal to noise ratio. These two types have very little difference regarding to the data preprocessing and following data analysis. In this chapter, unless declared, we present the details of data preprocessing and quality check based on raw data from dual color microarray hybridization.

4.1 Preprocessing of the Microarray Raw Data

We have the raw Agilent two-color microarray data for five different cell lines: A549, H522, H460, H322 and J42-L83. For each cell line, Bacterial plasmid DNA is used as common reference and samples come from four different treatment groups: None (48 hours after RNAi infection), mock treatment (MOCK), low dose drug treatment (LOW) and high dose drug treatment (HIGH). Each group has 3 or 4 biological replicates. In the two-color assay, test DNA sample is labeled with cy3, reference DNA sample is labeled with cy5. We analyze the data for the 5 cell lines separately.

Once we obtain the probe intensity files from Agilent's Feature Extraction software, we extract the column “gMeanSignal” as test channel signal, and column “rMeanSignal” as reference channel signal.

The microarray chip is designed for the RNAi library targeting eight sets of cancer genes: HS_Cancer, HS_Kinase, C600, Cellcycle, phosphotase_new, PI3K_cancer, Hemann_BC and roma_cancer. In our experiment, we only infect the cancer cells with RNAis in these three sets: HS_Cancer, HS_Kinase and Cellcycle. So from the raw data, we remove probes in the following sets: C600, phosphotase_new, PI3K_cancer and Hemann_BC. As probes in set roma_cancer have relatively low cross hybridization compared to the other four unused libraries, they are used as negative probes for background estimation.

On the microarray, there are two types of probes: barcode probes (60 mers) and half hairpin probes (21 mers). The structure of RNAi looks like a hairpin. It has two strings (sense string and anti-sense string) and a loop connecting them (see Figure 4.1). In our experiment, we used the RNAi library design from Dr. Gregory Hannon’s lab. In their design, besides the hairpin structure, each RNAi is attached to a barcode sequence as unique identification. The barcode probe binds to the barcode sequence and the half hairpin probe binds to the sense or anti-sense string. Our analysis indicates that their hybridization performances are different to some extent (data now shown here). So for each sample, we divide the data into barcode probe set and half hairpin probe set, and then process them separately during the background filter and normalization steps.

Figure 4.1 RNAi structure

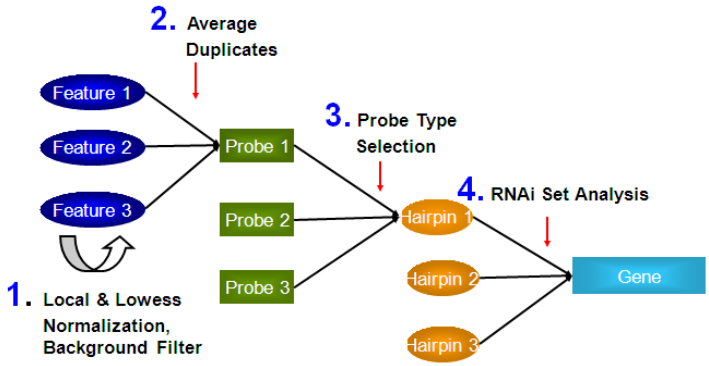

Figure 4.2 Data analysis illustration

Figure 4.2 illustrates the data analysis steps starting from the raw microarray intensity data. The raw microarray data is on the feature level. However, we are interested in which gene can enhance or weaken the drug effect. So our data analysis has to process from the feature level to gene level. Step1 to step3 are called as preprocessing steps. In the following, the details of the three preprocessing steps will be present.

4.1.1 Processing on Feature level

(1) Local Normalization

First we apply spatial correction to remove spatial effects resulting from uneven washing, evaporation edge effect and so on. The spatial correction uses a window of 300 probes, which is illustrated in Figure 4.3. Within each window, the intensity of the inside probes is scaled to make sure the median probe intensity in the window is equal to the median on the whole chip.

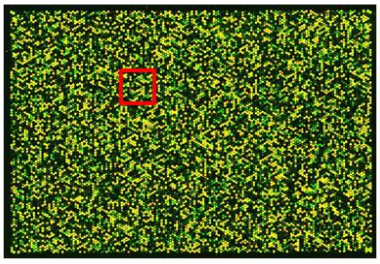

Figure 4.3 Spatial correction illustration

(2) Background Filter

At the time point of 48 hours after RNAi infection, when each RNAi has been integrated into the cancer cell but the target gene expression hasn't been silenced down thus the RNAi should have little effect on the cell proliferation, we believe each probe should have strong enough signal if its representative RNAi infection is successful and its hybridization works well. So we apply background filter to remove probes with very low intensity which we believe are badly designed probes.

The background filter works as follows. First, for each probe set, we calculate the median intensity of background probes in each channel for each sample in the NONE group. These medians are taken to be estimations of the background. Next, we remove probes from the dataset whose intensity is less than 1.5 times of background in red or green channel in more than half of the samples in the NONE group.

Figure 4.4 and Figure 4.5 show the intensity histogram plots of barcodes and half hairpin probes accordingly. The barcode probes are well separated from the background, while the half hairpin probes severely overlap with background. In our experiment, the background filter pass percentage for barcode probes is around 81%, compared to 35% for half hairpin probes. The reason might be due to the hybridization temperature setting which favors barcode probes. Once we lower the hybridization temperature from 65F to 52F, the background pass percentage of half hairpin probes increases to around 45%, while the percentage of barcode probes decreases to around 60%.

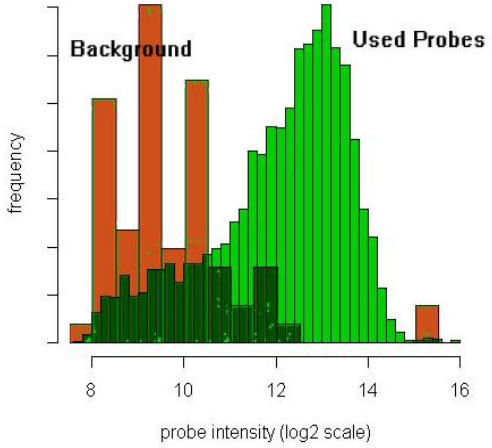

Figure 4.4 Probe intensity histrogram

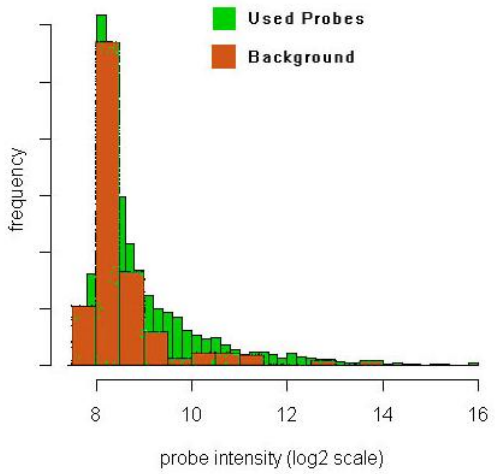

Figure 4.5 Probe intensity histrogram

(3) Normalization

As explained above, we first did dual color microarray hybridization and later switched to single color. The normalization is different between these two. For dual color microarray experiment, to adjust for the imbalance between the red and green dyes, which may arise from labeling or scanning. Figure 4.6 shows the probe intensity histograms of the two channels prior to lowess normalization. Clearly, there exists some difference between these two channels regarding to probe intensity distribution, whereas, the probe intensities in the two channels should follow almost the same distribution in theory. That's because we assume only a small percent of probes will exhibit different expression while the majority remains unchanged.

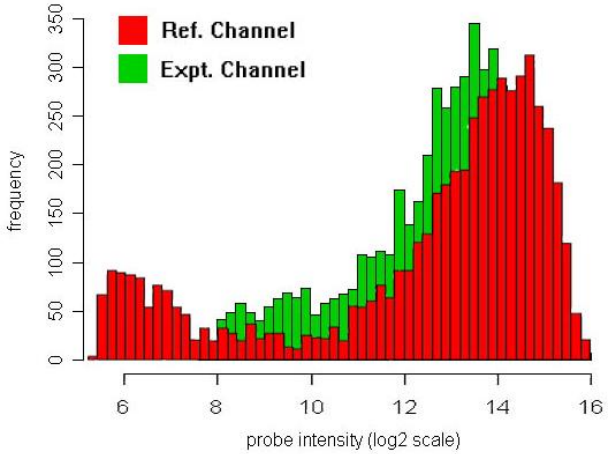

Figure 4.6 Probe intensity histogram

So for each microarray, we apply global loess normalization to remove the imbalance between the red and green channels [81]. This is done with R package limma (version 2.8.1) [82], using the function "normalizeWithinArrays". The parameter "method" in that function is chosen as "loess". After normalization, we obtain a log (base 2) ratio between green channel and red channel for each probe.

Figure 4.7 shows the MA plot before (left) and after (right) lowess normalization. Here A is the mean intensity of the two channels, and M represents the ratio of the intensity between the two channels. It can be seen that, after lowess normalization, the intensity distribution is balanced between red and green channels.

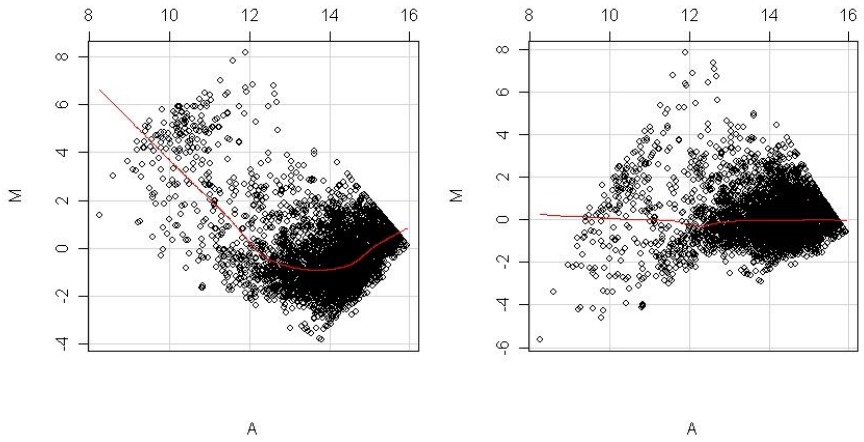

Figure 4.7 MA plot before and after lowess normalization

For single color hybridization microarray, we apply quantile normalization rather than lowess. First, the probe intensity is transferred into log2 scale. Then we perform quantile normalization. Quantile normalization is a technique for making two distributions identical in statistical properties [83]. To quantile-normalize two distribution of the same length, sort the two distributions separately. Then for both distributions, the highest entry takes the mean of the highest values, the second highest value becomes the mean of the second highest values, and so on. Quantile normalization is frequently used in microarray data analysis.

After quantile normalization, we median center the probe intensity to zero for each replicate sample. Figure 4.8 shows the probe intensity boxplot before (left) and after (right) quantile normalization. It can be seen that after quantile normalization, the probe intensity distribution is the same for all arrays. The only difference among these arrays is the rank. A probe might have different rank in different arrays in terms of the intensity value. For instance, a probe might rank top in the first array but rank bottom in the second array. Again, the assumption behind quantile normalization is that the probe intensity distribution remains unchanged on the whole even though some probes may be differently expressed across the arrays.

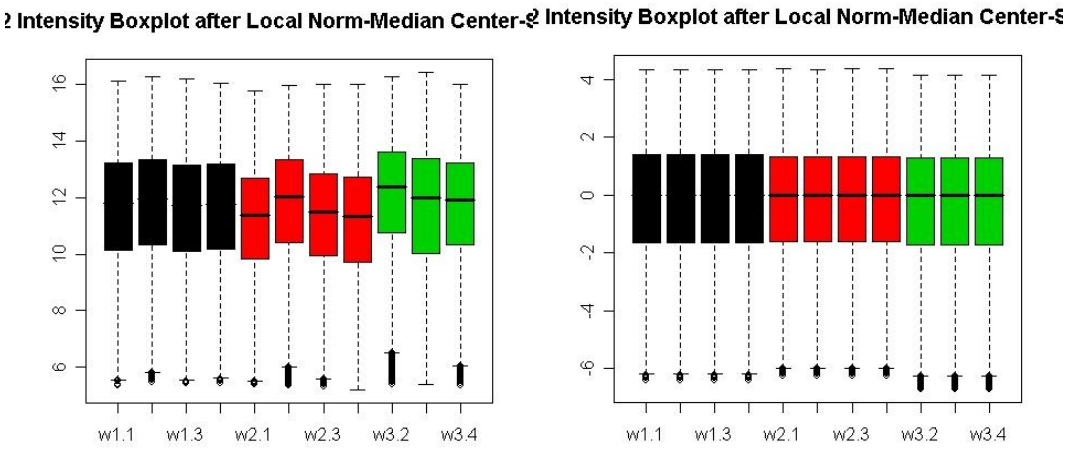

Figure 4.8 Quantile Normalization

4.1.2 From feature to probe level

In the microarray chip design, some probes may be printed in multiple spots (we call "features") on the chip to achieve a more reliable measurement. For these probes, we take the mean of the duplicate features log2 ratios as the probe log2 ratio.

4.1.3 From probe level to RNAi level

As mentioned above, most RNAis have both barcode probe and half hairpin probe printed on the chip. To collapse the data from the probe level into RNAi level, one choice is to take the mean of the probe ratios as the RNAi ratio. But as the two types of probes are different in probe length and nucleotide sequence, they might have different measurement errors and qualities. If we take the average, the good probe might be compromised by the relatively "bad" probe. So our choice is to use the probe with higher quality to represent the RNAi. The RNAi ratio is assigned as the ratio of the higher quality probe. To determine which probe has higher quality, we define an index SNR (signal to noise ratio) as:

$$\text{SNR}=\frac{\textit{between group variability}}{\textit{within group variability}}=\frac{\frac{\sum_{i=1}^{K} n_i(\overline{X}_{i\cdot}-\overline{X})^2}{K-1}}{\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})^2}{N-K}} \tag{4.1}$$

Where $\overline{X}_{i\cdot}$ denotes the sample mean in the i^{th} group indicator, n_i is the number of observations in the i^{th} group, $\overline{X}$ denotes the overall mean of the data, X_{ij} is the j^{th} observation in the i^{th} group out of K groups and N is the overall sample size.

Here the SNR is actually identical to the one-way ANOVA F-test statistic. We assume that the larger the SNR, the higher quality the probe has.

After the above steps, we obtain a 4500 by 9 matrix, each row represents a shRNA, and each column represents a replicate of NONE, MOCK, LOW or HIGH group. We use principal component analysis (PCA) to check the quality of the microarrays. Figure 4.9 show a PCA plot

example for one of our experiments. Black circles represent NONE group, red represents mock treatment group, green represents low dose drug treatment group and blue represents high dose drug treatment group. In this figure, the two drug treatment groups are well separated from the drug free groups in the direction of either the first principal component or the second principal component. This indicates that most of the variance comes from drug treatment, rather than the difference among technical replicates.

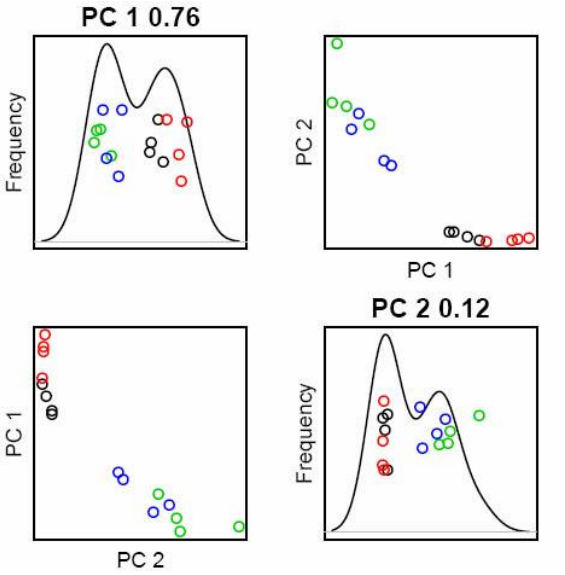

Figure 4.9 PCA plot example

4.2 Data Interpretation

In this section, we link the measurement data to the underlying biological property. Also, we will reveal how to estimate the RNAi drug sensitivity based on the data we have. To simplify the problem, we assume the data comes from the single color microarray, though the solution is the same for dual color microarray.

Let M_{ik} represent the measurement (proportional to cell number) of RNAi i in replicate sample k from the mock treatment group. The model can be written as:

$$\log_2 M_{ik} = \log_2 \beta_i + \varepsilon_{ik} \tag{4.2}$$

where $i = 1, \ldots, n$, n is the number of RNAis in the library. In our experiment, $n \approx 4500$, $k = 1, \ldots, m$, where m is the number of replicate samples in each group, β_i is a number proportional to the number of copies of RNAi i after infection and selection, and ε_{ik} is an error term added to represent the errors introduced from the experimental procedures. We assume the error term follows normal distribution $N(0, \sigma^2)$.

For the high dose drug treatment group, let H_{ik} represent measurement of RNAi i in replicate sample k from the high dose drug treatment group. It can be written as:

$$\log_2 H_{ik} = \log_2 \left(0.8\beta_i \times 2^{f_i} \frac{1}{0.8}\right) + \varepsilon_{ik} = \log_2 \beta_i + f_i + \delta_{ik} \tag{4.3}$$

Where f_i is the gene silencing effect of RNAi i on drug response. $f_i > 0$ means RNAi i is drug resistant and prevents the drug to kill cancer cells in high dose drug condition, and $f_i < 0$ means RNAi i is drug sensitive and help the drug to kill cancer cells in high dose drug condition. We can note that a normalization factor 1/0.8 is introduced in the above equation to balance the overall intensity of each sample and simplify the analysis afterwards. $\delta_{ik} \sim N(0, \sigma^2)$

As we focus on the comparison between HIGH and DMSO in this thesis, we don't present the model for low dose treatment group here. However, it's very similar to the high dose drug treatment.

By comparing high drug treatment group with mock treatment group, we can estimate the drug sensitivity f_i.

4.3 Cluster Analysis

Each RNAi is designed to silence down its target gene. However, different RNAis have different silencing efficiency even they might target the same gene. For example, one RNAi might reduce the Ras gene expression by 90% while another might reduce only 40%. In addition, almost every RNAi has off target effect, silencing down some non-specific genes besides its target gene [84]. Due to the silencing efficiency and off target effect problems, different RNAis targeting the same gene may show totally different drug responses. For example, for two RNAis designed to target the same gene, one might be tested as drug sensitive while the other as drug sensitive. This will muddle the real drug sensitivity of the target gene. Facing this challenge, we tried the cluster analysis to classify RNAis targeting the same gene. If for one gene, most of its RNAis fall into one specific cluster, then we will have more confidence to use this cluster to represents that gene.

4.3.1 Cluster method

Considering the five cell lines screened in our experiment have different drug sensitivity, we clustered RNAis in those cell lines separately. Before running the cluster analysis, we average the measurement of replicate samples in each group to remove the influence of batch effect on clustering. The cluster method we chose is K-Means clustering where K is set to 27. The Euclidean distance is chosen as distance measurement. Theoretically, there are 13 unique patterns when we only have three groups or conditions (see Figure 4.11). The reason why we choose K>13 is that we do not wish to miss any important pattern. After obtaining the 27 clusters, we assign each cluster into its corresponding pattern. For example, Figure 4.10 shows a cluster in cell line J42-L83. The left shows all RNAis measurement data in that cluster. And the right shows the averaged measurement in three experimental conditions. Based on this figure, this cluster clearly belongs to pattern #9. RNAis in this pattern are neutral in low dose drug condition, but resistant in high dose drug condition.

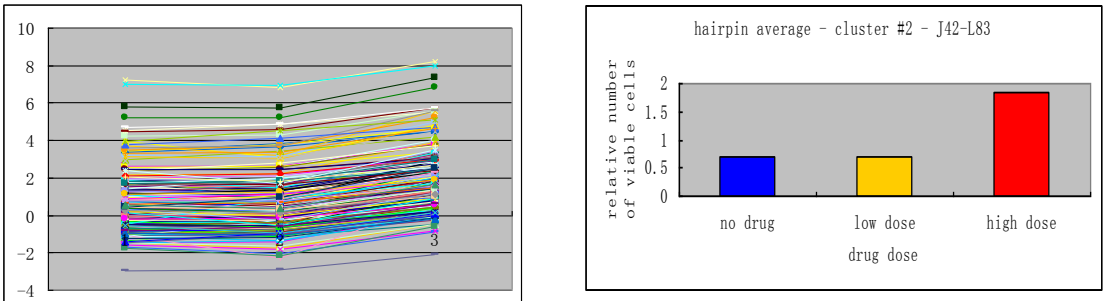

Figure 4.10 A cluster in J42-L83

4.3.2 Cluster Analysis Result Summary

Figure 4.11 shows the summary of the K-Means for 5 different cell lines. Obviously, patterns #10~13 can be neglected because of the very low representation, as expected. Pattern #7 is the most frequent pattern. This is also what we have expected -- that most RNAis would be neutral to the drug.

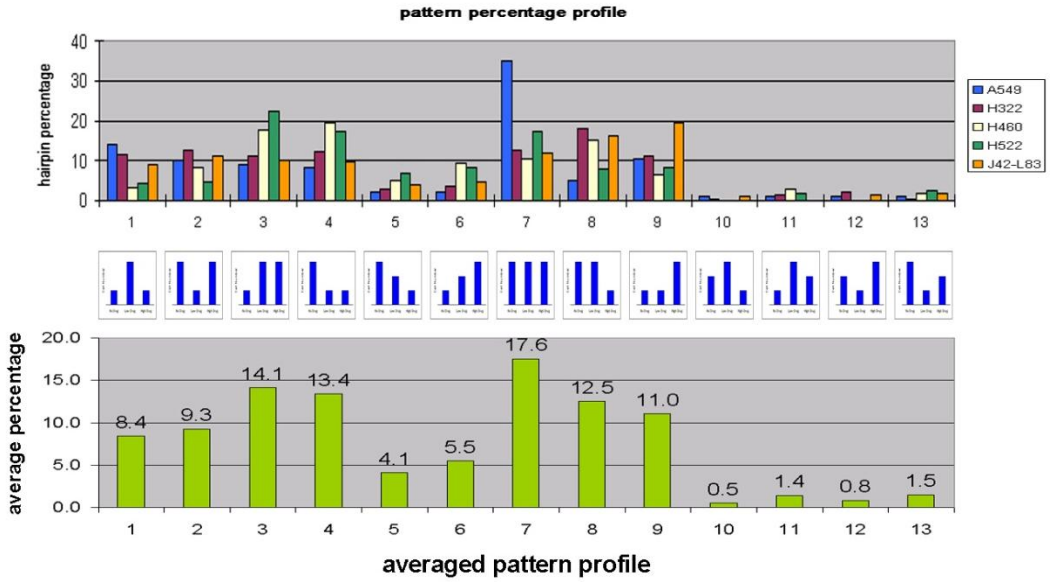

Figure 4.11 Pattern percentage profile in 5 cell lines

Chapter 5 . RNAi Set Enrichment Analysis

Gene set enrichment methods have been used successfully for gene expression microarray data. However, very few of them have been used for the analysis of large scale RNAi screens. Similar to the common modular framework for gene set enrichment analysis methods depicted in Ackermann's paper [45], we propose the general framework for RNAi Set Enrichment Analysis as Fig5.1. The scheme consists of two distinct ways to analyze RNAi screen data: Structural Equation Modeling (SEM) and RNAi set enrichment analysis. The details of the three SEM models will be discussed in next chapter. In this chapter, we focus on the RNAi Set Enrichment Analysis (RSEA). RSEA consists of three modules: the calculation of an RNAi level statistic, the computation of an RNAi set statistic and the significance assessment of the RNAi set statistic. In gene expression data analysis, multiple options have been proposed for each module. For our RNAi screen, we are going to evaluate and compare the statistical performance of these options in each module. First, we are going to give a brief introduction of these modules and the different options within each module.

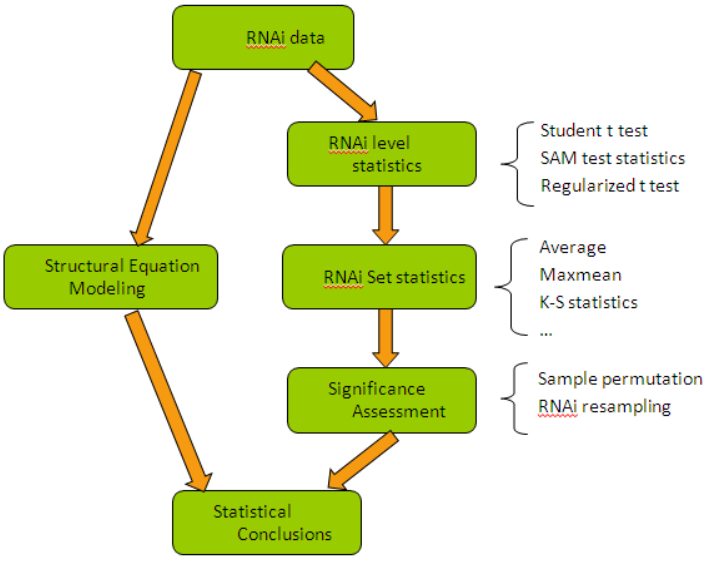

Figure 5.1 General framework for RNAi screen data analysis

5.1 RNAi Level Statistic

Similar to gene expression microarray study, in RNAi screen, we need some statistic to rank RNAis according to their activity. In the last few years, various statistics have been suggested, which may be classified as follows [85]:

(1) Simple methods: such as fold change and classical student's t statistic.

(2) The SAM ('significance analysis of microarrays') t statistic [86].

(3) (Penalized) likelihood methods. Please refer to [87-89] for q thorough review.

(4) Hierarchical Bayes methods, e.g.: [90-93] and "moderated t"[94-96].

For an introductory review of other approaches mentioned above, please see refer to Cui and Churchill [97] and Smyth [94].

Current good practice in gene expression case-control analysis favors the empirical or full Bayesian approaches over other competing methods [85]. The reason is that Bayesian methods naturally allow for information sharing across genes, which is essential when the number of sample is as small in typical genomic experiments. Specifically, the estimation of gene-specific variances profits substantially from pooling information across genes. On the other hand, Bayesian methods can become computationally quite expensive, and more importantly, typically rely on a host of very detailed assumptions concerning the underlying data and parameter generating models [85].

In our RNAi screen data analysis, we evaluated three options: the classical student's t statistic, SAM t statistic [86] and regularized t statistic [91]. A brief introduction of these statistics is given below:

5.1.1 Student's t statistic

In our experiment, we usually compare two classes: control and drug treatment. And we assume equal variance between these two classes. So we use unequal sample – equal variance t statistic in our RNAi screen:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{5.1}$$

Where

$$S_{X_1X_2} = \sqrt{\frac{(n_1-1)S_{X_1}^2 + (n_2-1)S_{X_2}^2}{n_1+n_2-2}} \tag{5.2}$$

$S_{X_1}^2$ is the estimated variance for group 1 and $S_{X_2}^2$ is the estimated variance for group. n_1 is the sample size for group 1 and n_2 is the sample size for group 2.

5.1.2 SAM t statistic

Compared to classical t statistic, SAM t statistic adds a constant S_0 to stabilize the estimation for standard deviation [62]:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S + S_0} \tag{5.3}$$

where $S = S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, S_0 is chosen as to minimize the coefficient of variation. For simplicity, in most case, S_0 is chosen as the 90th percentile of S of all RNAis.

5.1.3 Regularized t statistic

Baldi et al. developed a general Bayesian statistical framework for array data [91]. Here we adapt their method as follows:

We model the corresponding measurements of each RNAi in each situation (treatment or control) with a normal distribution $N(x;\mu,\sigma^2)$. For each RNAi and each condition, we have a two parameter model $w = (\mu,\sigma^2)$. Assuming that the observations are independent, the likelihood of the data D is given by:

$$P(D|\mu,\sigma^2) \approx \prod_{i=1}^{n} N(x_i;\ \mu,\sigma^2)$$
$$= C(\sigma^2)^{-n/2} e^{-\sum_i (x_i-\mu)^2/2\sigma^2}$$
$$= C(\sigma^2)^{-n/2} e^{-(n(m-\mu)^2+(n-1)S^2)/2\sigma^2} \quad (5.4)$$

Here C denotes the normalizing constant of the distribution, n is the sample size and $m = \overline{x}$. All the information about the sample that is relevant for the likelihood is summarized in the sufficient statistics n, m, and S^2.

A full Bayesian treatment requires introducing a prior distribution $P(\mu,\sigma^2)$. Several kinds of priors for the mean and variance of a normal distribution have been studied in the literature, including the non-informative improper prior and the conjugate prior [98, 99],. For our RNAi screen data, we choose the conjugate prior because of its convenient form. The form of the likelihood in Equation (5.4) shows that the conjugate prior density must also have the form $P(\mu|\sigma^2)P(\sigma^2)$, where the marginal $P(\sigma^2)$ is scaled inverse gamma and the conditional distribution $P(\mu|\sigma^2)$ is normal. This leads to a hierarchical model with a vector of four hyper parameters for the prior $\alpha = (\mu_0,\lambda_0,\upsilon_0,\sigma_0^2)$ with the densities:

$$P(\mu|\sigma^2) = N(\mu;\ \mu_0,\frac{\sigma^2}{\lambda_0}) \quad (5.5)$$
$$\text{And } P(\sigma^2) = I(\sigma^2;\ \upsilon_0,\ \sigma_0^2)\ . \quad (5.6)$$

The expectation of the prior is finite if and only if $\upsilon_0 > 2$.

The prior $P(\mu,\sigma^2) = P(\mu,\sigma^2|\alpha)$is given by:

$$C\sigma^{-1}(\sigma^2)^{-(\frac{\upsilon_0}{2}+1)} exp\left[-\frac{\upsilon_0}{2\sigma^2}\sigma_0^2 - \frac{\lambda_0}{2\sigma^2}(\mu_0-\mu)^2\right]. \quad (5.7)$$

The hyper parameters μ_0 and σ^2/λ_0 can be interpreted as the location and scale of μ, and the hyper parameters υ_0 and σ_0^2 as the degrees of freedom and scale of σ^2. Applying Bayes theorem, the posterior has the same functional form as the prior

$$P(\mu,\sigma^2|D,\alpha) = N(\mu;\ \mu_n,\sigma^2/\lambda_n)I(\sigma^2;\ \upsilon_n,\ \sigma_n^2) \quad (5.8)$$

With

$$\mu_n = \frac{\lambda_0}{\lambda_0+n}\mu_0 + \frac{n}{\lambda_0+n}m \quad (5.9)$$

$$\lambda_n = \lambda_0 + n \tag{5.10}$$

$$\upsilon_n = \upsilon_0 + n \tag{5.11}$$

$$\upsilon_n \sigma_n^2 = \upsilon_0 \sigma_0^2 + (n-1)s^2 + \frac{\lambda_0 n}{\lambda_0 + n}(m - \mu_0)^2 \tag{5.12}$$

While it is possible to use a prior mean μ_0 for gene expression data, in many situations it is sufficient to set $\mu_0 = m$. It can readily be shown that the conditional posterior distribution $P(\mu|\sigma^2, D, \alpha)$ of the mean is normal $N(\mu_n, \sigma^2/\lambda_n)$, the marginal posterior P(μ|D, α) of the mean is Student's $t(\upsilon_n, \mu_n, \sigma_n^2/\lambda_n)$, and the marginal posterior $P(\sigma^2|D, \alpha)$ of the variance is scaled inverse gamma $I(\upsilon_n, \sigma_n^2)$.

The posterior distribution $P(\mu, \sigma^2|D, \alpha)$ is the fundamental object of Bayesian analysis and contains the relevant information about all possible values of μ and σ^2. However, in the regularized t statistic in our RSEA framework, we take the mean of the posterior (MP) estimate as the single point estimate. By integration, the MP estimate is given by

$$\mu = \mu_n \text{ and } \sigma^2 = \frac{\upsilon_n}{\upsilon_n - 2}\sigma_n^2 \tag{5.13}$$

provided $\upsilon_n > 2$. If we take $\mu_0 = m$, we then get the following MP estimate:

$$\mu = m \text{ and } \sigma^2 = \frac{\upsilon_n \sigma_n^2}{\upsilon_n - 2} = \frac{\upsilon_0 \sigma_0^2 + (n-1)s^2}{\upsilon_0 + n - 2} \tag{5.14}$$

provided $\upsilon_0 + n > 2$. Based on the estimation in Equation (5.14), the regularized t statistic can be written as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{5.15}$$

Where

$$S_{X_1X_2} = \sqrt{\frac{(n_1-1)\sigma_{X_1}^2 + (n_2-1)\sigma_{X_2}^2}{n_1 + n_2 - 2}} \tag{5.16}$$

$$\sigma_{X_1} = \sqrt{\frac{\upsilon_0 \sigma_0^2 + (n_1-1)S_{X_1}^2}{\upsilon_0 + n_1 - 2}} \tag{5.17}$$

$$\sigma_{X_2} = \sqrt{\frac{\upsilon_0 \sigma_0^2 + (n_2-1)S_{X_2}^2}{\upsilon_0 + n_2 - 2}} \tag{5.18}$$

In fact, the regularized t statistic is a modification of the two sample equal variance student's t statistic, with the sample standard deviation replaced by Bayes MP estimate.

5.2 RNAi Set Statistic

A further step in an enrichment analysis is the computation of the RNAi set statistic. In gene set enrichment analysis, a lot of options have been proposed, such as [45]:

• the sum, mean or the median of the single RNAi statistics,

• the Kolmogorov-Smirnov statistic [7],

• the max-mean statistic [100], and

• the Wilcoxon rank sum test statistic.

The question of which statistic is optimal is subject to ongoing discussion. Efron and Tibshirani show that their max-mean statistic is more powerful than Kolmogorov-Smirnov test [100]. Jiang and Gentleman emphasized the problem of robustness against outliers and advocate summaries such as the median or the sign test statistic [54]. In our study, we chose and evaluated three kinds of RNAi set statistics: the mean of the single RNAi statistics, the max-mean statistic and the Kolmogorov-Smirnov statistic. These statistics are introduced below.

5.2.1 The mean of the single RNAi statistics:

It is the arithmetic mean value of the RNAi statistics of all the RNAis in the gene set.

5.2.2 Kolmogorov-Smirnov statistic

We take the data analysis procedure in GSEA as an example to describe the Kolmogorov-Smirnov statistic [7]. Suppose the genome wide expression profiles from samples belong to two classes, labeled 1 or 2. Genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric (see Figure 5.2).

Given a defined set of genes S, the goal of GSEA is to determine whether the members of S are randomly distributed throughout L, the ranked list of genes according to their differential expression between the classes, or primarily found at the top or the bottom. GSEA expects that sets related to the phenotypic distinction will tend to show the latter distribution. There are three key steps to the GSEA method [7]:

Step 1: Calculation of an Enrichment Score. First calculate an enrichment score (ES) that reflects the degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked list L. The score is calculated by walking down the list L, increasing a running-sum statistic when we encounter a gene in S and decreasing it when we encounter genes not in S. The magnitude of the increment depends on the correlation of the gene with the phenotype. The enrichment score is the maximum deviation from zero encountered in the random walk; it corresponds to a weighted Kolmogorov–Smirnov-like statistic [101] (see Figure 5.2).

Step 2: Estimation of Significance Level of ES. We estimate the statistical significance (nominal P value) of the ES by using an empirical phenotype-based permutation test procedure that preserves the complex correlation structure of the gene expression data. Specifically, we permute the phenotype labels and re-compute the ES of the gene set for the permuted data, which generates a null distribution for the ES. The empirical, nominal P value of the observed ES is then calculated relative to this null distribution. Importantly, the permutation of class labels preserves gene-gene correlations and, thus, provides a biologically more reasonable assessment of significance than would be obtained by permuting genes.

Step 3: Adjustment for Multiple Hypothesis Testing. When an entire database of gene sets is evaluated, we adjust the estimated significance level to account for multiple hypothesis testing. We first normalize the ES for each gene set to account for the size of the set, yielding a normalized enrichment score (NES). We then control the proportion of false positives by calculating the false discovery rate (FDR) [102, 103] corresponding to each NES. The FDR is the estimated probability that a set with a given NES represents a false positive finding; it is computed by comparing the tails of the observed and null distributions for the NES.

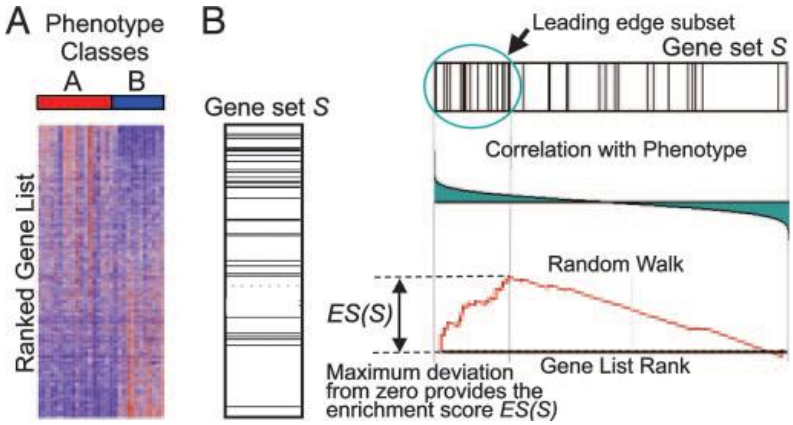

Figure 5.2 A GSEA overview illustrating the method [7]

5.2.3 Max-mean statistic

Max-mean statistic is implemented in GSA as below [100]:

1. Compute a summary statistic z_i for each gene, such as the two sample t-statistic for two-class data. Let $\mathbf{z}_S$ be the vector of z_i values for genes in a gene-set S.

2. For each gene-set S, choose a summary statistic $S = s(z)$: choices include the average of z_i or $|z_i|$ for genes in S, the max-mean statistic defined as $S_{max} = max\left\{\overline{S}_S^{(+)}, \overline{S}_S^{(-)}\right\}$, where $\overline{S}_S^{(+)}(z) = max\{z, 0\}$ and $\overline{S}_S^{(-)}(z) = -min\{z, 0\}$.

3. Standardize S by its randomization mean and standard deviation as $S' = (S - mean_s)/\sigma_s$.

4. Compute permutations of the outcome values (e.g. the class labels in the two-class case) and re-compute S' on each permuted dataset, yielding permutation values $S'^{*1}, S'^{*2}, \ldots, S'^{*B}$, where B is the total number of permutations.

We use these permutation values to estimate P-values for each gene-set score S', and false discovery rates applied to these P-values for the collection of gene-set scores.

5.3 Significance Assessment

The last step of RSEA is the assessment of significance of the observed RNAi set statistic. The calculation of the P-value can be done in two different ways:

5.3.1 RNAi resampling:

A large number of random RNAi sets of the same size as the set under investigation is drawn from all the RNAis and the RNAi set statistic is recomputed for every random set. The P-value is calculated as the fraction of resampled RNAi set statistics that exceed (or fall below) the observed value.

5.3.2 Sample label permutation:

The phenotypes of the subjects are permuted a large number of times and the RNAi level and RNAi set statistics are recomputed. The P-value is the fraction of permutation gene set statistics that exceed (or fall below) the observed value.

RNAi resampling implicitly assumes independent RNAis in the group, a prerequisite that is unlikely to hold, because they are designed to target the same gene. However, for sample label permutation, when the sample size is small, there might be not enough unique permutations to approximate the null distribution. So the choice of RNAi resampling or sample label permutation really depends on the experiment design. In this thesis, we evaluate these two options regarding to the statistical performance for different sample size and different redundancy in the RNAi library design. The evaluation results are presented in chapter 7.

Chapter 6 . Structural Equation Modeling

Structural Equation Modeling (SEM) can be applied to test and estimate causal relations using a combination of statistical data and qualitative causal assumptions. The definition of SEM was articulated by the geneticist Sewall Wright [104] and the cognitive scientist Herbert Simon [105], and formally defined by Judea Pearl using a calculus of counterfactuals [106].

SEM allows both confirmatory and exploratory modeling. SEM has the ability to construct latent variables: variables which are not measured directly, but are estimated in the model from measured variables. This allows the modeler to explicitly capture the unreliability of measurement in the model, which in theory allows the structural relations between latent variables to be accurately estimated. Factor analysis, path analysis and regression all represent special cases of SEM.

6.1 SEM Model

We propose the latent variable SEM for RNAi screen data analysis as in Figure 6.1.

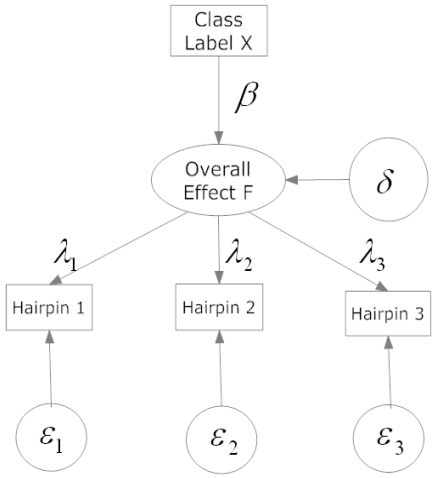

Figure 6.1 Network diagram for SEM

The above diagram can be represented by Equation (6.1):

$$\begin{cases} Y_1 = \mu_1 + \lambda_1 F + \varepsilon_1 \\ Y_2 = \mu_2 + \lambda_2 F + \varepsilon_2 \\ \quad\quad \vdots \\ Y_k = \mu_k + \lambda_k F + \varepsilon_k \\ F = \beta X + \delta \end{cases} \tag{6.1}$$

Where:

$Y_1 \sim Y_k$: measurements of RNAi 1~k.

$\mu_1 \sim \mu_k$: constants related to the initial number of copies of RNAi 1~k.

$\lambda_1 \sim \lambda_k$: coefficient of RNAi 1~k. This coefficient combines the RNAi silencing efficiency and RNAi off target effect into one single parameter.

$\varepsilon_1 \sim \varepsilon_k$: measurement errors of RNAi 1~k.

F: overall effect variable.

β: drug sensitivity of the gene.

X: class label. 0 represents control group and 1 represents drug treatment group.

δ: drug sensitivity variation across samples.

By setting different constraints on the RNAi coefficient and measurement error terms, Equation (6.1) can be turned into three models:

(1) Model A:

$\lambda_1 = 1.$

(2) Model B:

$\lambda_1 = \lambda_2 = \cdots \lambda_k = 1.$

This model is equivalent to multivariate analysis approach for repeated measures ANOVA. Theoretical proof is provided by Professor Wei Zhu in the following section.

(3) Model C:

$\lambda_1 = \lambda_2 = \cdots \lambda_k = 1 \; and \;\; \sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2, \; (i = 1, \cdots, m).$

This model is equivalent to the univariate analysis approach for repeated measures ANOVA. Refer to the following section for theoretical proof by Wu and Zhu [107]

6.2 RM ANOVA is a Special Case of Latent Variable SEM

This section is based on the theoretical development from our own laboratory (Wu and Zhu, 2010).

6.2.1 Latent variables in SM

Considering the situation by modeling one latent variable with measured variables as: $y = \Lambda_y \eta + \varepsilon$, i.e.

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_m \end{bmatrix} \cdot \eta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_m \end{bmatrix} \tag{6.2}$$

where $\underline{y} = (y_1,\ y_2, \cdots,\ y_m)'$ stands for m repeated measures or indicators for the latent variable η. $\Lambda_y = (\lambda_1,\ \lambda_2, \cdots,\ \lambda_m)'$ is the path coefficient matrix of indicators, and ε is the error matrix of indicators. For the model, we have the assumptions that $\varepsilon \sim N_m(0, \Theta_\varepsilon)$, $Var(\eta) = \sigma_\eta^2$, $Cov(\varepsilon_i, \eta_j) = 0$, where i, j = 1, 2, …, m (Figure 6.2).

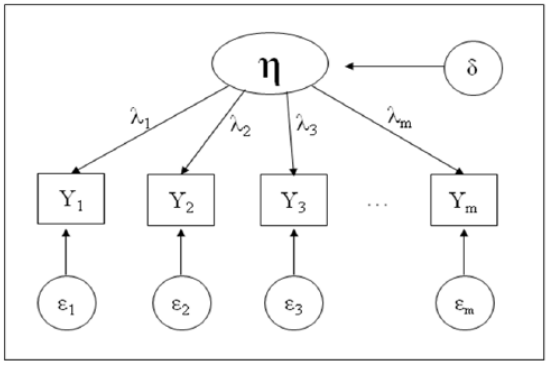

Figure 6.2 One latent variable with m measurements

Thus, the covariance matrix of y is

$$\begin{aligned}
&\Sigma_{yy} = E(yy') = E\left[(\Lambda_y\eta + \varepsilon)(\Lambda_y\eta + \varepsilon)'\right] \\
&= \Lambda_y E(\eta\eta')\Lambda_y{}' + \Theta_\varepsilon \\
&= \Lambda_y \sigma_\eta^2 \Lambda_y{}' + \Theta_\varepsilon \\
&= \sigma_\eta^2 \begin{bmatrix} \lambda_1^2 & \lambda_1\lambda_2 & \cdots & \lambda_1\lambda_m \\ \lambda_1\lambda_1 & \lambda_2^2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \lambda_m\lambda_1 & \lambda_m\lambda_2 & \cdots & \lambda_m^2 \end{bmatrix} + \begin{bmatrix} \sigma_{\varepsilon_1}^2 & & \cdots & \\ & \sigma_{\varepsilon_2}^2 & \cdots & \\ \cdots & \cdots & \cdots & \cdots \\ & & \cdots & \sigma_{\varepsilon_m}^2 \end{bmatrix} \\
&= \begin{bmatrix} \sigma_\eta^2\lambda_1^2 + \sigma_{\varepsilon_1}^2 & \sigma_\eta^2\lambda_1\lambda_2 & \cdots & \sigma_\eta^2\lambda_1\lambda_m \\ \sigma_\eta^2\lambda_1\lambda_1 & \sigma_\eta^2\lambda_2^2 + \sigma_{\varepsilon_2}^2 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_\eta^2\lambda_m\lambda_1 & \sigma_\eta^2\lambda_m\lambda_2 & \cdots & \sigma_\eta^2\lambda_m^2 + \sigma_{\varepsilon_m}^2 \end{bmatrix}
\end{aligned} \tag{6.3}$$

6.2.2 Repeated measures ANOVA

The univariate repeated measures ANOVA model is:

$$Y_{ij} = \mu_j + S_i + \varepsilon_{ij}, \tag{6.4}$$

where μ_j is the (fixed) effect of treatment j, S_i is the (random) effect of subject i, ε_{ij} is the random error independent of i S . With normality assumptions, we have

$S_i \overset{iid}{\sim} N(0, \sigma_s^2)$ and $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$. Let $\underline{Y_i} = (Y_{i1}, Y_{i2}, \cdots, Y_{im})'$, we have $\underline{Y_i} \overset{iid}{\sim} N_m(\underline{\mu}, \Sigma)$, $i = 1, \cdots, n$, where

$\underline{\mu} = (\mu_1, \mu_2, \cdots, \mu_m)'$ and

$$\Sigma = \begin{bmatrix} \sigma_s^2 + \sigma_\varepsilon^2 & \sigma_s^2 & \cdots & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 + \sigma_\varepsilon^2 & \cdots & \sigma_s^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_s^2 & \sigma_s^2 & \cdots & \sigma_s^2 + \sigma_\varepsilon^2 \end{bmatrix} \tag{6.5}$$

This particular structure of the variance covariance matrix is called "compound symmetry".

This compound symmetry matrix form in repeated measures ANOVA can be obtained for latent variable SEM by having constraints of equal path coefficients and equal error variance in (*), i.e. $\lambda_i = 1,\ \sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2,\ (i = 1, \cdots, m)$.

The covariance structure is reduced to the compound symmetry form as below.

$$\Sigma = \begin{bmatrix} \sigma_\eta^2 + \sigma_\varepsilon^2 & \sigma_\eta^2 & \cdots & \sigma_\eta^2 \\ \sigma_\eta^2 & \sigma_\eta^2 + \sigma_\varepsilon^2 & \cdots & \sigma_\eta^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\eta^2 & \sigma_\eta^2 & \cdots & \sigma_\eta^2 + \sigma_\varepsilon^2 \end{bmatrix} \tag{6.6}$$

Alternatively for repeated measures ANOVA, we can use the multivariate approach where no structure, other than the usual symmetry and non-negative definite properties, is imposed on the variance covariance matrix Σ in $\underline{Y_i} \overset{iid}{\sim} N_m(\underline{\mu}, \Sigma),\ i = 1, \cdots, n$. Thus the multivariate approach has no requirement on equal error variance, which is equivalent to latent variable SEM when only restriction of equal path coefficients is applied, that is,

$$\lambda_i = 1, (i = 1, \cdots, m). \tag{6.7}$$

The corresponding covariance matrix is:

$$\Sigma = \begin{bmatrix} \sigma_\eta^2 + \sigma_{\varepsilon_1}^2 & \sigma_\eta^2 & \cdots & \sigma_\eta^2 \\ \sigma_\eta^2 & \sigma_\eta^2 + \sigma_{\varepsilon_2}^2 & \cdots & \sigma_\eta^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\eta^2 & \sigma_\eta^2 & \cdots & \sigma_\eta^2 + \sigma_{\varepsilon_m}^2 \end{bmatrix} \tag{6.8}$$

In summary, we have shown from the covariance structure that the repeated measures ANOVA in the univariate analysis approach, is a special case of latent variable SEM when the error variances are equal and the path coefficients are all set to 1. In addition, the repeated measure ANOVA in the multivariate analysis approach is also a special case of latent variable SEM when all the indicator path coefficients are set to be one.

Chapter 7 . Simulation Study

As described in previous chapters, we propose RSEA and SEM for RNAi screen data analysis. Before we adapt them in biological study, their statistical performances have to be evaluated first. As there are very little published RNAi screen data, in this chapter, we access the specificity and sensitivity performance of RSEA and SEM on simulated data.

To make sure the simulated data is as close as to the real RNAi screen data, we develop a model for the data simulation. This model takes most characters of the RNAi screen experiment into consideration.

7.1 Data Simulation Model

Fig.2 illustrates the model used to generate data. We consider the situation of two classes: control class and treatment class. For the control group, we assume the RNAis follow $N(0,\sigma_0^2)$ distribution. σ_0^2 denotes the variance of measurement error. For treatment groups, the RNAi might work (with probability p) or might not work (with probability 1-p).

In the treatment group, if the RNAi works, then its total effect on the drug response will be the silencing efficiency times the gene sensitivity plus off target effect. So the RNAi follows distribution $N(\lambda\beta+\delta,\sigma_0^2)$. λ denotes silencing efficiency, which is assumed to follow uniform distribution U(0,1). β denotes the gene sensitivity or effect on drug response of cancer cells.

If the RNAi doesn't work in the treatment, then we treat it as in the control group. It follows the same distribution as control group: $N(0,\sigma_0^2)$

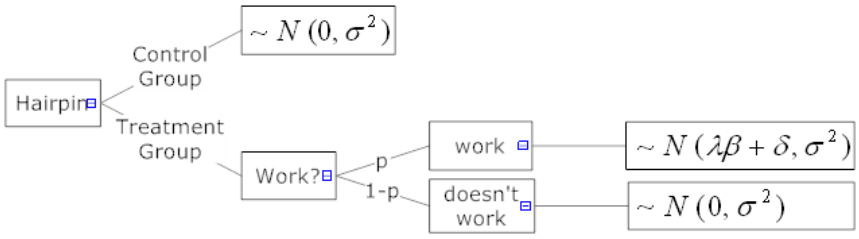

Figure 7.1 Illustration for data simulation

More details about the data simulation are described as follows. Let X_{ij} denotes the measurement value of RNAi i in replicate j of control group. Based on Figure 7.1, X_{ij} can be written as:

$$X_{ij}=\varepsilon_{ij},\qquad \varepsilon_{ij}\sim N(0,\sigma_0^2) \tag{7.1}$$

where σ_0^2 denotes the variance of measurement error. Unless declared explicitly, we set $\sigma_0^2=1$ in all following simulation studies in this chapter.

Let Y_{ij} denotes the measurement value of RNAi i in replicate j of drug treatment group. Based on Figure 7.1, Y_{ij} can be written as:

$$Y_{ij} = I_i(\lambda_i\beta + \delta_i) + \varepsilon_{ij} \quad (7.2)$$

Where I_i denotes RNAi working index, following Bernoulli distribution with $p = p_0$. Unless declared explicitly, we set $p_0 = 0.8$ in all following simulation studies in this chapter. λ_i denotes RNAi silencing efficiency, following uniform distribution U(0,1). β_i denotes the gene sensitivity or effect on drug response of cancer cells. Unless declared explicitly, we set $\beta_i = 3$ in all simulation studies in this chapter.

δ_i denotes the RNAi off target effect, following normal distribution $N(0, \sigma_{OTE}^2)$. σ_{OTE} denotes the standard deviation of off target effect. Unless declared explicitly, we set $\sigma_{OTE} = 0.2$ in all following simulation studies in this chapter.

7.2 RSEA Simulation study

In this section, we evaluated the performances of different statistic options in RSEA regarding to their power and type I error rate performances when drug treatment group is compared to control group. For the power study, the significance level is chosen as 5% in all simulation studies in this chapter.

7.2.1 RNAi Level Statistic Comparison

For RNAi level statistic, we implemented three options: student's t, regularized t and SAM t. We compared their power and type I error rate curves in different situations based on simulated data. The situations tested include the four combinations of varying library redundancy or sample size, and mean or max-mean as RNAi set statistic. The comparison results are present below.

(1) RNAi redundancy as independent variable

We fist compare the performance of the three RNAi level statistics under different RNAi redundancy (number of RNAis targeting each gene). The parameters used for this comparison are listed in Table 7.1. In each simulation run, we simulated RNAis targeting 2000 genes in total. 25% of these genes are drug resistant, 25% are drug sensitive and the left 50% are neutral. To obtain more accurate estimate, we run the simulation five times and estimate the power and type I error rate estimation as the average of the 5 runs. In all simulation results in this chapter, "power" actually refers to the percentage of genes detected as significant with 5% as the p value cutoff when each gene is assigned with a certain amount of effect on drug response. The "type I error rate" actually refers to the percentage of genes detected as significant with 5% as the p value cutoff when each gene has no effect on drug response.

Table 7.1 Parameters used for data generation.

Parameter	Description	Value
SimulationRun	number of simulation runs	5
RNAi level statistic	-	Bayes t
RNAi set statistic	-	Mean
Significance assessment	-	permutation
Perm/Resample Num	number of permutation or resampling times	1000
SensitiveGenePct	percentage of drug sensitive genes	25%
ResistantGenePct	percentage of drug resistant genes	25%
Sample size	number of replicates in each group	10
TotalGeneNum	total number of genes in each simulation run	2000

The power simulation results are listed in Table 7.2 and plotted in Figure 7.2. They indicate that SAM t has the highest power than the others even though the power difference between SAM and regularized t is small (around 1%).

Table 7.2 Statistical power comparison among RNAi level statistics

RNAi redundancy	Regularized t	SAM t	student's t
2	77.1%	78.1%	74.8%
3	86.6%	88.1%	85.1%
4	91.8%	93.4%	91.4%
5	95.5%	95.9%	95.1%
6	97.6%	98.2%	97.3%
7	99.0%	98.9%	98.3%

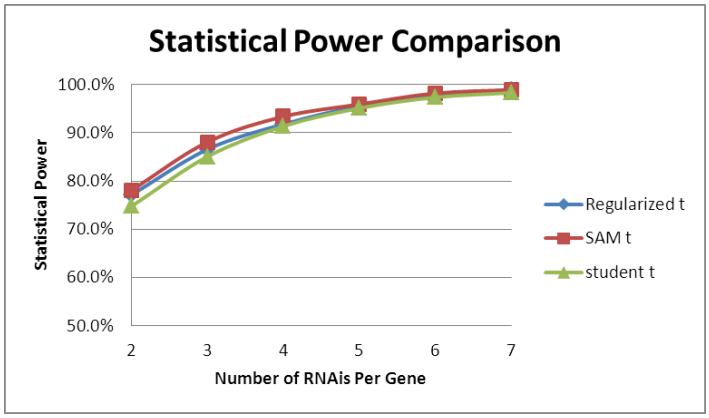

Figure 7.2 Comparison among RNAi level statistics

The type I error rate simulation results are listed in Table 7.3 and plotted in Figure 7.3. They indicate that SAM t has the highest type I error rate than the others (around 4~6% higher). The performances for regularized and student's t are close (the difference is around 2%).

Table 7.3 Type I error rate comparison among RNAi level statistics

RNAi redundancy	Regularized t	SAM t	student's t
2	8.2%	12.1%	6.7%
3	8.2%	12.4%	6.4%
4	8.9%	11.3%	7.0%
5	8.1%	12.1%	6.8%
6	7.7%	11.6%	7.1%
7	8.9%	11.3%	7.7%

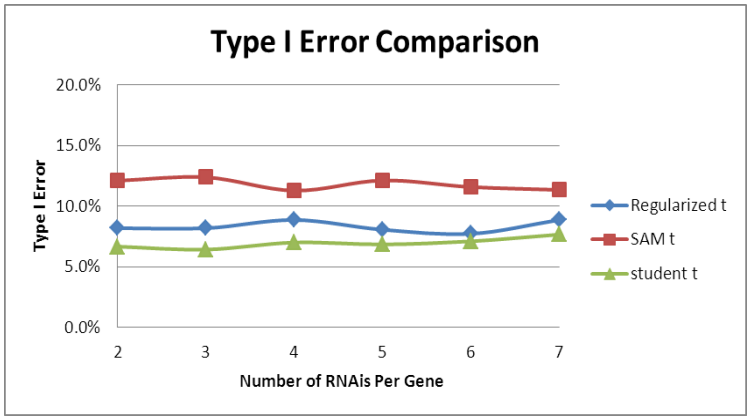

Figure 7.3 Type I error rate comparison

These results suggest regularized t might the best choice as RNAi level statistic if we take both sensitivity and specificity into consideration for the simulated data with "mean" chosen as RNAi set statistic. Below, we continue to investigate their performance when "max-mean" is chosen as RNAi set statistic.

Then we compare the RNAi level statistics when the RNAi set statistic is chosen as max-mean. The simulation settings are the same as above except that RNAi set statistics is chosen as max-mean.

The power simulation results are listed in Table 7.4 and plotted in Figure 7.4. They indicate that SAM t has the highest power than the others even though the power difference between SAM and regularized t is small (around 1%).

Table 7.4 Statistical power comparison among RNAi level statistics

RNAi redundancy	Regularized t	SAM t	student's t
2	78.0%	80.8%	76.9%
3	89.0%	90.2%	87.8%
4	94.4%	95.3%	93.3%
5	96.8%	97.6%	96.3%
6	98.6%	99.0%	98.0%
7	99.3%	99.4%	99.2%

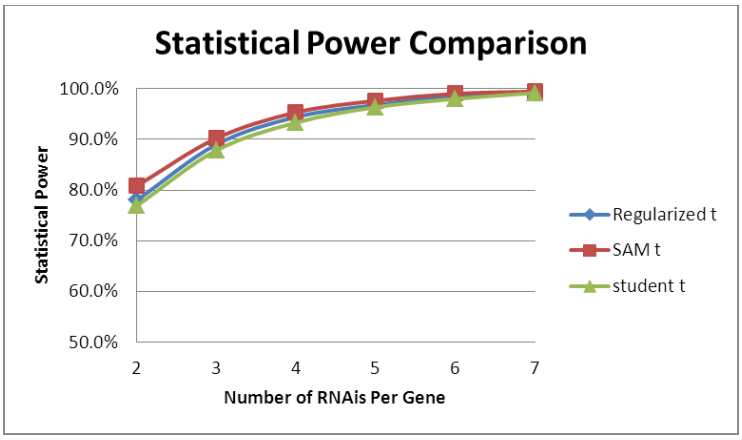

Figure 7.4 Comparison among RNAi level statistics

The type I error rate simulation results are listed in Table 7.5 and plotted in Figure 7.5. They indicate that SAM t has the highest type I error rate than the others (around 4~10% higher). The performances for regularized and student's t are close (the difference is around 1~3%).

Table 7.5 Type I error rate comparison among RNAi level statistics

RNAi redundancy	Regularized t	SAM t	student's t
2	8.4%	12.6%	7.6%
3	9.5%	14.2%	7.8%
4	9.9%	14.4%	7.7%
5	10.4%	16.2%	7.6%
6	9.8%	16.6%	7.8%
7	11.4%	18.6%	7.8%

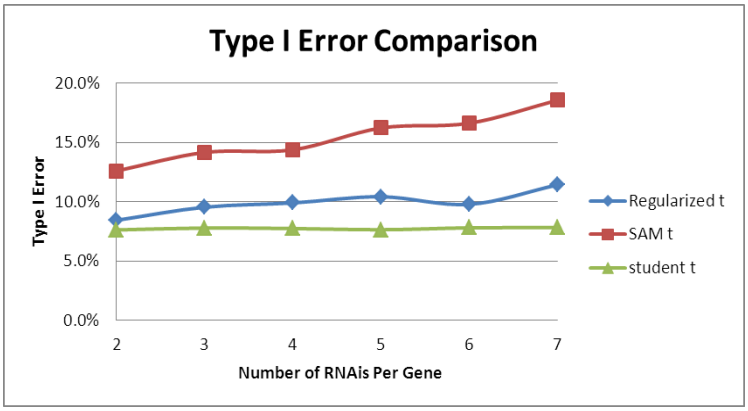

Figure 7.5 Comparison among RNAi level statistics

So these results suggest the same conclusion: regularized t might be the best choice as RNAi level statistic if we take both sensitivity and specificity into consideration. Below, we continue to investigate their performance when "max-mean" is chosen as RNAi set statistic.

(2) Sample size as independent variable

We then compare the performance of the three RNAi level statistics under different sample

size. The parameters used for this comparison are listed in Table 7.6. Note that different from the last section, we use RNAi resampling as the significance assessment method here because the sample size could be as small as 3. In each simulation run, we simulated RNAis targeting 2000 genes in total. 10% of these genes are drug resistant, 10% are drug sensitive and the left 80% are neutral. To obtain more accurate estimate, we run the simulation five times and estimate the power and type I error rate estimation as the average of the 5 runs.

Table 7.6 Parameters used for data generation

SimulationRun	5
RNAi set statistic	mean
significance assessment	RNAi resampling
Perm/Resample Num	5000
SensitiveGenePct	10%
ResistantGenePct	10%
TotalGeneNum	2000
RNAi redundency	4

The power simulation results are listed in Table 7.7 and plotted in Figure 7.6. They indicate that regularized t has the highest power than the others and student's t has the lowest power, with the difference as big as 13% for certain sample size.

Table 7.7 Statistical power comparison among RNAi level statistics

Sample Size	Regularized t	SAM t	student's t
3	57.5%	54.8%	44.4%
4	62.6%	59.2%	54.7%
5	63.6%	63.1%	59.8%
6	66.1%	66.4%	65.6%
7	67.2%	67.1%	64.3%
8	69.5%	69.5%	67.4%
9	70.3%	70.8%	69.1%

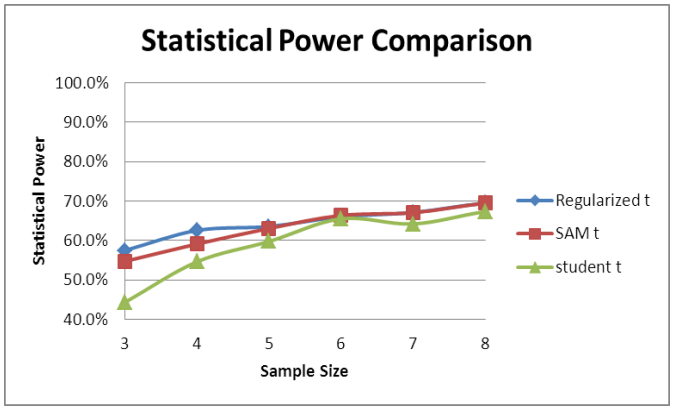

Figure 7.6 Comparison among RNAi level statistic options

The type I error rate simulation results are listed in Table 7.8 and plotted in Figure 7.7. They

indicate that regularized t has the lowest type I error rate even the magnitude is small (less than 2%) for all three methods.

Table 7.8 Type I error rate comparison among RNAi level statistics

Sample Size	Regularized t	SAM t	student's t
3	0.938%	0.913%	1.638%
4	0.463%	0.775%	1.050%
5	0.400%	0.363%	0.500%
6	0.163%	0.125%	0.300%
7	0.150%	0.113%	0.175%
8	0.025%	0.100%	0.125%
9	0.013%	0.050%	0.125%

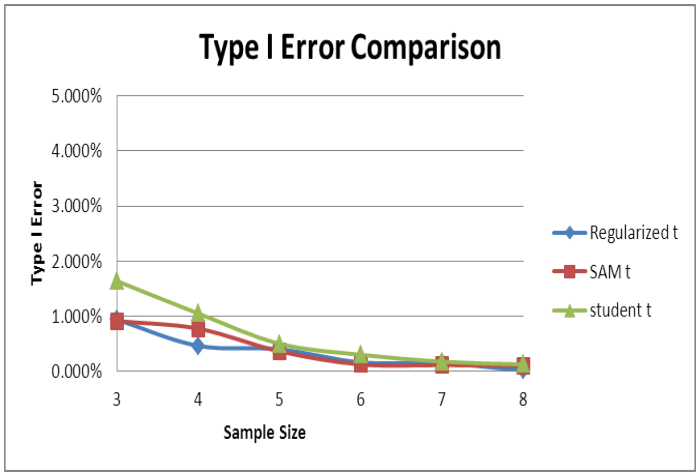

Figure 7.7 Comparison among RNAi level statistic options

In summary, all the above comparisons among the three RNAi level statistics suggest regularized t might be the best choice as RNAi level statistic if we take both sensitivity and specificity into consideration. Below, we continue to compare performances of the RNAi set statistics.

7.2.3 RNAi Set Statistic Comparison

For RNAi set statistic, we have three options: mean, max-mean and Kolmogorov –Smirnov statistics. We implemented the first two with Matlab. The third one is implemented in GSEA software package [7]. In this section, we compared their power and type I error rate curves for different RNAi redundancy for the first two. The performance of Kolmogorov –Smirnov statistics will be investigated in next section. In this simulation study, the RNAi level statistic is chosen as regularized t statistic, which is indicated to have best performance compared to the other two RNAi level statistics. Sample permutation is utilized in the significance assessment procedure and the number of replicates in each group is 10.

The power simulation results are listed in Table 7.9 and plotted in Figure 7.8. They indicate that max-mean has a little higher power than the mean, the difference ranging from 0.3% to 2.6%.

Table 7.9 Statistical power comparison between RNAi set statistics

RNAi redundancy	mean	max-mean
2	77.1%	78.0%
3	86.6%	89.0%
4	91.8%	94.4%
5	95.5%	96.8%
6	97.6%	98.6%
7	99.0%	99.3%

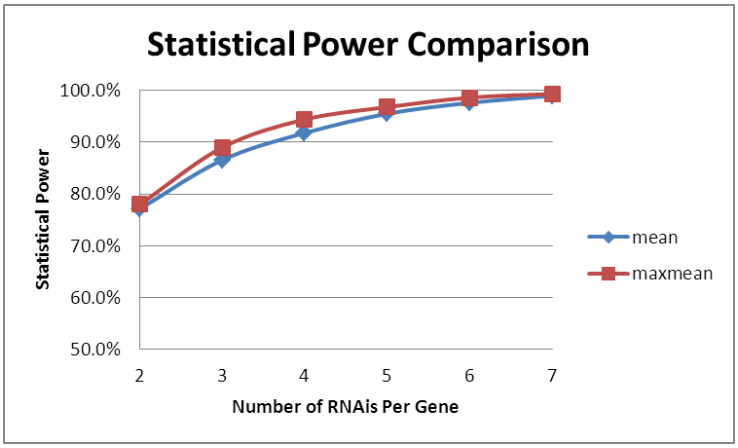

Figure 7.8 Comparison among RNAi set statistic options

The type I error rate simulation results are listed in Table 7.10 and plotted in Figure 7.9. They indicate that max-mean has a little higher type I error rate than mean, with the difference ranging from 0.2% to 2.5%.

Table 7.10 Type I error rate comparison between RNAi set statistics

RNAi redundancy	mean	max-mean
2	8.2%	8.4%
3	8.2%	9.5%
4	8.9%	9.9%
5	8.1%	10.4%
6	7.7%	9.8%
7	8.9%	11.4%

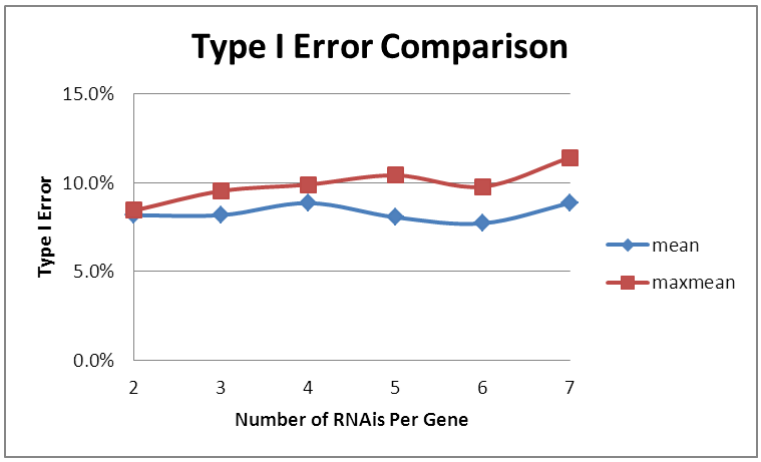

Figure 7.9 Comparison among RNAi set statistic options

In RNAi screen, many times the control of false negative rate is more important than the false positive rate because false positive candidates can be removed in following validation experiments. So based on the above comparison results, max-mean might be a choice compared to mean as RNAi set statistic if we take both sensitivity and specificity into consideration. Below, we are going to compare the performances of regularized t - maxmean combination to GSEA method. The reason why we don't compare max-mean Kolmogorov–Smirnov directly is that it's hard to integrate regularized t statistic into GSEA software package for us.

7.2.4 Regularized t & Max-mean vs GSEA

We compared their power and type I error rate curves based on simulated data. The comparison results are present below.

(1) Sample Permutation

First, we compare them based on sample permutation. The number of permutation times is 1000 for both.

The power simulation results are listed in Table 7.11 and plotted in Figure 7.10. They indicate that regularized t - Maxmean combination has much higher power than GSEA.

Table 7.11 Statistical power comparison between GSEA and RSEA

RNAi redundancy	Regularized t - maxmean	GSEA
2	78.0%	40.9%
3	89.0%	56.6%
4	94.4%	70.6%
5	96.8%	81.8%
6	98.6%	85.0%
7	99.3%	90.5%

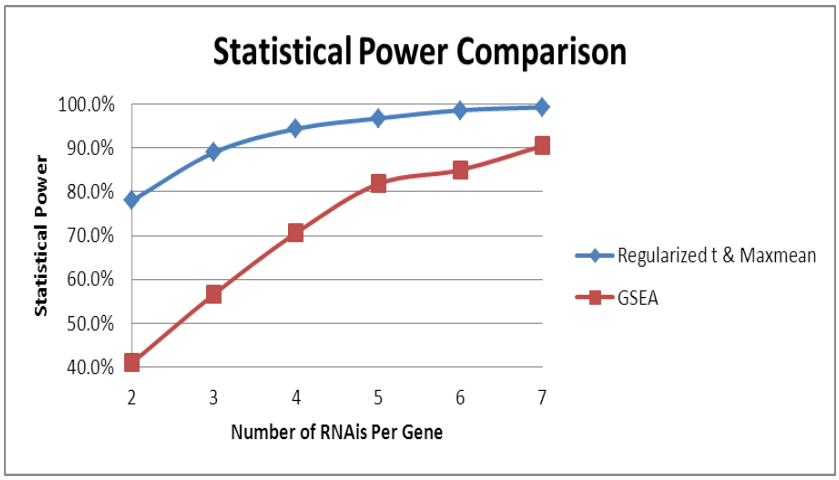

Figure 7.10 GSEA vs. regularized t & max-mean

The type I error rate simulation results are listed in Table 7.12 and plotted in Figure 7.11. They indicate that GSEA has lower type I error rate than the other. The type I error rate for GSEA is around 2.5% while for the other one, it's around 10%.

Table 7.12 Type I error rate comparison between GSEA and RSEA

RNAi redundancy	Regularized t & Max-mean	GSEA
2	8.4%	1.4%
3	9.5%	2.0%
4	9.9%	2.2%
5	10.4%	2.1%
6	9.8%	2.0%
7	11.4%	2.4%

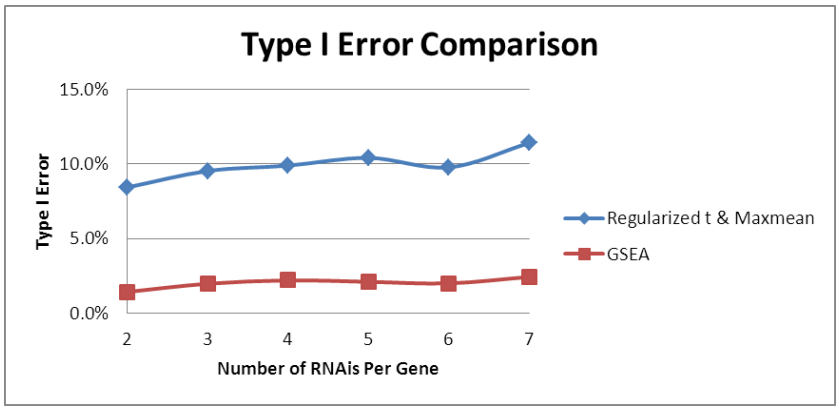

Figure 7.11 GSEA vs. regularized t & maxmean

As mentioned before, the control of false negative rate is more important than the false positive rate in RNAi screen at most times. So based on the above comparison results, regularized t - maxmean combination might be a better choice compared to GSEA. Below, we continue to compare their performances based on RNAi resampling.

(2) RNAi Resampling

To achieve a better approximation of the null distribution, the number of resampling times for regularized t – maxmean is set as 5000 in the simulation study. For GSEA, we choose 1000 times of resampling in consideration of the time cost (GSEA R package is much slower than our Matlab implemented regularized t – maxmean). We choose the default parameter setting for GSEA in their software package. The parameter settings for regularized t & max-mean are listed in Table 7.13:

Table 7.13 Parameter setting for regularized t – maxmean combination

SimulationRun	10
RNAi level statistic	Bayes t
RNAi set statistic	Max-mean
Perm/Resample Num	5000
SensitiveGenePct	0.1
ResistantGenePct	0.1
Sample size	10
TotalGeneNum	4000

The power simulation results are listed in Table 7.14 and plotted in Figure 7.12. They indicate that regularized t - Maxmean combination has much higher power than GSEA.

Table 7.14 Statistical power comparison between GSEA and RSEA

RNAi redundancy	Regularized t & Max-mean	GSEA
2	40.2%	38.2%
3	56.9%	46.8%
4	69.3%	61.3%
5	78.7%	73.8%
6	84.1%	81.0%
7	89.4%	84.8%

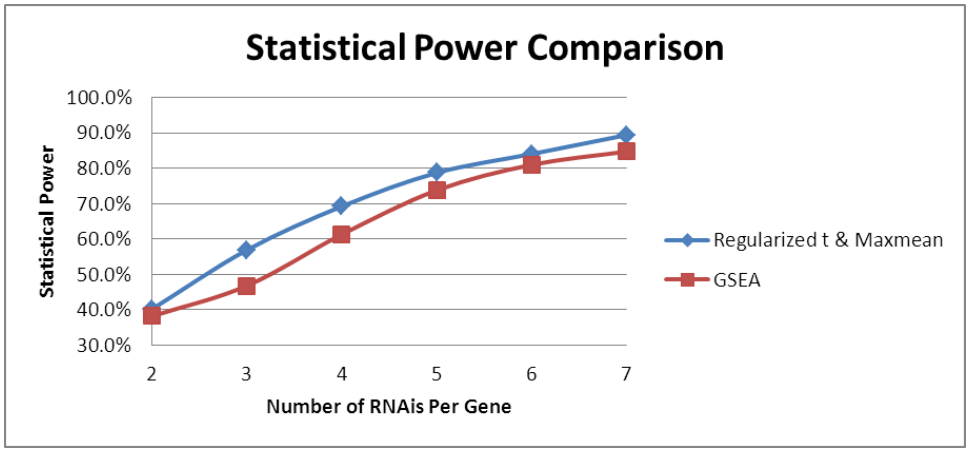

Figure 7.12 GSEA vs. regularized t & max-mean

The type I error rate simulation results are listed in Table 7.15 and plotted in Figure 7.13. They indicate that both have very lower type I error (less than 1%).

Table 7.15 Type I error rate comparison between GSEA and RSEA

RNAi redundancy	Regularized t & Max-mean	GSEA
2	0.009%	0.650%
3	0.003%	0.725%
4	0.000%	0.857%
5	0.000%	0.708%
6	0.000%	0.750%
7	0.006%	0.438%

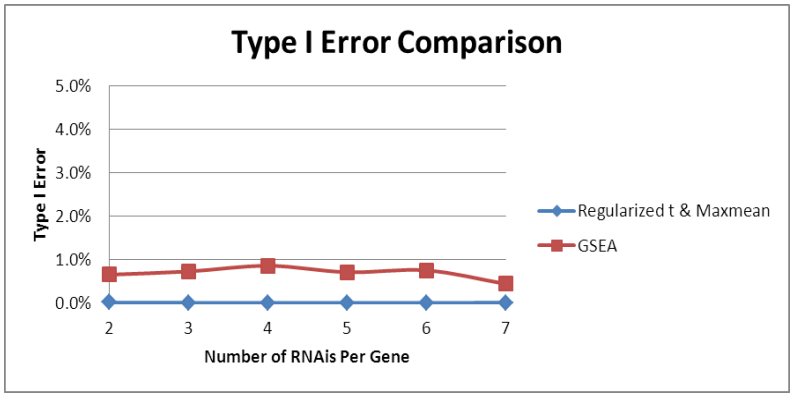

Figure 7.13 GSEA vs. regularized t & max-mean

So based on the above comparisons utilizing sample permutation or RNAi resampling, regularized t - maxmean combination might be a better choice compared to GSEA. Below, we are going to evaluate the performances of different Significance Assessment Methods.

7.2.5 Significance Assessment Methods Comparison

We studied the performance of resampling and permutation under different RNAi redundancy and sample size.

(1) RNAi redundancy as independent variable:

We first present the result under different RNAi redundancy. Parameters used to generate the data are listed in Table 7.16:

Table 7.16 Parameters used for data simulation

	Permutation	RNAi resampling
SimulationRun	5	10
RNAi level statistic	Bayes t	Bayes t
RNAi set statistic	Max-mean	Max-mean
Perm/Resample Num	1000	5000
SensitiveGenePct	25%	10%
ResistantGenePct	25%	10%
Sample size	10	10
TotalGeneNum	2000	4000

The power simulation results are listed in Table 7.17 and plotted in Figure 7.14. They indicate that permutation has much higher power than resampling, especially when the RNAi redundancy is low.

Table 7.17 Power comparison between resampling and permutation

RNAi redundancy	resampling	permutation
2	40.2%	78.0%
3	56.9%	89.0%
4	69.3%	94.4%
5	78.7%	96.8%
6	84.1%	98.6%
7	89.4%	99.3%

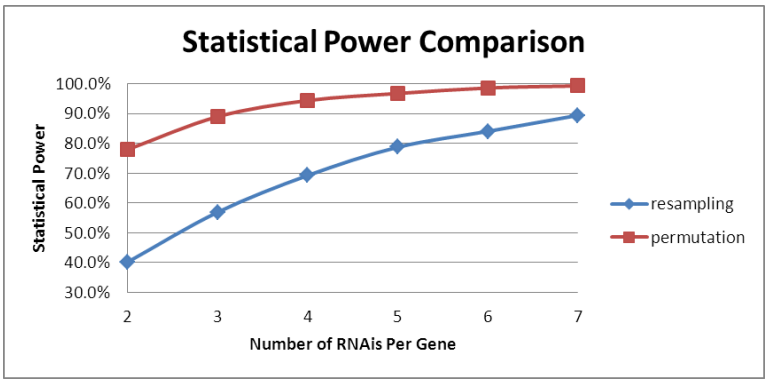

Figure 7.14 Comparison between resampling and permutation

The type I error rate simulation results are listed in Table 7.18 and plotted in Figure 7.15. They reveal that resampling has lower type I error rate than permutation. The type I error rate for resampling is almost 0 while for the other one, it's around 10%.

Table 7.18 Type I error rate comparison between resampling and permutation

RNAi redundancy	resampling	permutation
2	0.9%	8.4%
3	0.3%	9.5%
4	0.0%	9.9%
5	0.0%	10.4%
6	0.0%	9.7%
7	0.6%	11.4%

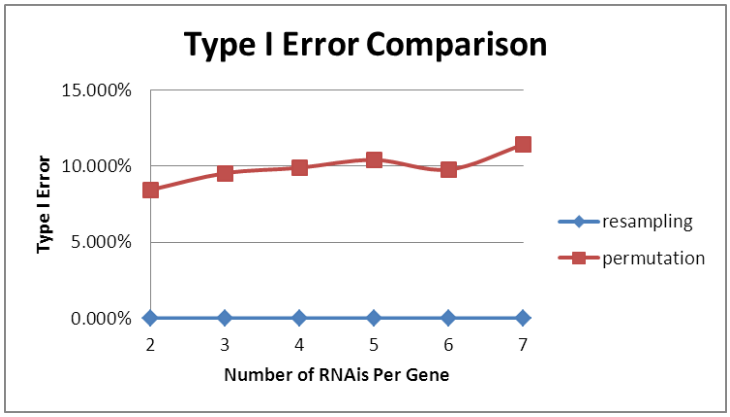

Figure 7.15 Comparison between resampling and permutation

(2) Sample size as independent variable:

We then present the result under different sample size. Parameters used to generate the data are listed in Table 7.19:

Table 7.19 Power comparison between resampling and permutation

	RNAi resampling	Sample permutation
SimulationRun	5	2
RNAi level statistic	Bayes t	Bayes t
RNAi set statistic	Max-mean	Max-mean
Perm/Resample Num	5000	1000
SensitiveGenePct	10%	25%
ResistantGenePct	10%	25%
TotalGeneNum	4000	1000
RNAi redundency	4	4

The power simulation results are listed in Table 7.20 and plotted in Figure 7.16. They indicate that permutation has much higher power than resampling for most cases, except when sample size =3. In that case, the power almost reduces to 0. This result matches our expectation. When sample size=3, there are only 20 unique permutations can be generated by the permutation method. So it's rare to have P-value less than 5%.

Table 7.20 Power comparison between resampling and permutation

Sample Size	resampling	permutation
3	54.3%	0.6%
4	59.9%	77.0%
5	63.5%	86.3%
6	64.9%	89.4%
7	63.3%	89.9%
8	65.5%	92.4%
9	68.9%	92.7%
10	69.3%	94.4%

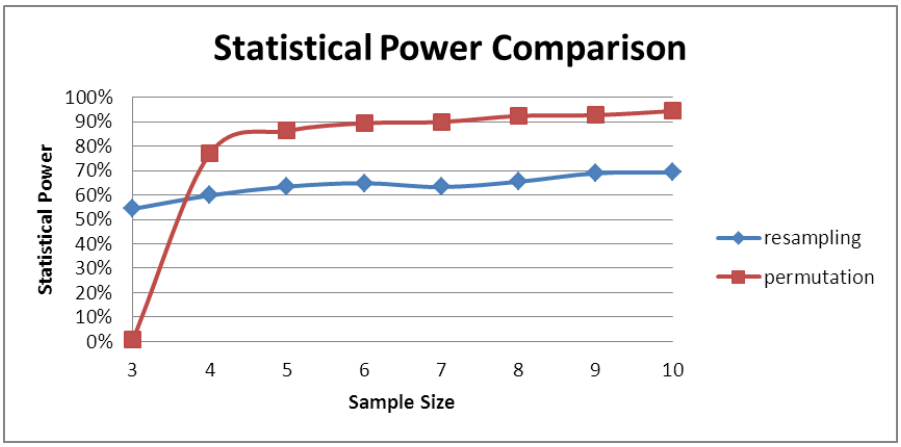

Figure 7.16 Comparison between resampling and permutation

The type I error rate simulation results are listed in Table 7.21 and plotted in Figure 7.17. They reveal that resampling has lower type I error rate than permutation. The type I error rate for resampling is almost 0, while for the other, it's still around 10%.

Table 7.21 Type I error rate comparison between resampling and permutation

Sample Size	resampling	permutation
3	0.45%	0.10%
4	0.28%	8.30%
5	0.19%	8.20%
6	0.09%	9.60%
7	0.05%	8.20%
8	0.01%	10.70%
9	0.00%	9.80%
10	0.00%	9.90%

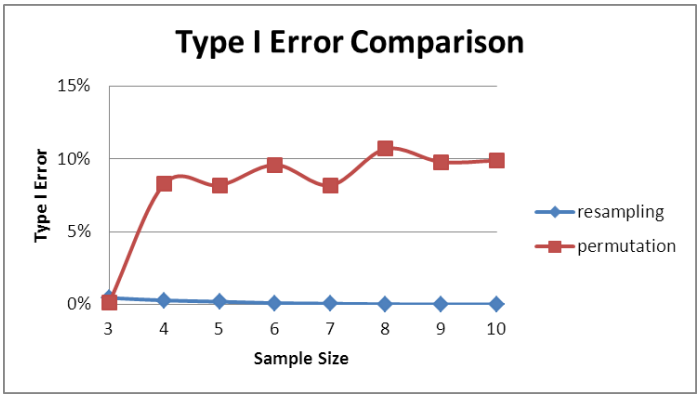

Figure 7.17 Comparison between resampling and permutation

So based on the above comparison studies, for RNAi screen data analysis, permutation might be a better choice compared to resampling, especially when sample size is larger than 6 (there are more than 1000 unique permutations when sample size is larger than 6). When sample size is small, resampling might a better choice especially for the extreme case when sample size equal to 2 or 3.

7.3 SEM Simulation Result

We compare the power and type I error of the three models based on simulated data. In each simulation, the total number of genes is 1000.

7.3.1 Powers Comparison (RNAi Redundancy)

The result is shown in Figure 7.18. The RNAi redundancy ranges from 2 to 7. From that figure, we can see RMANOVA have slightly higher statistical power than SEM and significantly higher power than MANOVA. The regular latent SEM is believed to match the RNAi screen mechanism more closely. However in Figure 7.18 its power is lower than uni-variate approach for repeated measures ANOVA. The explanation might be that regular latent SEM has too many parameters to estimate, which results in the frequent LEVMAR Optimization failure in TCALIS procedure of SAS software. When we increase the RNAi off target effect, the power comparison between regular latent SEM and uni-variate approach for repeated measures ANOVA is shown in Figure 7.19. In this simulation, gene effect is set as 0.8, variance of measurement error is set as 0.16, and variance of off target effect is set as 0.25. Figure 7.19 reveals that for large off target effect, regular latent SEM has higher power than uni-variate approach for repeated measures ANOVA. This matches our expectation. Regular latent SEM allows different coefficient for the RNAis targeting the same gene, compared to the equal coefficient assumption in uni-variate approach for repeated measures ANOVA. Thus it should has higher power to detect the gene effect signal behind the off target effect signal.

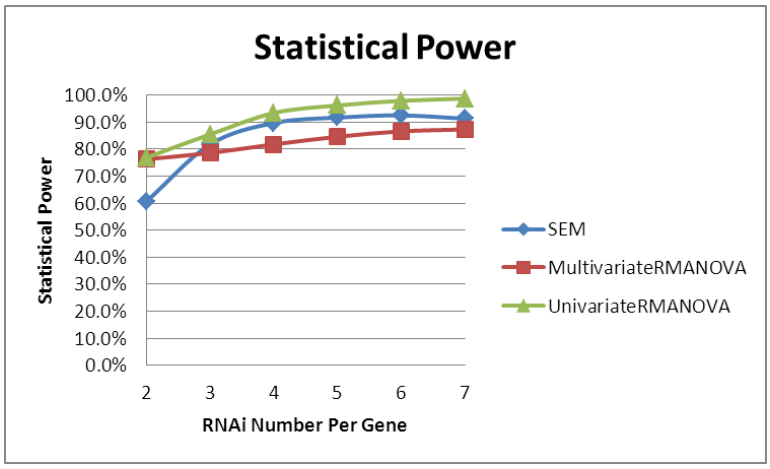

Figure 7.18 Statistical power varies with RNAi redundancy

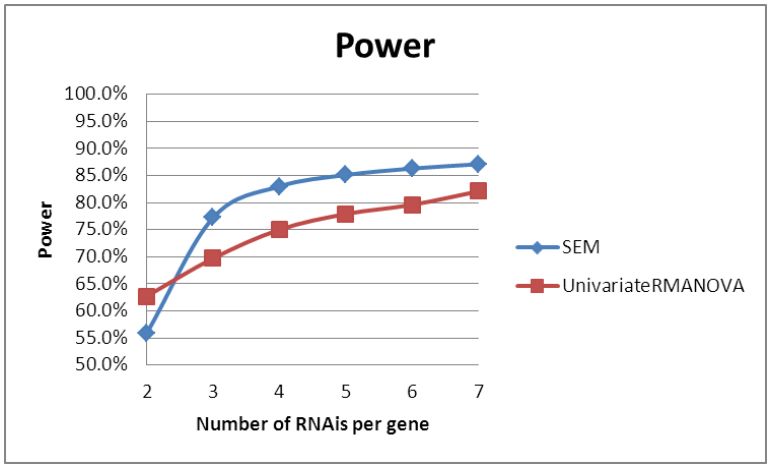

Figure 7.19 Power comparison under large off target effect

7.3.2 Power Comparison (Sample Size)

The experiment has two groups, treatment group and control group. In the simulation study, we let the two groups have equal sample size n. We investigate the influence of sample size n on statistical power in the SEM model. The result is shown in Figure 7.20. In our simulation study, the sample size in each group ranges from 5 to 30. As we can see from the plot, the statistical power of all the three models increases with sample size. However, when sample size is fixed, RMANOVA and SEM have significantly higher power than MANOVA. And for SEM and RMANOVA, the statistical power increases relatively fast from 79% to 89% when the sample size increases from 5 to 10. After that, the power stabilizes even if the sample size continues to increase. For MANOVA, the statistical power continues to grow with the increase of sample size.

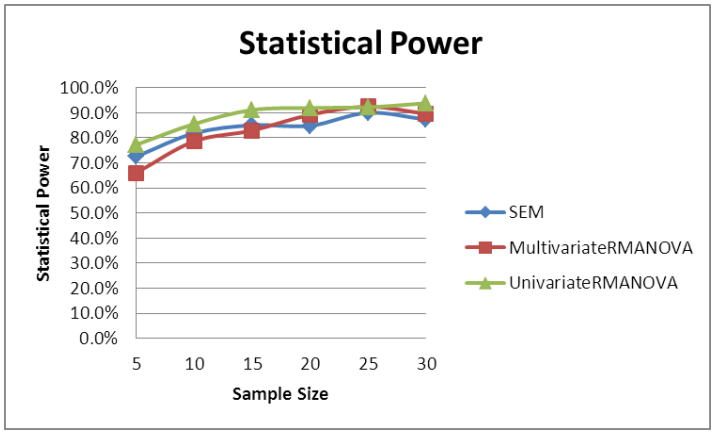

Figure 7.20 Statistical power varies with sample size

7.3.3 Type I Error Rate Comparison (RNAi Redundancy)

I also compared the different performance of the three models on type I error rate. First, I varied the RNAi redundancy with other parameters fixed. Figure 7.21 shows the result. We can see that RMANOVA has lowest type I error rate compared to the other two models. SEM has the highest type I error rate in most cases except when RNAi number = 2. And we can see that for

MANOVA, the RNAi number has significant effect on type I error rate. MANOVA's type I error rate increases from 6% to 24% as the RNAi number increases from 2 to 4. But after that, the error rate goes all the way down to16% as the RNAi number continues to increase to 7. Compared to MANOVA, the type I error rate changes slightly as the RNAi redundancy increase, for both SEM and RMANOVA models.

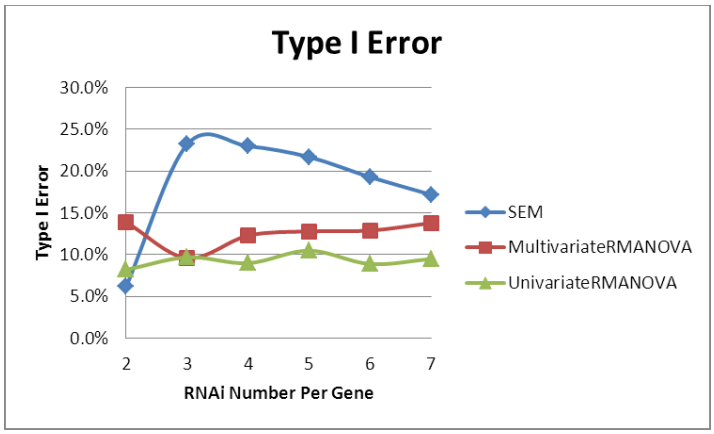

Figure 7.21 Type I error rate varies with RNAi redundancy

7.3.4 Type I Error Rate Comparison (Sample Size)

We also studied the sample size influence on type I error rate. Figure 7.22 shows the result for all three models when RNAi number is fixed at 2. We can see that SEM has the lowest type I error rate and MANOVA has the highest one. As sample size increase, the type I error rate decreases for SEM and RMANOVA. However, there is oscillation for RMANOVA.

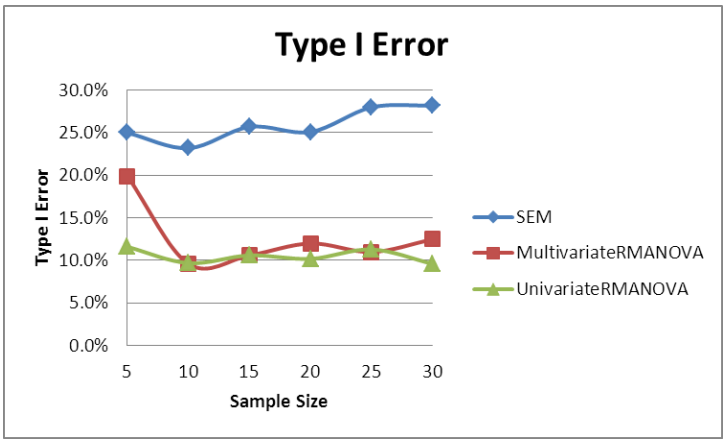

Figure 7.22 Type I error rate varies with sample size

7.4 RSEA - SEM Comparison

We have done the comparisons within RSEA and with SEM. Here we are going to compare RSEA directly to SEM. Again, we compared their performances under different RNAi redundancy and sample size separately.

7.4.1 RNAi Redundancy as Independent Variable

First we compared their performances under different RNAi redundancy. The parameter setting for SEM is described in Table 7.22 . And for RSEA, it's listed in Table 7.23.

Table 7.22 Paremeter setting for SEM

model	UniVariateRMANOVA
genenum	1000
repnum	10

Table 7.23 Parameter setting for RSEA

SimulationRun	5
RNAi level statistic	Bayes t
RNAi set statistic	Max-mean
significance assessment	permutation
Perm/Resample Num	1000
SensitiveGenePct	25%
ResistantGenePct	25%
Sample size	10
TotalGeneNum	2000

The power simulation results are listed in Table 7.24 and plotted in Figure 7.23. They indicate that SEM and RSEA have very similar power for most cases. For RNAi redundancy =4, there seems to be a perturbation on the curve. This perturbation might come from the estimation variation considering the SEM simulation, the power is calculated on only 1000 genes.

Table 7.24 Power comparison between Univariate RMANOVA and RSEA

RNAi redundancy	Univariate RMANOVA	RSEA
2	76.9%	78.0%
3	85.5%	89.0%
4	93.4%	94.4%
5	96.2%	96.8%
6	97.9%	98.6%
7	98.7%	99.3%

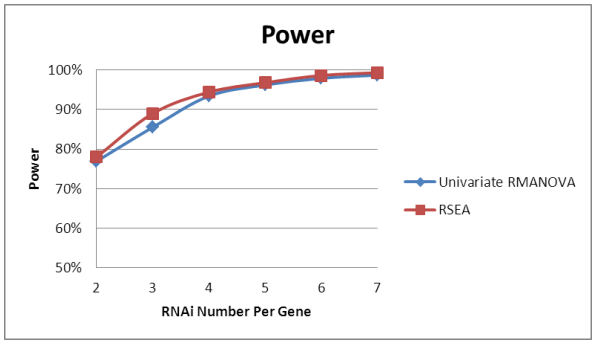

Figure 7.23 Univariate RMANOVA vs. RSEA

The type I error rate simulation results are listed in Table 7.25 and plotted in Figure 7.24. They

reveal that RSEA and SEM have very similar type I error rate performances. The type I error of RSEA goes up slowly as RNAi redundancy increases. While for SEM, this rate remains constant with small oscillation.

Table 7.25 Comparison between Univariate RMANOVA and RSEA

RNAi redundancy	Univariate RMANOVA	RSEA
2	8.2%	8.4%
3	9.7%	9.5%
4	9.0%	9.9%
5	10.5%	10.4%
6	8.9%	9.8%
7	9.5%	11.4%

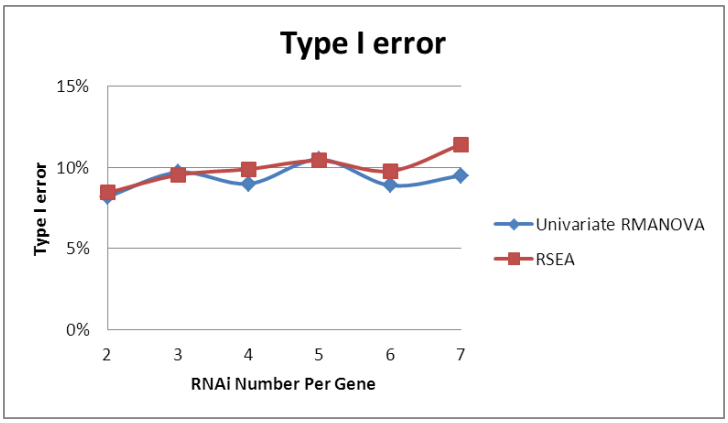

Figure 7.24 Univariate RMANOVA vs. RSEA

7.4.2 Sample Size as Independent Variable

Then we compared their performances under different sample size. The parameter setting for SEM is described in Table 7.26. And for RSEA, it's listed in

Table 7.27.

Table 7.26 Parameters used for SEM

Model	UNIRMANOVA
Genenum	1000
Hpnum	3
Workhpct	0.8

Effect	3
Exptsigma	1
offtgsigma	0.2

Table 7.27 Parameters used for RSEA

SimulationRun	2
RNAi level statistic	Bayes t
RNAi set statistic	Max-mean
significance assessment	permutation
Perm/Resample Num	1000
SensitiveGenePct	25%
ResistantGenePct	25%
TotalGeneNum	1000
RNAi redundency	3

The power simulation results are listed in Table 7.28 and plotted in Figure 7.25. They indicate that RSEA has a little bit higher power than SEM for most cases (around 2%).

Table 7.28 Power comparison

Sample Size	Univariate RMANOVA	RSEA
5	77.2%	79.4%
10	85.5%	89.0%
15	91.1%	90.9%
20	91.9%	93.2%
25	92.2%	94.7%
30	93.8%	95.7%

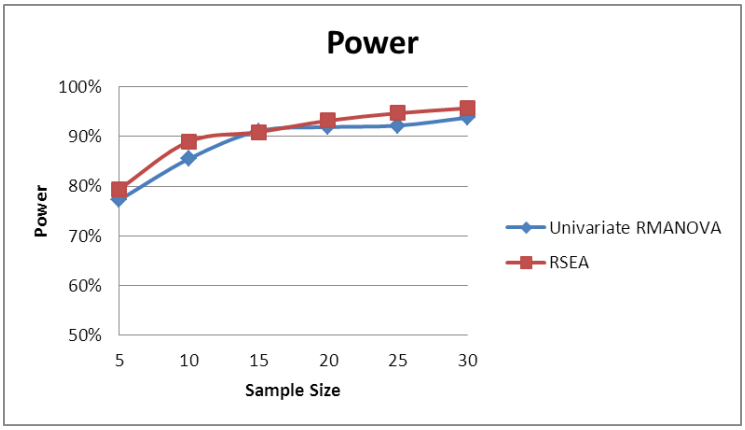

Figure 7.25 Univariate RMANOVA vs. RSEA

The type I error rate simulation results are listed in Table 7.29 and plotted in Figure 7.26. They reveal that the type I error of SEM is somehow lower than RSEA at mast cases. And the type I error of SEM is more stable compared to RSEA. From Figure 7.26, it's indicated that the type I error of RSEA goes up gradually as sample size increase. While for SEM, the curve remains flat.

Table 7.29 Type I error rate comparison

Sample Size	Univariate RMANOVA	RSEA
5	11.6%	9.9%
10	9.7%	9.5%
15	10.6%	11.8%
20	10.2%	11.0%
25	11.3%	12.1%
30	9.6%	14.1%

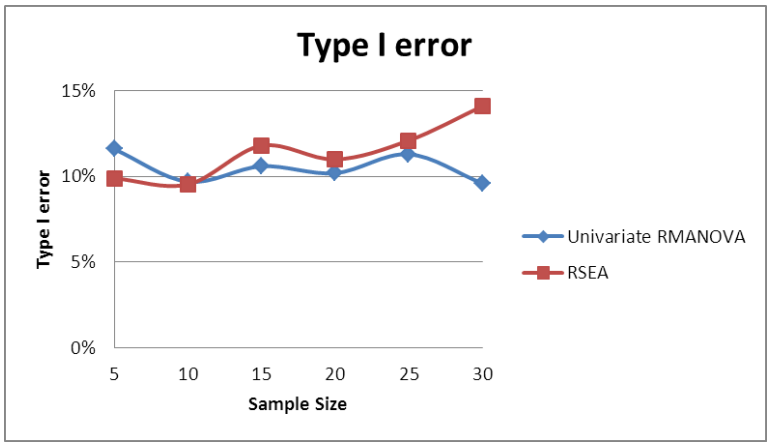

Figure 7.26 Univariate RMANOVA vs. RSEA

In summary, SEM and RSEA have very similar and comparable statistical performance. The RSEA might have a little bit higher power in some cases while the type I error rate of SEM is more controllable, without increase with sample size or RNAi redundancy.

Chapter 8 . Case Study

In the previous chapter, we investigate the performance of RSEA and SEM based on simulated data. In this chapter, we will apply the RSEA in the data analysis of PLKi RNAi screen accomplished in our lab. Important results of data analysis will be presented and discussed, including the preprocessing part.

8.1 Preprocessing

After we get the raw microarray data from Agilent feature extraction software, the first step we did is to generate boxplots to check the quality of each array.

8.1.1 Boxplot

Figure 8.1 to Figure 8.5 show the boxplots for all 5 cell lines. The left represents the barcode probes and the right represents half hairpin probes. The green color represents the cy3 channels and the red represents the cy5 channels. Based on these graphs, there seem to be one outlier in H460. After careful check with the outlier, we decided to keep but will pay special attention to it in the following analysis.

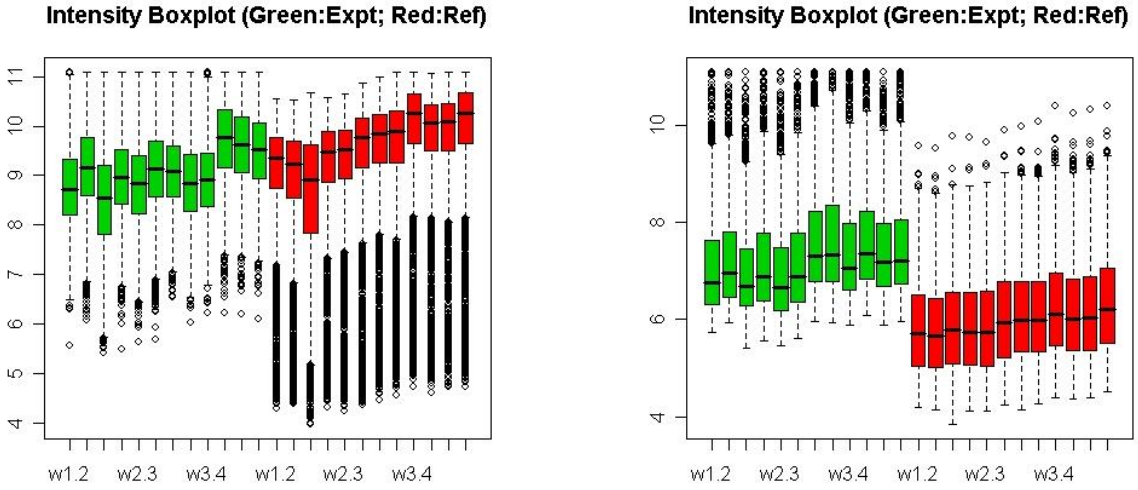

Figure 8.1 Intensity boxplot for A549

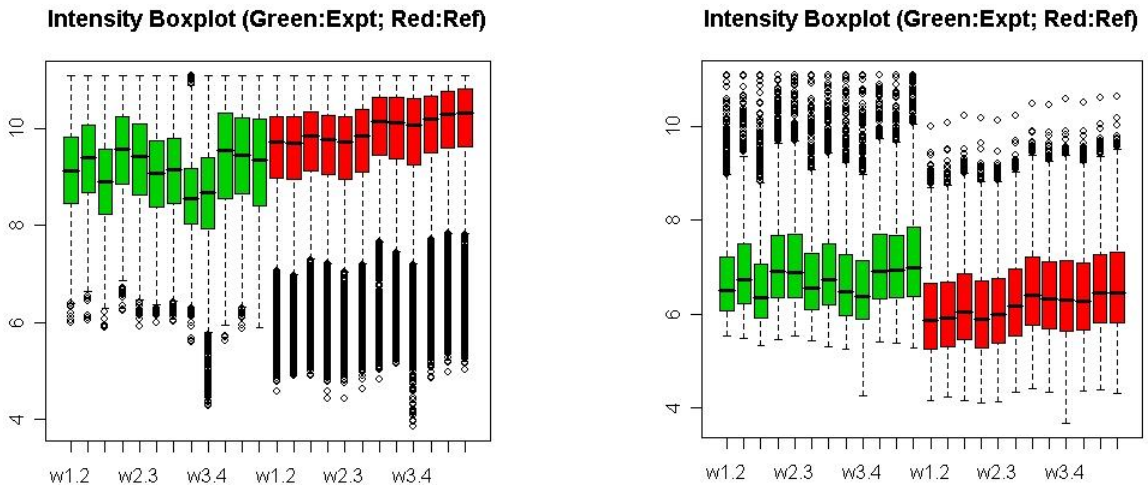

Figure 8.2 H322 intensity boxplot

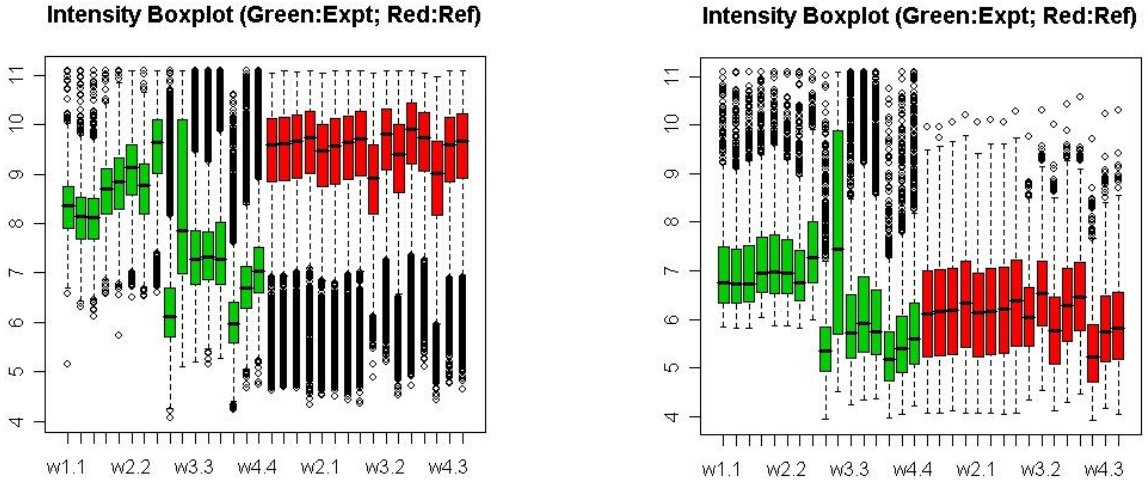

Figure 8.3 H460 intensity boxplot

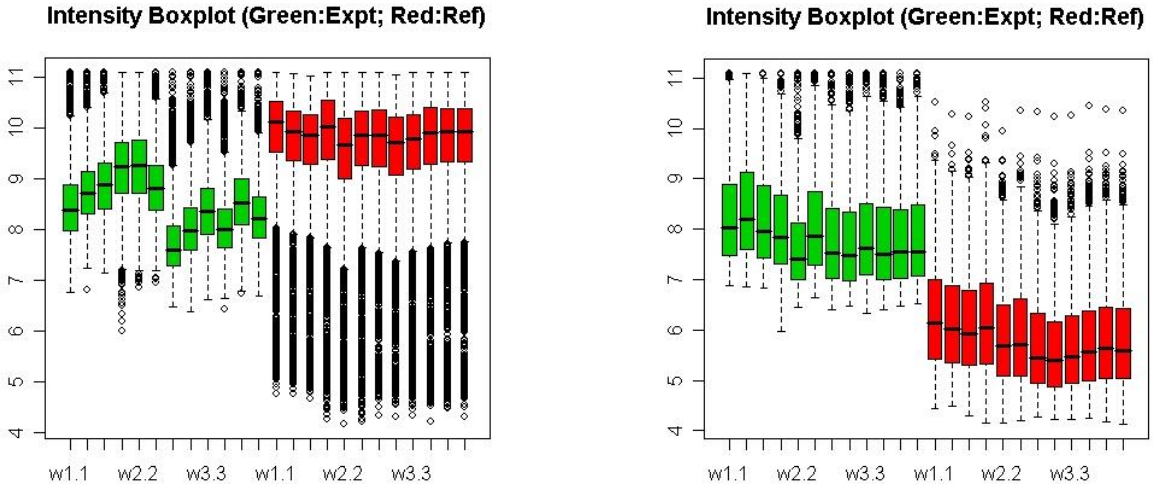

Figure 8.4 H522 intensity boxplot

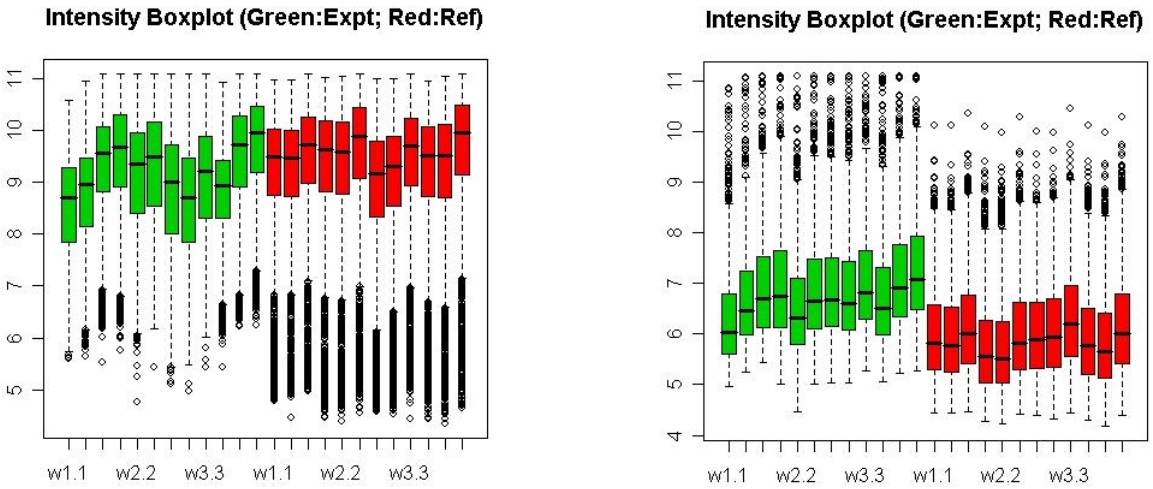

Figure 8.5 J42-L83 intensity boxplot

8.1.2 PCA

After preprocessing as described in chapter 4, we checked the PCA plots of these cell lines (shown in Figure 8.6 to Figure 8.8). In these plots, black and green circles represents arrays from NONE and DMSO groups accordingly, and green and blue circles represents arrays from LOW and HIGH groups accordingly. In these plots, the drug treatment groups are partially or well separated from control groups, which indicates that drug treatment is an important source of variation in these experiments.

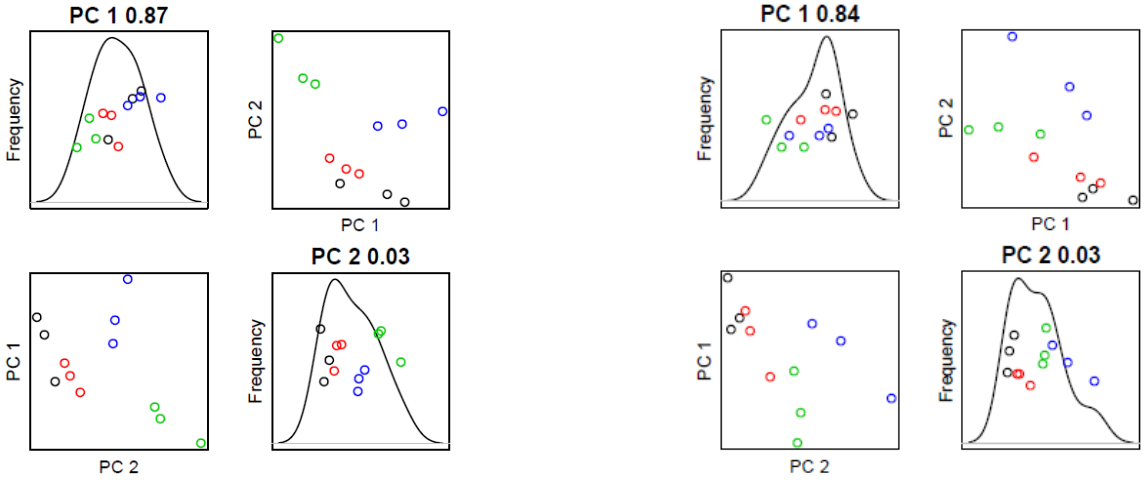

Figure 8.6 A549 (left) and H322 (right)

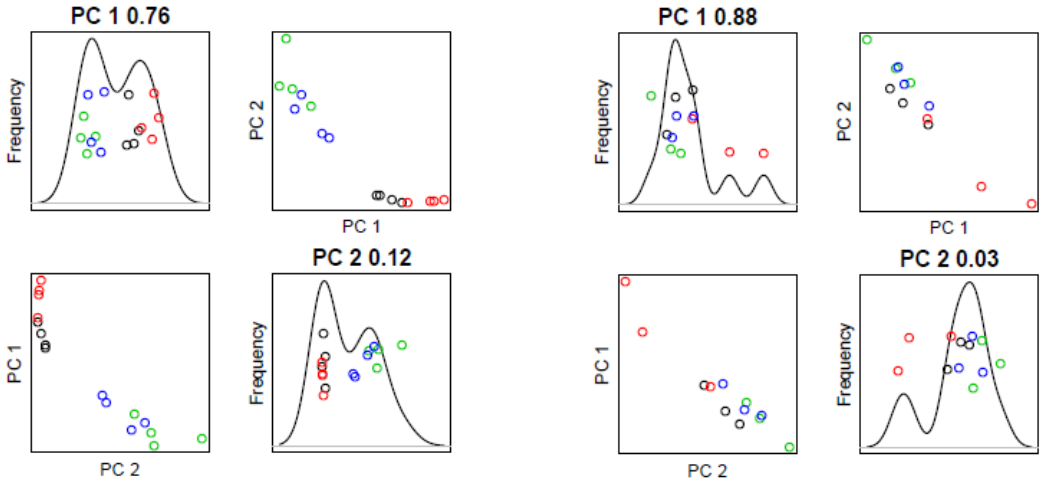

Figure 8.7 H460 (left) and H522 (right)

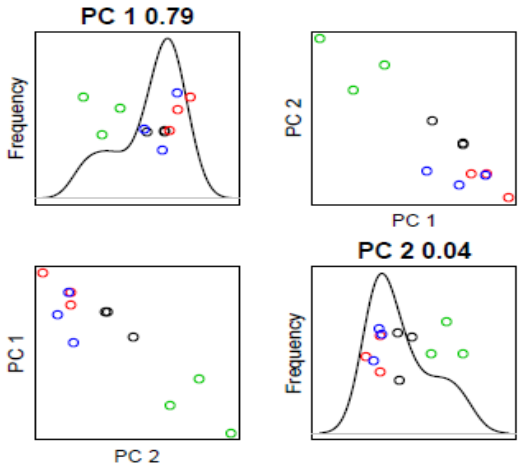

Figure 8.8 J42-L83 PCA plot

8.2 GSA analysis

Previously, we used Tibshirani's Gene Set Analysis [100] to analyze all the five cell lines. The default GSA parameters are used when running the R software package "GSA". Number of permutation times is set as 1000. Based on the analysis, gene RARA ("retinoic acid receptor alpha") is the only one significantly enriched in four out of all five cell lines at 5% significance level. The test detail is as follows:

Table 8.1 GSA test result for RARA in all five cell lines

CellLine	P value	Total RNAi Num	Depleted RNAi Num	Enriched RNAi Num	Fold Change
A549	0	4	1	3	1.5
H322	10.2%	4	1	3	1.6
H460	0	3	0	3	4.8
H522	0	3	1	2	1.6
J42-L83	0	4	1	3	1.6

In the RNAi library, gene "RARA" has four RNAis each targeting different positions of this gene, which are listed in Table 8.2. The details such as the sequence design of these RNAis can be found at http://cancan.cshl.edu/cgi-bin/Codex/Codex.cgi.

There is one issue needs to be clarified regarding to Table 8.1. In the column "Total RNAi Num", the values changes between 3 and 4, rather than being constant. This is due to the background filter. For example, in cell line H460, RNAi "V2HS_239390" is removed from the GSA analysis by the background filter due to low probe intensity.

Table 8.2 Four RNAis targeting RARA.

V2HS_131541	V2HS_131536	V2HS_239486	V2HS_239390

8.3 RSEA analysis

Besides the GSA, we use the general RSEA framework to test the significance of gene RARA. Based on the same data, we get the result as shown in Table 8.3.

Table 8.3 RSEA result for RARA in all five cell lines.

Cell Line	P-value based on sample permutation	P-value based on RNAi resampling
A549	6.6%	5.1%
H322	11.3%	4.6%
H522	9.1%	28.7%
J42	9.5%	10%
H460	2.5%	2.8%

In RSEA, the RNAi level statistic is chosen as "Regularized t test", and the RNAi set statistic is chosen as "Max-mean" statistic. For significance assessment, we tried both sample permutation and RNAi resampling, and the p values are reported in the second and third columns separately. As we can see, the p values in Table 8.3 are much higher than those in Table 8.1. However, considering the small sample size (four cell lines has only three replicates in each condition and the other has four replicates), p values in Table 8.3 seems to be more trustable. And if we choose 10% significance level, RARA will still be significantly enriched in four out of all five cell lines, which means RARA might be a drug resistant gene.

8.4 Validation

To verify our finding, Nancy Liu, a former postdoc in our lab who did most if not all of the experiments in the PLKi RNAi screen project, did the validation experiments.

8.4.1 Silencing RARA Confers PLKi Resistance

First validation experiment is to test the hypothesis that silencing RARA confers PLKi resistance. H460 cancer cells are devided into three groups: mock treatment, low dose drug treatment and high dose drug treatment. For each group, we have three replicate plates, and we use RNAi C11 to silence down RARA gene expression in one plate, and then use another RNAi G6 to

silence down RARA in the second plate. The third plate is left alone without RNAi silencing. Then for each plate, we measure the percent of variable cells relative to control plate (without drug treatment or shRNA silencing). The measurement result is illustrated in Figure 8.9. The left shows the Effect of RARA RNAis on PLKi resistance and the right shows the western blot analysis.

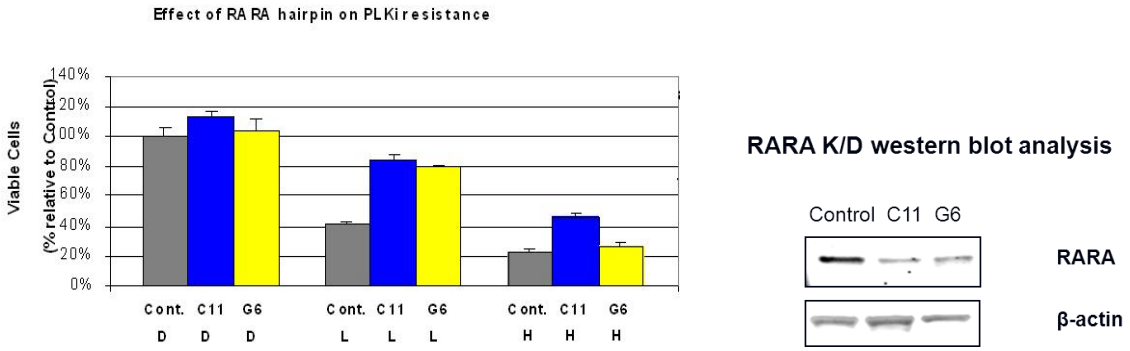

Figure 8.9 Silencing RARA confers PLKi resistance

From Figure 8.9(a), we can see that in both low and high dose drug treatment groups, the number of viable cells in the RNAi C11 silenced plates increases compared to the plate without RNAi silencing. And Figure 8.9(b) western blot confirms that RARA gene expression is significantly reduced by RNAi C11.

8.4.2 RARA Activation Confers PLKi Sensitivity

The second validation experiment is to test the hypothesis that RARA activation confers PLKi sensitivity. Retinoic acid receptor (RAR) belongs to a gene superfamily of hormone nuclear receptors that act as ligand-dependent transcriptional factors [108]. In our validation experiment, we use three different agents to activate RARA expression separately: All-trans retinoic acid (ATRA) which is a ligand of RAR and exerts its biologic effect by binding to RAR, 9-cis retinoic acid (9-cis-RA)which binds only to retinoid X receptor, and Am80 which is RAR alpha-specific agonist. The experiment result is shown in Figure 8.10~Figure 8.12.

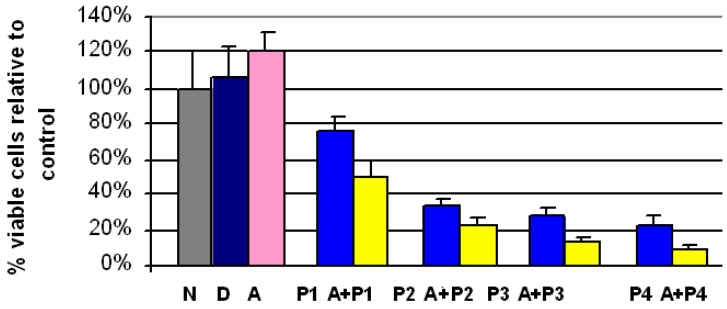

Figure 8.10 Combination treatment of ATRA+PLKi in H460

In Figure 8.10, H460 cells were treated with ATRA, PLKi, or a combination of ATRA and PLKi. N = untreated. D = DMSO. A = ATRA (1 μM). P1 = PLKi (20nM). P2 = PLKi (30nM). P3 = PLKi (100nM). P4 = PLKi (300nM).

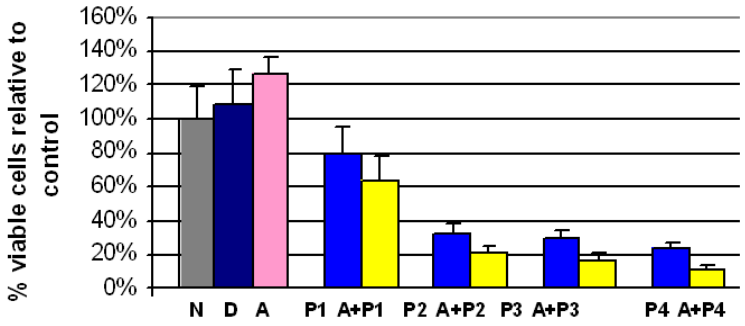

Figure 8.11 Combination treatment of 9-cis-RA +PLKi in H460

In Figure 8.11, H460 cells were treated with 9-cis-RA, PLKi, or a combination of 9-cis-RA and PLKi. N = untreated. D = DMSO. A = 9-cis-RA (1 μM). P1 = PLKi (20nM). P2 = PLKi (30nM). P3 = PLKi (100nM). P4 = PLKi (300nM).

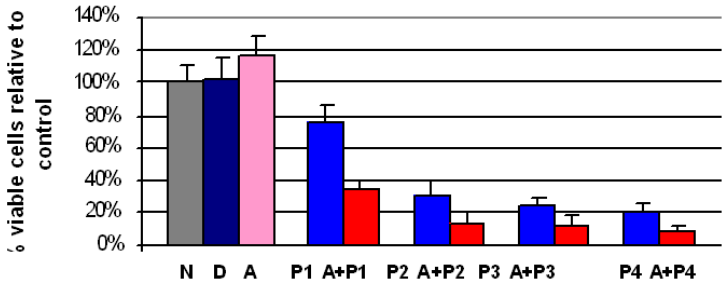

Figure 8.12 Combination treatment of Am80 + PLKi in H460

In Figure 8.12, H460 cells were treated with ATRA, PLKi, or a combination of Am80 and PLKi. N = untreated. D = DMSO. A = Am80 (1 μM). P1 = PLKi (20nM). P2 = PLKi (30nM). P3 = PLKi (100nM). P4 = PLKi (300nM).

We can clearly see that all three agents sensitize H460 cells to PLKi. This result gives us more confidence that RARA activation confers PLKi sensitivity.

Chapter 9 . Discussion and Conclusion

In this thesis, we propose RNAi Set Enrichment Analysis method and Structural Equation Modeling to take multiple RNAis into consideration to access the gene effect on drug response.

RSEA has multiple modules. And each module has multiple statistic choices. Based on simulation studies, we tested which statistic choice has the best performance in terms of statistical power and type I error rate. The results indicate that the combination of regularized t static as RNAi level statistic, max-mean as RNAi level, and permutation as significance assessment method might achieve the best performance in most cases. However, if the sample size is small, resampling, instead of permutation, might be the better choice especially for the extreme case when sample size equal to 2 or 3.

For SEM, our simulation studies reveal that Uni-variate analysis approach for repeated measures ANOVA might be the better choice for small target effect. However, for large off target effect, regular latent SEM has higher statistical power than univariate approach for repeated measures ANOVA

Compared to RSEA, SEM has very similar statistical performance in our simulations. The RSEA might have a little bit higher power in some cases while the type I error rate of SEM is more stable, without increase with sample size or RNAi redundancy.

To verify our models, we apply them with real data from our experiments. The result shows that the drug resistant candidate gene RARA identified by our models is highly likely to be a true positive based on the validation experiments.

Of course, most of our current simulation results are based on the small off target effect and hairpin independence structure assumption. In the future, simulations under different assumptions should be done. For example, we can vary the standard deviation of measurement error, standard deviation of the off target effect, hairpin working probability and hairpin correlation structure. Under these situations, the performance of the RSEA and SEM models might be different from current results.

In addition, the validity and usefulness of our models still need to be tested with more real biological data. In the future, we hope more and more RNAi screen data will be published and can be publically accessed.

References

1. Degenhardt YY, Wooster R, McCombie RW, Lucito R, Powers S: **High-content analysis of cancer genome DNA alterations**. *Curr Opin Genet Dev* 2008, **18**(1):68-72.

2. Schlabach MR LJ, Solimini NL, Hu G, Xu Q, Li MZ, Zhao Z, Smogorzewska A, Sowa ME, Ang XL, et al.: **Cancer proliferation gene discovery through functional genomics**. *Science* 2008(**319**):5.

3. Silva JM, Marran K, Parker JS, Silva J, Golding M, Schlabach MR, Elledge SJ, Hannon GJ, Chang K: **Profiling essential genes in human mammary cells by multiplex RNAi screening**. *Science* 2008, **319**(5863):617-620.

4. Boutros M, Kiger AA, Armknecht S, Kerr K, Hild M, Koch B, Haas SA, Paro R, Perrimon N: **Genome-wide RNAi analysis of growth and viability in Drosophila cells**. *Science* 2004, **303**(5659):832-835.

5. Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepfer AM, Hinkle G, Piqani B, Eisenhaure TM, Luo B, Grenier JK *et al*: **A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen**. *Cell* 2006, **124**(6):1283-1298.

6. Kittler R, Putz G, Pelletier L, Poser I, Heninger AK, Drechsel D, Fischer S, Konstantinova I, Habermann B, Grabner H *et al*: **An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division**. *Nature* 2004, **432**(7020):1036-1040.

7. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.

8. Konig R, Chiang CY, Tu BP, Yan SF, DeJesus PD, Romero A, Bergauer T, Orth A, Krueger U, Zhou Y *et al*: **A probability-based approach for the analysis of large-scale RNAi screens**. *Nat Methods* 2007, **4**(10):847-849.

9. Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang X, Hinkle G, Boehm JS, Beroukhim R, Weir BA *et al*: **Highly parallel identification of essential genes in cancer cells**. *Proc Natl Acad Sci U S A* 2008, **105**(51):20380-20385.

10. Scholl C, Frohling S, Dunn IF, Schinzel AC, Barbie DA, Kim SY, Silver SJ, Tamayo P, Wadlow RC, Ramaswamy S *et al*: **Synthetic lethal interaction between oncogenic**

KRAS dependency and STK33 suppression in human cancer cells. *Cell* 2009, **137**(5):821-834.

11. Sawyers C: **Targeted cancer therapy**. *Nature* 2004, **432**(7015):294-297.

12. Karnoub AE, Weinberg RA: **Ras oncogenes: split personalities**. *Nat Rev Mol Cell Biol* 2008, **9**(7):517-531.

13. Roberts PJ, Der CJ: **Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer**. *Oncogene* 2007, **26**(22):3291-3310.

14. Hartwell LH, Szankasi P, Roberts CJ, Murray AW, Friend SH: **Integrating genetic approaches into the discovery of anticancer drugs**. *Science* 1997, **278**(5340):1064-1068.

15. Kaelin WG, Jr.: **The concept of synthetic lethality in the context of anticancer therapy**. *Nat Rev Cancer* 2005, **5**(9):689-698.

16. Bender A, Pringle JR: **Use of a screen for synthetic lethal and multicopy suppressee mutants to identify two new genes involved in morphogenesis in Saccharomyces cerevisiae**. *Mol Cell Biol* 1991, **11**(3):1295-1305.

17. Lucchesi JC: **Synthetic lethality and semi-lethality among functionally related mutants of Drosophila melanfgaster**. *Genetics* 1968, **59**(1):37-44.

18. Simons A, Dafni N, Dotan I, Oron Y, Canaani D: **Establishment of a chemical synthetic lethality screen in cultured human cells**. *Genome Res* 2001, **11**(2):266-273.

19. Stockwell BR, Haggarty SJ, Schreiber SL: **High-throughput screening of small molecules in miniaturized mammalian cell-based assays involving post-translational modifications**. *Chem Biol* 1999, **6**(2):71-83.

20. Bernards R, Brummelkamp TR, Beijersbergen RL: **shRNA libraries and their use in cancer genetics**. *Nat Methods* 2006, **3**(9):701-706.

21. Downward J: **Use of RNA interference libraries to investigate oncogenic signalling in mammalian cells**. *Oncogene* 2004, **23**(51):8376-8383.

22. Westbrook TF, Stegmeier F, Elledge SJ: **Dissecting cancer pathways and vulnerabilities with RNAi**. *Cold Spring Harb Symp Quant Biol* 2005, **70**:435-444.

23. Ngo VN, Davis RE, Lamy L, Yu X, Zhao H, Lenz G, Lam LT, Dave S, Yang L, Powell J *et al*: **A loss-of-function RNA interference screen for molecular targets in cancer**. *Nature* 2006, **441**(7089):106-110.

24. Turner NC, Lord CJ, Iorns E, Brough R, Swift S, Elliott R, Rayter S, Tutt AN, Ashworth A: **A synthetic lethal siRNA screen identifying genes mediating sensitivity to a PARP inhibitor**. *EMBO J* 2008, **27**(9):1368-1377.

25. Whitehurst AW, Bodemann BO, Cardenas J, Ferguson D, Girard L, Peyton M, Minna JD, Michnoff C, Hao W, Roth MG *et al*: **Synthetic lethal screen identification of chemosensitizer loci in cancer cells**. *Nature* 2007, **446**(7137):815-819.

26. Rottmann S, Wang Y, Nasoff M, Deveraux QL, Quon KC: **A TRAIL receptor-dependent synthetic lethal relationship between MYC activation and GSK3beta/FBW7 loss of function**. *Proc Natl Acad Sci U S A* 2005, **102**(42):15195-15200.

27. Zhang XD, Kuan PF, Ferrer M, Shu X, Liu YC, Gates A, Kunapuli P, Stec EM, Xu M, Marine SD *et al*: **The Use of Strictly Standardized Mean Difference for Hit Selection in Primary RNA Interference High-Throughput Screening Experiments**. *Journal of Biomolecular Screening* 2007, **12**(4):497-509.

28. Xiaohua Douglas Zhang PFK, Marc Ferrer, Xiaohua Shu, Yingxue C. Liu, Adam T. Gates, Priya Kunapuli, Erica M. Stec, Min Xu, Shane D. Marine, Daniel J. Holder, Berta Strulovici, Joseph F. Heyse and Amy S. Espeseth: **The Use of Strictly Standardized Mean Difference for Hit Selection in Primary RNA Interference High-Throughput Screening Experiments. J***ournal of Biomolecular Screening* 2007, **12**(4):13.

29. Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R: **Statistical practice in high-throughput screening data analysis**. *Nat Biotechnol* 2006, **24**(2):167-175.

30. Zhang XD: **A new method with flexible and balanced control of false negatives and false positives for hit selection in RNA interference high-throughput screening assays**. *J Biomol Screen* 2007, **12**(5):645-655.

31. Gou D, Narasaraju T, Chintagari NR, Jin N, Wang P, Liu L: **Gene silencing in alveolar type II cells using cell-specific promoter in vitro and in vivo**. *Nucleic Acids Res* 2004, **32**(17):e134.

32. Hsieh AC, Bo R, Manola J, Vazquez F, Bare O, Khvorova A, Scaringe S, Sellers WR: **A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens**. *Nucleic Acids Res* 2004, **32**(3):893-901.

33. Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, Heimerikx M, Kerkhoven RM, Madiredjo M, Nijkamp W, Weigelt B *et al*: **A large-scale RNAi screen in human cells identifies new components of the p53 pathway**. *Nature* 2004, **428**(6981):431-437.

34. Boutros M, Bras LP, Huber W: **Analysis of cell-based RNAi screens**. *Genome Biol* 2006, **7**(7):R66.

35. Chung N, Zhang XD, Kreamer A, Locco L, Kuan PF, Bartz S, Linsley PS, Ferrer M, Strulovici B: **Median absolute deviation to improve hit selection for genome-scale RNAi screens**. *J Biomol Screen* 2008, **13**(2):149-158.

36. Espeseth AS, Huang Q, Gates A, Xu M, Yu Y, Simon AJ, Shi XP, Zhang X, Hodor P, Stone DJ *et al*: **A genome wide analysis of ubiquitin ligases in APP processing identifies a novel regulator of BACE1 mRNA levels**. *Mol Cell Neurosci* 2006, **33**(3):227-235.

37. Haney SA: **Increasing the robustness and validity of RNAi screens**. *Pharmacogenomics* 2007, **8**(8):1037-1049.

38. Pelkmans L, Fava E, Grabner H, Hannus M, Habermann B, Krausz E, Zerial M: **Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis**. *Nature* 2005, **436**(7047):78-86.

39. Zhang XD, Yang XC, Chung N, Gates A, Stec E, Kunapuli P, Holder DJ, Ferrer M, Espeseth AS: **Robust statistical methods for hit selection in RNA interference high-throughput screening experiments**. *Pharmacogenomics* 2006, **7**(3):299-309.

40. Zhang XD, Kuan PF, Ferrer M, Shu X, Liu YC, Gates AT, Kunapuli P, Stec EM, Xu M, Marine SD *et al*: **Hit selection with false discovery rate control in genome-scale RNAi screens**. *Nucleic Acids Res* 2008, **36**(14):4667-4679.

41. Gunter B, Brideau C, Pikounis B, Liaw A: **Statistical and graphical methods for quality control determination of high-throughput screening data**. *J Biomol Screen* 2003, **8**(6):624-633.

42. Glivenko V: **Sulla determinazione empirica della legge di probabilita**. *Giorn Ist Ital Attuari* 1933, **4**:92-99.

43. Cantelli FP: **Sulla determinazione empirica delle leggi di probabilita**. *Giorn Ist Ital Attuari* 1933, **4**:221-424.

44. Donsker MD: **Justification and extension of Doob's heuristic approach to the Kolmogorov–Smirnov theorems**. *Annals of Mathematical Statistics* 1952, **23**:277–281.

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

46. Draghici S, Khatri P, Martins R, Ostermeier G, Krawetz S: **Global functional profiling of gene expression**. *Genomics* 2003, **81**:98 - 104.

47. Hosack D, Dennis G, Sherman B, Lane H, Lempicki R: **Identifying biological themes within lists of genes with EASE**. *Genome Biol* 2003, **4**:R70.

48. Zhang B, Schmoyer D, Kirov S, Snoddy J: **GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies**. *BMC Bioinformatics* 2004, **5**:16.

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

50. Vencio R, Shmulevich I: **ProbCD: enrichment analysis accounting for categorization uncertainty**. *BMC Bioinformatics* 2007, **8**:383.

51. Pavlidis P, Lewis D, Noble W: **Exploring gene expression data with class scores**. *Pac Symp Biocomput* 2002:474 - 485.

52. Tian L, Greenberg S, Kong S, Altschuler J, Kohane I, Park P: **Discovering statistically significant pathways in expression profiling studies**. *Proc Natl Acad Sci USA* 2005, **102**:13544 - 13549.

53. Smyth G: **Limma: linear models for microarray data**. *Bioinformatics and Computational Biology Solutions using R and Bioconductor* 2005:397 - 420.

54. Jiang Z, Gentleman R: **Extensions to gene set enrichment**. *Bioinformatics* 2007, **23**(3):306-313.

55. Gentleman R: **Category: using categories to model genomic data**. *Bioconductor Package Vignette* 2008.

56. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E *et al*: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nat Genet* 2003, **34**(3):267-273.

57. Virtaneva K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfield CD, de La Chapelle A, Krahe R: **Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics**. *Proc Natl Acad Sci U S A* 2001, **98**(3):1124-1129.

58. Zahn JM, Sonu R, Vogel H, Crane E, Mazan-Mamczarz K, Rabkin R, Davis RW, Becker KG, Owen AB, Kim SK: **Transcriptional profiling of aging in human muscle reveals a common aging signature**. *PLoS Genet* 2006, **2**(7):e115.

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

60. Keller A, Backes C, Lenhof H: **Computation of significance scores of unweighted gene set enrichment analyses**. *BMC Bioinformatics* 2007, **8**:290.

61. Dinu I, Potter J, Mueller T, Liu Q, Adewale A, Jhangri G, Einecke G, Famulski K, Halloran P, Yasui Y: **Improving gene set analysis of microarray data by SAM-GS**. *BMC Bioinformatics* 2007, **8**:242.

62. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci U S A* 2001, **98**(9):5116-5121.

63. Newton M, Quintana F, Boon J, Sengupta S, Ahlquist P: **Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis**. *Ann Appl Statist* 2007, **1**:85 - 106.

64. Goeman J, van de Geer S, de Kort F, van Houwelingen H: **A global test for groups of genes: testing association with a clinical outcome**. *Bioinformatics* 2004, **20**:93 - 99.

65. McCullagh P, Nelder JA: **Generalized linear models**, 2nd edn. London ; New York: Chapman and Hall; 1989.

66. le Cessie S, van Houwelingen HC: **Testing the fit of a regression model via score tests in random effects models**. *Biometrics* 1995, **51**(2):600-614.

67. Mansmann U, Meister R: **Testing differential gene expression in functional groups**. *Methods Inf Med* 2005, **44**(3):449 - 453.

68. Kong S, Pu W, Park P: **A multivariate approach for integrating genome-wide expression data and biological knowledge**. *Bioinformatics* 2006, **22**:2373 - 2380.

69. Rahnenfuhrer J, Domingues F, Maydt J, Lengauer T: **Calculating the statistical significance of changes in pathway activity from gene expression data**. *Statistical applications in genetics and molecular biology* 2004, **3**:Article16.

70. Edelman E, Porrello A, Guinney J, Balakumaran B, Bild A, Febbo P, Mukherjee S: **Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles**. *Bioinformatics* 2006, **22**(14):e108 - e116.

71. Lewin A, Grieve I: **Grouping gene ontology terms to improve the assessment of gene set enrichment in microarray data**. *BMC Bioinformatics* 2006, **7**:426.

72. Nacu S, Critchley-Thorne R, Lee P, Holmes S: **Gene expression network analysis and applications to immunology**. *Bioinformatics* 2007, **23**(7):850 - 858.

73. Adewale A, Dinu I, Potter J, Liu Q, Yasui Y: **Pathway analysis of microarray data via regression**. *J Comput Biol* 2008, **15**(3):269 - 277.

74. Lauter J, Horn F, Rosolowski M, Glimm E: **High-dimensional data analysis: selection of variables, data compression, and graphics - application to gene expression**. *Biometrical J* 2009, **51**.

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

76. Liu Q, Dinu I, Adewale A, Potter J, Yasui Y: **Comparative evaluation of gene-set analysis methods**. *BMC Bioinformatics* 2007, **8**:431.

77. Chen J, Lee T, Delongchamp R, Chen T, Tsai C: **Significance analysis of groups of genes in expression profiling studies**. *Bioinformatics* 2007, **23**:2104 - 2112.

78. Nam D, Kim S: **Gene-set approach for expression pattern analysis**. *Brief Bioinform* 2008, **9**:189 - 197.

79. Song S, Black M: **Microarray-based gene set analysis: a comparison of current methods**. *BMC Bioinformatics* 2008, **9**:502.

80. Dopazo J: **Formulating and testing hypotheses in functional genomics**. *Artif Intell Med* 2008.

81. Smyth GK, Speed T: **Normalization of cDNA microarray data**. *Methods* 2003, **31**(4):265-273.

82. Smyth GK: **Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor**. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Edited by R. Gentleman VC, S. Dudoit, R. Irizarry, W. Huber. New York: Springer; 2005: 397-420.

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

84. Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, Mao M, Li B, Cavet G, Linsley PS: **Expression profiling reveals off-target gene regulation by RNAi**. *Nat Biotechnol* 2003, **21**(6):635-637.

85. Opgen-Rhein R, Strimmer K: **Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach**. *Statistical applications in genetics and molecular biology* 2007, **6**:Article9.

86. Bradley Efron RT, Virginia Goss, Gil Chu: **microarrays and their use in a comparative experiment**. In.; 2000.

87. Ideker T, Thorsson V, Siegel AF, Hood LE: **Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data**. *J Comput Biol* 2000, **7**(6):805-817.

88. Wright GW, Simon RM: **A random variance model for detection of differential gene expression in small microarray experiments**. *Bioinformatics* 2003, **19**(18):2448-2455.

89. Wu B: **Differential gene expression detection using penalized linear regression models: the improved SAM statistics**. *Bioinformatics* 2005, **21**(8):1565-1571.

90. Newton MA, Noueiry A, Sarkar D, Ahlquist P: **Detecting differential gene expression with a semiparametric hierarchical mixture method**. *Biostatistics* 2004, **5**(2):155-176.

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

92. Lonnstedt I, Speed T: **Replicated microarray data**. *Stat Sinica* 2002, **12**(1):31-46.

93. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW: **On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data**. *J Comput Biol* 2001, **8**(1):37-52.

94. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Statistical applications in genetics and molecular biology* 2004, **3**:Article3.

95. Cui X, Hwang JT, Qiu J, Blades NJ, Churchill GA: **Improved statistical tests for differential gene expression by shrinking variance components estimates**. *Biostatistics* 2005, **6**(1):59-75.

96. Fox RJ, Dimmic MW: **A two-sample Bayesian t-test for microarray data**. *BMC Bioinformatics* 2006, **7**:126.

97. Cui X, Churchill GA: **Statistical tests for differential expression in cDNA microarray experiments**. *Genome Biol* 2003, **4**(4):210.

98. Box GEP, Tiao GC: **Bayesian inference in statistical analysis**, Wiley classics library edn. New York: Wiley; 1992.

99. Pratt JW, Raiffa H, Schlaifer R: **Introduction to statistical decision theory**. Cambridge, Mass. ; London: MIT Press; 1995.

100. Efron B, Tibshirani R: **On testing the significance of sets of genes**. In.; 2006.

101. Hollander M, Wolfe DA: **Nonparametric Statistical Methods**. New York: Wiley; 1999.

102. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I: **Controlling the false discovery rate in behavior genetics research**. *Behav Brain Res* 2001, **125**(1-2):279-284.

103. Reiner A, Yekutieli D, Benjamini Y: **Identifying differentially expressed genes using false discovery rate controlling procedures**. *Bioinformatics* 2003, **19**(3):368-375.

104. Wright SS: **Correlation and causation**. *Journal of Agricultural Research* 1921, **20**:29.

105. Simon H: **Causal ordering and identifiability**. In: *Studies in Econometric Method.* Edited by Hood WCK, T.C. New York: Wiley; 1953.

106. Pearl J: **Causality : models, reasoning, and inference**. Cambridge, U.K. ; New York: Cambridge University Press; 2000.

107. Wu X, Zhu W: **Latent SEM and repeated measures ANOVA**. In.: Stony Brook University; 2010.

108. Altundag O, Altundag K, Morandi P, Gunduz M: **Adjuvant targeted therapy with trastuzumab may decrease metastatic capacity in specific group of oropharyngeal cancer patients: downregulation of E-cadherin-catenin complex by cooperative effect of erbB-2 and human papillomavirus type 16 E6/E7 protooncogenes**. *Med Hypotheses* 2004, **63**(2):277-280.