# INTEGRATED GENOMIC MARKERS FOR CHEMOTHERAPEUTICS


**WU SONG**

M.Sc., High Performance Computation for Engineered Systems, Singapore-MIT Alliance, National Univ. of Singapore (2003)

B.Sc, Chemical Engineering, East China Univ. of Sci&Tech (2000)

Minor Major, Applied Mathematics, East China Univ. of Sci&Tech (2000)

**A THESIS SUBMITTED**

**FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**

**DEPARTMENT OF BIOLOGICAL SCIENCE**

**NATIONAL UNIVERSITY OF SINGAPORE**

**2011**

# Integrated Genomic Markers for Cancer Therapeutics

By

Wu Song

## Abstract

As translational research has created opportunity for an increasing number of anticancer agents, the need to develop computational methods to identify and understand predictive biomarkers has become emergent. This dissertation introduces a generic and systematic bioinformatics method to develop biomarker(s) for cancer therapeutics. The overall methodology includes the conceptualization of general types of biomarkers, implementation of algorithms, a uniqueness test of the signature markers in the test data using a novel computational algorithm and innovative bioinformatics algorithms to detect the presence of the signature with the pattern remained in the test data. An integrated genomic analysis to model gene expression and genomic aberrations is proposed to identify the minimal marker sets for clinical translation.

We then study a novel biological phenomenon in cancer therapeutics, that cancer cells may show concordant chemo-response to multiple anticancer agents. The representative preclinical models (both cell lines and primary tumor derived explants)

are selected to reflect concordant sensitive and concordant resistant tumor cells. Moreover, we developed the gene expression signature of concordant chemotherapeutics using NCI60 data to characterize the concordance of chemotherapeutics. A high predictive value (AUC = 0.88±0.10) is observed in an independent validation using Oncotest tumor clonogenic assay and gene expression data from primary xenograft tumor models. When the signature is applied to expression data from tumors of breast cancer patients treated with (TFAC) combination chemotherapy, the signature predictor predicts treatment outcome (pCR vs RD) with a p-value=0.017. We also find that the signature predictor is able to predict the survival of patients in breast cancer and lung cancer. Meta-analysis using Oncomine$^{TM}$ tools shows that more than 20 unique drug sensitivity concepts are significantly associated with the developed signature of concordant chemotherapeutics. These results demonstrate that concordance of chemotherapeutics is present in both preclinical models and clinical patients; the developed signature may have clinical utility for patients treated with standard of care chemotherapeutic agents in solid tumors.

In summary, we present innovative bioinformatics methods to develop genomic markers for cancer therapeutics and we identify a novel biological problem in cancer therapeutics using translational research methods.

**Thesis Supervisors: Greg Tucker-Kellogg**

# Acknowledgements

I owe a significant debt of gratitude to many people whom I would like to acknowledge for their contributions to my thinking, my writings, my career development, and my personal life.

First and foremost, I would like to acknowledge Dr. Greg Tucker-Kellogg, who provided me with strong support, invaluable advice and suggestions throughout my PhD study and my work at Lilly. I benefited greatly from his experience throughout my time at Lilly Systems Biology (LSB, 2005-2007) and Lilly Singapore Centre for Drug Discovery (LSCDD, 2007-2010). It is sad that LSCDD was closed and Systems Biology team was hence dismissed, but I did receive the best training and practice in this field. When Greg left Lilly, Prof. Paul Matsudaira and Greg are my NUS supervisors. From Prof Matsudaira, I was inspired by his wisdom, and I learned a lot about pursuing the scientific spirit from him.

I would also like to acknowledge Christoph Reinhard who is also my Lilly supervisor. He supported, mentored me and kept on encouraging me to finish my PhD within Lilly. I also want thank Jonathon Sedgwick and other LSCDD management members. Without their support, I would not have this chance to finish my PhD research while working for Lilly.

Amit Aggarwal, who worked with me closely in Lilly, had provided me many opportunities and excellent guidance, helping me develop my capabilities. He generously spent countless hours sitting down with me, even when I just turned up in

his office - unannounced. He has been most patient with me, explaining all kinds of concepts to me, especially, when I first entered this field. To him, I owe lots of thanks and beers.

In addition to Greg, Amit and Christoph, I would like thank a bunch of people in Lilly: Dan Li, Li Heng, Li Yue, Yu Kun, Kevin Gao, Jian Yan, Chen Peng, Julian Lee, Seppo Karina and Yang Feng, my wonderful colleagues.

Lastly, I have saved the most important acknowledgement for my family: my wife Lei and my son Leoson. I owe them too much! Their support has been unsurpassable and their love is unconditional. My deepest thanks go to my wife Lei who patiently encouraged me throughout my studies. I would also like to thank my parents and parents-in-laws for their courage and love.

# Table of Contents

# List of Tables

# List of Figures

# List of Acronyms

| | |
|---|---|
| SVD | Singular Value Decomposition |
| NMF | Non-Negative Matrix Factorization |
| PCA | Principle Component Analysis |
| ICA | Independent Components Analysis |
| FC | Fold Change |
| FDR | False Discovery Rate |
| CS | Concordant Sensitive |
| CR | Concordant Resistant |
| POS | Probability of Sensitivity |
| POA | Probability of Activity |
| AUC | Area Under Curve |
| pCR | Pathological Complete Response |
| RD | Residual Disease |
| NCI | National Cancer Institute |
| DTP | Drug Therapeutics Program |
| RECIST | Response Evaluation Criteria In Solid Tumors |
| FCM | Fuzzy c-means |

# Chapter 1 Introduction

## 1.1 Background

Cancers are diseases characterized by the uncontrolled growth and spread of abnormal cells. According to a World Health Organization report, over 10 million cancer cases occur annually, and deaths resulting from cancer worldwide are projected to continue rising to an estimated 11 million deaths by the year 2030. Although chemotherapy is a well-established therapeutic method for treatment of cancers, the clinical pathological response rate to the chemotherapy is usually low due to its limited efficacy and adverse effects. Resistance to chemotherapy is the most important factor contributing to the low response rate, and it remains a major obstacle in the treatment of cancer patients. Chemotherapy resistance occurs when cancers cells that have been responding to a therapy -- as evidenced by either growth delay or arrest -- begin to grow despite continued treatment. In other words, the cancer cells have acquired resistance to the effects of the chemotherapy, and the cancer treatment by way of chemotherapy from that point on is ineffective[1, 2].

The mechanisms of resistance of chemotherapy are very complex. Some cancer cells that are not killed by the chemotherapy may carry gene mutations that confer resistance to a drug. Some cancer cells may produce many copies of some particular genes, and then trigger an overproduction of protein that may render the anticancer drug ineffective. A well known resistance mechanism is when a

chemotherapeutic drug is pumped out of the cell by ATP transporter molecules, such as p-glycoprotein(ABCB1)[3]. These proteins are able to transport the drugs out of the cell, thus preventing drug activity. Some cancer cells may adapt to repair the DNA damage caused by some anticancer drugs. As a result, these cancer cells may be able to develop a mechanism to make the drugs ineffective[2].

Although cancer is highly complex and the disease etiologies are still unclear; we can now define, and characterize, different types of cancer cells by thousands of genetic aberrations, epigenetic changes, post-transcriptional modifications, and combinations of these mechanisms, rather than by site of origin using modern biology technologies. By unraveling these complexities and decoding cancer pathways, we hope to understand why some cancer cells are specifically resistant to a course of treatment, and why some cancer cells are highly sensitive to the same treatment. The choices of anticancer therapy for individual patients can thus be optimized by "translational research" methods. When using the terms "translational science" and "translational medicine"; we should define these terms carefully. We define these terms as effective translation of the new knowledge, mechanisms, and techniques generated in basic science research and clinical research into new applicable approaches for prevention, diagnosis, and treatment of disease and to provide better healthcare for patients[4, 5].

In order to thoroughly understand the cytogenetic and molecular alterations in cancer cells, various data types such as large-scale karyotype changes, sequence

alterations on protein-coding or regulatory regions, DNA copy number variations, epigenetic modification changes, mRNA, protein and microRNA expression are needed. Integrative analysis of these data can lead to a comprehensive molecular genetics and characterization of cancer cells. Apart from understanding the molecular genetics of cancer cells, we are also studying the activities of anticancer drugs in different types of cancer cells using comprehensive assays, such as proliferation and colony formation assays, which are widely used to screen chemo-response of anticancer drugs in both in-vitro and in-vivo models. For example, NCI-DTP (National Cancer Institute – Drug Therapeutics Protram) has screened more than 10,000 anticancer drugs in 60 cell lines using a Sulforhodamine B colorimetric assay, and 0 has screened more than 200 anticancer drugs in more than 100 human derived explants models using a tumor clonogenic assay[6, 7].

Not only do these measurement technologies enable us to observe more about cancer cells, but we also now have access to ever-increasing computational power to process the vast quantities of information, and assist in the identification and characterization the various pieces of the scientific puzzles. The developments here point to the need to form hypotheses from all these basic measurements of the cancer cells and anticancer drugs at an integrated systems level.

The work presented in this dissertation is an attempt to address three levels of systematic analysis: 1) the development of principled computational methods for developing biomarkers and generating hypothesis for cancer therapeutics; 2) the

discovery of the association between gene or protein expression and genetic aberrations using integrated analysis of different types of complex measurement; and 3) translation. There are two types of translation: a) translating the generated biological hypothesis into a clinical hypothesis and validating it using clinical data; and b) translating the learned biology during this process and forming clinical hypotheses for validation.

In this introductory chapter, we will briefly elaborate on some of the points we have alluded to above and provide a detailed context for understanding and motivating this work. In the later chapters, we will explore the various steps and describe how we understand and utilize different types of genomic measurements to develop genomic markers for chemotherapeutics, and how we decode complex biological data using mathematical and computational methods.

## 1.2 Principle of Scientific inquiry

The process of scientific inquiry is a repeated cycle of observation and explanation. When handling biological data, the earliest stages of the cycle sometimes consist of pure observation. The first step is to gather the raw material, out of which to create questions and then formulate the right questions and seek the answers.

Today, as we are in the genomic era, overwhelming biological data has been and is being generated like a continuing flood. After a period of observation, we may naturally begin to ask what kind of biological phenomena exist in the experimental

data that, and in some cases, we may notice different biological phenomena from what we have learned from textbooks. Followed by observations, critical dry bench analysis with mathematical and computational approaches and questions are formulated and possible explanations or hypotheses are postulated. Frequently, once integrated with multiple biological data, a number of hypotheses that are not consistent with the phenomena observed in different data can be rejected immediately; however, on the other hand, there may be a number of hypothesis that are consistent with the data and should be moved forward to the next stage. The next step is to gather more data for further analysis. Usually, we call this "validation", or at a minimum, "hypothesis testing". If the gathered data for validation is from same study, carefully-executed N-fold cross validation will give fairly robust validation results. But it is surprisingly easy to mistakenly "peek ahead" in the many steps of model building and testing. For that reason, the best standard for validation is when the gathered data is taken from another independent study. In the event that the hypotheses remain or pass the validation test, more experimentation and observation are necessary to distinguish between alternative explanations of the phenomena. In a biological study, if the hypothesis is generated from *in vitro* data, we strongly prefer furthering an *in-vivo* test before translating the hypothesis to the clinic. More on this will be discussed in later chapters. For use of predictions in appropriate and ethical human clinical studies, not only the *in vitro* validation, de-novo *in-vivo* test and

de-novo prediction in clinical data must be sound, all the experimental data generated and analyzed must also be based on robust statistical theory.

## 1.3 Genomic biology

The structure of DNA was discovered by James Watson and Francis Crick in 1953. Since that, scientists have started to decode human DNA by developing and exploiting an increasing understanding of DNA sequence technology and quickly applied this knowledge into drug development. Just as John Sulston thinks “Science is essentially a cultural and dynamic activity. It generates pure knowledge about ourselves and about the universe we live in, knowledge that continually reshapes our thinking”[8], the genomic understanding is the basement of today’s drug development.

A grand milestone of genome research effort was the "Human Genome Project". When first proposed, many scientific researchers argued that deciphering the human genome would lead to new understanding and benefits for human health. In 1990, these advocates won over detractors, and the Human Genome Project (HGP) was officially launched with funding from the US National Institutes of Health (NIH) and Department of Energy (DOE). Labs from all over the world collaborated with the NIH and DOE and resolved to sequence 95% of the DNA in human cells. In 2003, with heavy involvement from major partners Wellcome Trust (U.K.) and

contributions from Japan, France, Germany, China, and other countries, the Human Genome Project was successfully completed.

Today, vast quantities of genomic data are being generated throughout genomic biology, including DNA/RNA sequence data, mRNA/miRNA expression data, DNA methylation data, DNA copy number variation data and single polymorphism data, and protein expression data. The availability of genomic data will have a profound impact on biomedical research, diagnosis and therapeutic treatment. This also requires scientific researchers to advance systematic computational methods to better understand the data. In the following subsections, we will discuss different types of genomic data for context.

### 1.3.1 Genome sequencing

The DNA or RNA sequence is the primary structural description of a nucleic acid which composes of sequential nucleotides connected by chemical bonds. It can be written as a succession of letters representing the nucleotides of a DNA molecule or strand. By convention, the primary structure of a DNA or RNA molecule is reported from the 5' end to the 3' end. The sequence is considered to have capacity to carry information. The DNA genetic sequence carries the inherited information content of living functions.

Sequences can be reported by reading biological raw material through DNA sequencing methods. The principle objective of sequencing genomics is to determine the sequences of nucleotides that comprise the genomes of various living organisms.

In detail, sequencing genomics not only sequences the organism's chromosomes, it also identifies the organism's genes, introns and exons, proteins coding sequences and regulatory elements through genomic analyses. The Human Genome Project has completely sequenced the DNA sequence of *homo sapiens* in 2003. To date, the genomes of a large number of other organisms have also been sequenced, including panda, yeast, many types of bacteria and dozens of fishes and plants[9].

In past two decades, the Sanger method of sequencing by capillary electrophoresis using the ABI 3730xL(ABI Sanger) platform was widely used as the major solution for large-scale sequencing projects, and is recognized as the "gold standard" in terms of both read length and sequencing accuracy[10-13].

However, the high cost of Sanger sequence based method has greatly limited high-throughput sequence data generation. The increasing demand for low-cost sequencing is the key driver of the development of high-throughput sequencing technologies which dramatically parallelize the sequencing process and are able to produce thousands or millions of sequences at once. Several next generation sequencing (NGS) technologies have recently emerged, including Roche 454, Illumina Genomic Analyzer (GA), and Applied Biosystem (ABI) SOLiD, which are able to generate more than three to four orders of magnitude sequence and are considerably less expensive than the Sanger method on the ABI Sanger platform. To date, these new technologies have been successfully applied towards ChIP-sequencing to identify binding sites of DNA-associated proteins, RNA-sequencing to profile the

mammalian transcriptome, as well as whole human genome sequencing. Currently there is much interest in applying NGS platforms for targeted sequencing of specific candidate genes, intervals identified through single nucleotide polymorphism (SNP)-based association studies, or the entire human genome in large numbers of individuals[14-18].

Today, the low cost of high-throughput sequencing technologies is bringing the idea of “personalized medicine” closer to reality. The cost of sequencing a single whole genome has dropped to within several thousand US dollars. One day, full genome sequence data may allow healthcare researchers to investigate an individual’s entire genome and therefore detect all disease-related genetic variants, regardless of the genetic variant's prevalence or frequency. This will enable the rapidly emerging medical fields of “personalized medicine”, and will lead to a revolution in clinical genetics. Full genome sequencing is an important step towards better understanding the basis of genetic disease. However, it should be recognized that despite advancements in genome sequencing technology, incomplete understanding of the significance of individual variants or combinations of variants will limit the widespread usefulness of full genome sequencing in medicine until its clinical utility can be demonstrated [19].

### 1.3.2 DNA copy number variations

DNA copy number variations (CNV) are alterations of genomic DNA in which a certain region of the chromosome has been deleted or amplified. The size of

DNA copy number variation ranges from 1 kb to few mb in a single chromosome. Copy number variations include deletions, insertions, duplications and more complex variants, and have been shown to affect gene expression, phenotypic variation and adaptation. The importance of DNA copy number variation in the human genome has become increasingly apparent over the last few years. The study of genome-wide copy number variation has shown that SNPs will have to share their place in the spotlight when it comes to studies of human genetic variation, disease and population structure[20, 21].

DNA Copy Number Variation(CNV) is caused by genomic rearrangements such as deletions, duplications, inversions, and translocations. Segmental Duplications (SD) is the typical explanation of genomic rearrangements. Segmental duplications are operationally defined as >1 kb stretches of duplicated DNA with high sequence identity, for example, Low Copy Repeat (LCR) is a DNA genome region specific sequence repeat and is susceptible to result in DNA copy number variations. Any change between two copies of DNA sequence, for instance, size, orientation and percentage similarity or distance, is susceptible to change in LCR therefore leading to genomic rearrangement[22-24].

Copy Number Variation(CNV) can be discovered by cytogenetic techniques such as fluorescent in situ hybridization, comparative genomic hybridization, array comparative genomic hybridization, and by virtual karyotyping with SNP arrays. To evaluate gain or loss of specific human samples, Wellcome Trust Sanger Institute

researchers have created the human CNV project, which has generated the most complete map within the human genome of variation in copy number between healthy individuals. To date, the human CNV project has detected CNVs in the genomes of 270 individuals (the HapMap collection) with mixed ancestry of Europe, Africa and East Asia race (http://www.sanger.ac.uk/humgen/ cnv/).

### 1.3.3 DNA methylation

DNA methylation is a type of chemical modification of a DNA sequence. This modification can be inherited through cell division and it is subsequently removed without changing the original DNA sequence during zygote formation. It is one of the several types of epigenetic changes for DNA. DNA methylation typically occurs in the context of CpG (cytosine followed by guanine) dinucleotides. Localized regions of high CpG frequency (or CpG islands) are located around the promoters of the genes that are frequently expressed in cells. Methylated CpG sequence suppresses the corresponding genes' expression. Cytosine methylation is the major form of DNA methylation in many mammalians. The methylated cytosine can be converted to thymine by accidental deamination, and the methylated CpG sequence will be transformed into the TpG sequence, and making the gene inactive[25, 26].

There are several assays to discover DNA methylation, such as Methylation-Specific PCR (MSP) assay, the HELP assay, and Methylated DNA immunoprecipitation (MeDIP) assay. Methylation-Specific PCR assay is based on the chemical reaction of sodium bisulfite with DNA that converts unmethylated cytosines

of CpG dinucleotides to uracil or UpG, followed by traditional PCR experiment. The HELP assay is able to detect and cleave methylated and unmethylated CpG sites using restriction enzymes. Methylated DNA immunoprecipitation first isolates methylated DNA fragments, and then and puts the DNA fragments into DNA microarrays or DNA sequencing (MeDIP-seq)[27, 28].

Changes in the pattern of DNA methylation data have been identified consistently in cancer cells. In the past decades, DNA methylation was speculated to play an important role in the onset or course of cancer. Recently, various changes in the DNA methylation patterns or in DNA methyltransferase expression levels in cancer cells have been reported. These changes provide a direct and indirect link between DNA methylation and cancer cell proliferation. In particular, DNA methylation might play a critical role in oncogene and tumor suppress gene mutations[29-31].

### 1.3.4 Gene expression

Gene expression is the phenotypic expression of gene products. It means from transcription, through RNA processing to translation and post-translational modifications. Alternatively, gene expression refers to the process by which information, which is carried by a gene (DNA sequence), is translated to synthesize a functional gene product.

Regulation of gene expression is the very important in the majority of cellular activities. It gives the cell control over all structures and functions, such as cellular

differentiation, morphogenesis and organism development. Gene expression can be modulated, from the DNA-RNA transcription process to post-translational modification of a protein. To better understand the regulation of gene expression usually requires the exploitation of the genome profiling at DNA, RNA and protein levels. In the previous part of this chapter, we have discussed DNA sequencing data – gene mutation, DNA copy number variation data – gene copy numbers and DNA methylation data – gene methylation, the combined changes of these factors actively lead to the variation of gene expression and functional gene products, which are often proteins. In addition, small non-coding RNAs, such as microRNA, and various classes of short or long non-coding RNAs may be constantly involved in a variety of gene regulatory functions[32, 33].

Changes in gene expression underlie many biological phenomena. In cancer research, the study of the gene expression, especially the expressions of oncogenes and tumor suppress genes, is the critical step. Ideally, gene expression is measured by detecting the final gene product, usually the coded protein expression; however it is much easier to detect one of the precursors, typically mRNA, to infer levels of gene expression, especially in high throughput screening. Therefore, the expression pattern of a particular gene or set of genes, such as increases and decreases, measure the relative abundance of the gene specific mRNA transcript. In modern molecular biology, the high throughput screening of measuring thousands of genes concurrently

is called gene expression profiling, and it has been shown to be a powerful tool in cancer research[34-42].

To date, there are three major technologies have been adopted to profile gene expression: DNA microarray technology[43], serial analysis of gene expression (SAGE) and RNA-Seq technology[44-46]. Microarrays measure the relative activity of previously identified target genes by designing thousands of DNA oligonucleotides in the array. However, sequencing based techniques, like SAGE and SuperSAGE, are detecting the expression of level of full genomic transcripts instead of predefined set genes. Since 2006, the advent of next-generation sequencing techniques has made sequence based expression analysis more popular and more accurate when compared with microarrays. RNA-seq technology is also called “Whole Transcriptome Shotgun Sequencing” method. It is basically using high-throughput sequencing technologies to sequence cDNA in order to get information about a sample's RNA content. Till now, microarrays are still far more common because of its well validated reproducibility[47]. However, RNA-Seq is becoming widely used with reducing cost and the technology seems to be more reliable. What is important in this dissertation is that microarray technology allows comparison with a vast reference data set of published clinical gene expression data generated using the same method.

In last decade, the technologies of DNA olignucleotide based microarrays have advanced tremendously. Both Affymetrix and Illumina have developed arrays for different types of scientific research. In a microarray chip, the probes are attached

via surface engineering to a solid surface by a covalent bond to a chemical matrix such as epoxy-silane, amino-silane, lysine or polyacrylamide. The solid surface is usually glass, plastic or a silicon chip. In Affymetrix, the solid surface is the array itself, while in Illumina, the solid support is microscopic beads distributed onto the array. The quality of DNA microarray data may potentially become problematic due to false cross-hybridizations between the design probes and mRNAs[43, 48]. As we discussed in our general scientific research methodology, validation is always necessary when confirming the basic discovery results. A low-throughput but highly validated approach for measuring mRNA abundance is the reverse transcription quantitative polymerase chain reaction(RT-PCR), followed with real-time polymerase chain reaction (qPCR). RT-PCR generates a DNA template from the mRNA by reverse transcription (cDNA). This cDNA template is then used for qPCR, where the intensity of fluorescence on the probe changes as the DNA amplification process progresses[49].

### 1.3.5 MicroRNA expression

MicroRNAs or miRNAs are short ribonucleic acid molecules, and on average they are only 22 nucleotides long. MicroRNAs are usually found in eukaryotic cells. MicroRNA is one of the post-transcriptional regulators that bind to complementary sequences on target gene mRNA transcripts, and it usually plays a translational repressing role that silences gene expression. The human genome has roughly over 1000 microRNAs and they are predicted or validated to target about 60% of

mammalian genes. Some microRNAs are present upon human cell types[50, 51]. Some microRNAs have been shown to play either the role of oncogenes or tumor suppressing genes. For example, miR-17-5p and miR-20a are shown to mediate E2F1 pathway activity by transcriptionally repressing E2F1[52, 53].

Similar to mRNA expression, microRNA expression can be detected using microRNA microarrays. MicroRNA is hybridized to microarrays, slides or chips with probes to hundreds of microRNA targets, so that relative levels of microRNAs can be determined in different samples. More accurately, microRNA expression can be quantified with a polymerase chain reaction process of modified RT-PCR followed by quantitative real-time PCR. Variations of this method achieve absolute or relative quantification.

### 1.3.6 Protein expression

Proteins are the final product of gene expression system. In the cell machinery, proteins are the actual workers, and they are synthesized and regulated depending on the functional need in the cell. Protein expression refers to the way in which proteins are synthesized, modified, and regulated in living organisms. The blueprints for proteins are stored in DNA and decoded by highly regulated transcriptional processes to produce mRNA, and the mRNA is then translated into a protein. Transcription passes the information from DNA to mRNA, and translation is the synthesis of protein polypeptides based on a sequence specified by mRNA. After translation, protein polypeptides are modified in various ways to complete their structure,

designate their location or regulate their activity within the cell. Post-translational modifications are various additions or alterations to the chemical structure, and are critical features of the overall cell biology. There are a number of post-translational modification of proteins in the cell, which include: a) polypeptide folding into a globular protein to arrive at the lowest energy state; b) modifications of the amino acids; c) disulfide bridge formation or reduction; and d) protein modifications to facilitate binding functions, such as glycosylation and acetylation of histone to modify DNA-histone interactions[54, 55].

During the progression of cancer cells, many signaling proteins are activated through genetic, epigenetic and post-translational events[56-60]. The quantitative detection of proteins in cells and tissues to discover the expression patterns or changes in different conditions, such as health, disease, differentiation and drug treatment, is a central aim of proteomics research. The array format is well established for the rapid analysis of protein expression. There are three general types of protein arrays: large-scale functional chips, analytical capture arrays and lysate arrays. Large-scale functional chips immobilize large numbers of purified proteins and are able to assay a wide range of biochemical functions, such as protein-protein, protein-DNA, protein-small molecule interactions and enzyme activity, and are able to detect antibodies and demonstrate their specificity. Analytical capture arrays, also called antibody arrays, usually array antibodies, but may also use alternative protein scaffolds, peptides or nucleic acid aptamers. They are able to detect and quantify

analytes in complex mixtures such as plasma/serum or tissue extracts. Lysate arrays -- also called reverse phase proteins arrays – use cell samples, such as tissue lysates, printed on an array surface, and the protein targets are then detected with high quality antibodies overlaid on them[61, 62]. All three types of assays are widely used in diagnostics, clinical biomarkers and discovery research.

Beside the array techniques, another label-free detection method is mass spectrometry. Mass spectrometry sequences the amino acids in a protein and compares its amino acid sequence with known proteins. The amino acid sequence also can be used to predict the charge of the molecule, its size, and its probable three-dimensional structure. The charge and size is confirmed experimentally using SDS-PAGE and double-dimension gels. The three-dimensional structure of the protein is determined through X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR).

Proteins play a number of different roles within cells, and their interactions are the most important biological activities. In this dissertation, we are going to apply our developed computational methods for the protein expression data and to develop protein expression markers for cancer therapeutics. The integrated analysis with other genomic data, such as DNA copy number variation data, DNA methylation data and mRNA and miRNA expression, permits us to decode the resistance of chemotherapeutics and develop the potential biomarkers for clinical application.

### 1.3.7 Integrated analysis of genomic data

The first foray into analysis of genomics data are the phenotypic expression data, such as gene expression and protein expression. Challenges to the study of gene expression and protein expression data are very similar, including data normalization, data smoothing, correlating between genes or samples, clustering data, categorizing data and using hierarchical clustering methods, principle component analysis or singular value decomposition(SVD), or non-negative matrix factorization(NMF) to pull out patterns in the data. These enriched analysis methods have successfully identified a number of remarkable patterns in the phenotypic expression data[63-65].

However, these analyses based on pure phenotypic expression data, such as gene expression markers, lack concrete biological explanations of genetic changes, such as mutations, copy number variation or methylation, and regulations between transcripts or proteins and miRNAs. To better understand the developed phenotypic expression markers and apply them in clinical practice, it is necessary to decode the over or under expression for specific genes or proteins using the corresponding sequence genomics data, copy number variation, methylation data and miRNA data. Therefore, we propose a simple mathematical model to describe how genetic aberrations change the transcripts expression.

| | |
|---|---|
| $Expression\ (g_i) = f\left(Mutation\ (g_i), CNV\ (g_i), Methylatio\ n(g_i)\right)$ | Eq 1-1 |

Here $Expression(g_i)$ represents the gene or protein expression value of gene *i,* usually it is the normalized readouts from large scale screening array data;

$Mutation(g_i)$ is the mutation of the gene *i*, especially deletions or frame shift mutation; $CNV(g_i)$ is the DNA copy number variation value or calls for gene *i*; $Methylation(g_i)$ is the DNA methylation value for gene *i.* By using linear programming techniques, we can select the meaningful genes whose transcript expression is well explained by the corresponding genetic aberrations. However, this simple mathematical model does not consider the regional or distant genomic loci that regulate the expression of gene *i*. The gene or protein expression of gene *i* maybe regulated by the approximate location of the gene-of-origin, which is called cis or trans regulation. These genomic loci is usually called expression quantitative trait loci (eQTLs)[66]. To improve the proposed mathematical model, including the cis and trans regulations of the corresponding genomic loci is the direction, however, many eQTLs show tissue dependent variation[67], and this may set up a barrier in our analysis since we started the biomarker development with the pre-clinical models which include multiple cell types. If the biomarker development is focusing on specific tumor type, we highly suggest changing the mathematical models as:

$$Expression\ (g_i) = f\left(Mutation\ (g_i), CNV\ (g_i), SNP\ (g_i), Methylatio\ n(g_i)\right) \quad \text{Eq 1-2}$$

Here, $SNP(g_i)$ is the gene *i* associated SNP's variation.

## 1.4 Chemotherapeutics

Chemotherapy is a distinct treatment from surgery and radiation therapy in treating cancer. Rather than physically removing or destroying a tumor or a part of it,

chemotherapy uses anticancer drugs to interact with cancer cells to eradicate or control the growth of cancer cells. The anticancer drugs used in chemotherapy are able to reach most parts of the body. Therefore, chemotherapy is likely to be recommended for cancer that has already spread to other areas of the body, for tumors that have occurred at more than one site, or for tumors that cannot be removed surgically. It is also used when a patient has the recurrent disease after initial treatment with surgery or radiation therapy.

A single drug may be given or a combination of several different drugs may be given together, and this latter approach is called combination chemotherapy. The mechanism to kill or stop cancer cells growth is different for different anticancer drugs, and combination therapies exploit these differences. Cells divide by going through a cell cycle, following an ordered set of events that include the synthesis of the DNA (S-phase) and the mitosis (M-phase), when the cell is thereafter divided into two daughter cells. Normal cells grow and die in a precise and controlled way, but when cancer occurs, the cells growth process becomes abnormal, with cells dividing and forming more cells without control and order. In chemotherapy, anticancer drugs that interfere primarily with DNA synthesis and mitosis (the S and M phases of the cell cycle) are used to destroy cancer cells. Different drugs work through different mechanisms: some damage a cell's genetic material (DNA), while others prevent the cell from dividing.

Typically, there are two types of anticancer drugs: cytotoxic drugs and targeted drugs. Chemotherapy with cytotoxic drugs cannot distinguish between normal cells and cancer cells other than in gross features of replication; hence both types of cells are affected by chemotherapy. The toxicity of chemotherapeutic drugs to normal cells is the cause of unwanted side effects. Targeted drugs are intended to target the genetic lesions specific to cancer cells and to make cancer cells stop accumulating in tumor progression. Targeted drugs thus only target cancer cells and are intended to have much less effects on normal cells. Some targeted drugs show much better benefits than cytotoxic drugs for certain tumors in clinical use. However, the intrinsic value of chemotherapy lies in the fact that the killing effect of chemotherapeutic agents has a definite selectivity for cancer cells over normal host cells. Normal tissues are able to repair themselves and continue to grow, so the injury caused by chemotherapy is rarely permanent.

In general, cytotoxic agents can be classified into alkylating agents, antimetabolites, topoisomerases, vinca alkaloids and taxanes. Alkylating agents are so named because of their ability to alkylate many nucleophilic functional groups under conditions present in cells. Cisplatin and carboplatin, as well as oxaliplatin, are alkylating or alkylating like agents. They impair cell functions by forming covalent bonds with the amino, carboxyl, sulfhydryl, and phosphate groups in biologically important molecules. Other alkylating agents are mechlorethamine, cyclophosphamide, chlorambucil, and ifosfamide. They work by chemically

modifying a cell's DNA. Antimetabolites agents include 5-fluorouracil, mercaptopurine, trimethoprim, pyrimethamine and pemetrexed. Some of these agents replace natural substances as building blocks in DNA molecules, thereby altering the function of enzymes required for cell metabolism and protein synthesis. Antimetabolites of this sort are cell cycle specific, and are most effective during the S-phase of cell division because they primarily act upon cells undergoing synthesis of new DNA for formation of new cells. Topoisomerases are essential enzymes that maintain the topology of DNA. Inhibition of type I or type II topoisomerases interferes with both transcription and replication of DNA by upsetting proper DNA supercoiling. Type I topoisomerase inhibitors include Camptothecins: Irinotecan and Topotecan. Type II inhibitors include Amsacrine, Etoposide, Etoposide phosphate, and Teniposide. Vinca alkaloids such as Vincristine, Vinblastine, Vinorelbine and Vindesine, bind to specific sites on tubulin, inhibiting the assembly of tubulin into microtubules (M phase of the cell cycle). They are derived from the Madagascar periwinkle and Catharanthus roseus. Taxane is the natural product Paclitaxel, originally known as Taxol and first derived from the bark of the Pacific Yew tree. Docetaxel is a semi-synthetic analogue of Paclitaxel. Taxanes enhance stability of microtubules, preventing the separation of chromosomes during anaphase.

One of the most important decisions for an oncologist is to prescribe the right drug with the right amount of anticancer drugs to treat the cancer patients at the right time. In this dissertation, one of the main aims is to develop a systematic methodology

to identify molecular markers which may help to stratify patients groups for specific treatment using anticancer agents. An important phenomenon is that some cancer cells show sensitivity to almost all kinds of cytotoxic chemotherapeutic agents. While conversely, some cancer cells show resistance to many types of cytotoxic chemotherapeutic agents. In this dissertation, we will focus on developing genomic markers to elucidate the concordant chemotherapeutics using preclinical materials.

## 1.5 Biomarker discovery

### 1.5.1 Biomarker for cancer therapeutics

The development of molecular biological techniques for genetic analysis has led to a great increase of our knowledge of genomics in general, and specifically, to our understanding of the structure and behavior of cancer genomics. These molecular techniques are being used to study biomarkers to stratify cancer patients groups in cancer chemotherapeutics, and have been shown great potential to improve the quality of patients' lives. A 70-gene MammaPrint signature (Agendia Inc, Huntington Beach, CA) measures the gene expression profile of 70 genes and uses its expression pattern to predict the likelihood of distant metastases for early stage breast cancer. This is the first molecular marker approved by the US Food and Drug Administration (FDA). Another molecular marker, Oncotype DX (Genomic Health), uses its 21 candidate genes to estimate likelihood of recurrence. The gene expression signature is composed by ER and HER2, as well as ER-regulated genes and several proliferation-related

genes. Increasingly, such emerging molecular markers could influence clinical care[68-70].

Prognostic and predictive efficacy markers are the most important markers to guide the selection of the most appropriate chemotherapy for individual cancer patients. In the last decades, a lot of retrospective studies on many markers have been performed, but only few had been validated in prospective therapeutic trials or prospective studies from an accurately selected patient population. The predictive efficacy markers or response markers are used to predict the potential "responders" for specific chemotherapy or evaluate the probability of "sensitivity/response" for the individual patient to the chemotherapy. Prognostic markers, such as Agendia's MammaPrint signature and Genomic Health's Oncotype DX signature, are used to estimate the likely outcome of treatment, for instance, the recurrence of tumor growth after primary treatment for the cancer patients. Prognostic markers play a key role in clinical practice in distinguishing patients into different risk groups and providing guidance for doctors to design treatment strategies in the care for patients. For example, the amplification of MYCN proto oncogene is a known indicator of poor outcome in neuroblastoma patients, therefore, patients with MYCN amplication need more challenging clinical care[71].

Typically, there are two types of biomarkers in clinical applications: prognostic markers and efficacy markers. Prognostic markers may be generic and can predict the response of multiple chemotherapies; while efficacy markers may be

specific and it can only predict one or a combination of anticancer agent(s). Although prognostic biomarkers that provide information on the natural course of disease after standard treatments are useful, predictive biomarkers are of greater value in clinical decision making and will be essential tools to provide the tailoring treatments for cancer patietns. The measurement of efficacy of anticancer agents in solid tumors is based on pathological response using the RECIST criteria, categorized as complete pathological response, partial pathological response, stable disease or progression. The measurement of general prognosis, however, is based on patients' survival, such as progression free survival, disease free survival and overall survival. Patient survival is highly correlated with the patient's chemotherapeutic response to anticancer agent(s), if chemotherapy was chosen as the primary care for a cancer patient. The aim of our research is to develop molecular markers to stratify cancer patients for standard of care chemotherapies. In this dissertation, we will not specifically discuss whether the developed markers have prognostic value, or if they are purely efficacy markers.

Tumor response to chemotherapy varies from one patient to another. It would be extremely useful to know ahead of time whether tumor cells of an individual patient would respond to chemotherapeutic agents, or whether an individual patient would show resistance to the chemotherapy. There are three types of molecular markers that provide guidance for chemotherapy treatment on clinical practice: a) the targeted genes; b) the activity of the targeted pathways; c) the genes are indirectly

related with the agent's or targeted pathway. The drug targets are the most important markers for the targeted anticancer agent. The clinical response rate of Gleevec is highly dependent on the targeted genes, mutation of BCR-ABL gene and over expression of C-KIT. Patients with HER2 amplification show 20% higher response rate than patients with HER2 normal copy number to the Herceptin[72, 73]. The activity of the targeted pathway is also the key marker for the drug's clinical response. For example, a patient with K-Ras mutation may show much worse response than a patient with wild type K-Ras to Panitumumab, which is a fully human monoclonal antibody specific to the epidermal growth factor receptor (EGFR). This is because a patient with a K-Ras mutation will lead to constitutively active downstream signaling of the pathway unaffected by the drug target[74, 75]. Apart from drug targeted genes and targeted pathways, there are also some molecular markers which are not directly related with the targeted pathways for example, ABCB1, TOP2A, ERBB2, and BCL2 are candidates of predictive markers to predict the chemosensitivity for cytotoxic chemotherapy in breast cancer[76].

### 1.5.2 Biomarker development for cancer therapeutics

The current primary focus of the translational research is to improve clinical outcomes by utilizing clinical validated biomarkers. According to the guidance from FDA pharmacogenomics, biomarkers are classified into three types: a) exploratory marker; b) probable validated marker and c) validated marker (FDA report 2004). An exploratory marker is more like a clinical hypothesis, for example, Taxane shows

resistance in the patients with β-tubulin mutations. A probable validated marker means that marker has been validated by some biological or clinical data, for example, Taxane shows higher IC50 in cell lines with β-tubulin mutations than cell lines without β-tubulin mutations. A validated marker means that a probable validated marker is confirmed in clinical trials, and has shown usefulness in improving clinical outcomes.

Remarkable advances in the understanding of neoplastic progression at the cellular and molecular levels have spurred interest in molecularly targeted cancer therapeutics. New imaging and bioassay technologies are providing the basis for developing biomarkers that will facilitate development of these molecularly targeted drugs. Biomarkers may be used in early drug development to elucidate the mechanism of action of a drug and provide preliminary evidence of its effect. As the relationship between a drug or class of drugs and a biomarker becomes better understood, there is a hope that clinical assays can be developed to identify patients most likely to benefit from the drug. These biomarkers are termed predictive biomarkers. Although prognostic biomarkers that provide information on the natural course of disease after standard treatments are useful, predictive biomarkers are of greater value in clinical decision making and will be essential tools for tailoring treatments.

NCI, FDA and drug makers have consensus about the facing challenges in drug and biomarker co-development in an seminar discussion[77]. The critical issue to develop both drug and corresponding biomarkers are highly depended on the level

of understanding of the biology of the drug target and its interaction with the drug. This means to understand the underlying biology of the drug target and the mechanism of action of the drug is the still the key to speed up the process of drug and biomarker co-development. Therefore, the pre-clinical models are the good starting point for the biomarker development, especially to study the biology of the drug target and the mechanism of action of the drug, and to discover the probable marker and to develop and to evaluate biomarker assay performance.

In the recent few years, remarkable advances in the understanding of cancerous progression at the cellular and molecular levels have spurred interest in oncogene or oncogene like addicted molecular targeted therapeutics. The latest news is the approvement of Zelboraf by FDA[78]. Zelboraf, which is specifically inhibiting mutated B-Raf protein, offers significant survival benefit in metastatic melanoma patients. The BRAF protein is involved in cell signaling pathway and promotes cell proliferation if over active. It is mutated in 50 percent of late-stage melanoma patients. Since Zelboraf selectively inhibits the mutated BRAF V600 protein, therefore, patients with B-Raf V600 mutation will likely show response to Zelboraf. Roche also developed Cobas 4800 BRAF V600 mutation assay to test the mutation in clinical patients. However, other than Zelboraf, there are many other targeted drugs are not “lucky” to have 50 percent of mutations for their drug targets in patients, and the tumor cells in patients may not show apparent tumor growth addiction on the drug targets as well. Therefore, the development of biomarkers for these drugs is also very

challenging, especially the number of available pre-clinical models is limited. What is needed here is the sophisticated methodology to identify the potential markers based on small size pre-clinical samples. Advanced bioinformatics analysis with new genomic and bioassay technologies are providing the basis for developing biomarkers that will facilitate development of these molecularly targeted drugs. In this dissertation, we will try to propose a systematic methodology to develop probably validated biomarkers for anticancer agents using a model driven analysis of genomic data and chemo-sensitivity data. Although there are some pioneer bioinformatics works done in the past decade [70, 79] [80], there is no systematic procedure to develop the probable validated markers for most standard of care chemotherapeutics and targeted agents.

Data driven analysis methodologies have been useful in uncovering interesting patterns to form exploratory hypothesis in the biological and clinical data. We use the following framework to integrate genomic data and chemo-response data in order to develop and validate biomarkers for cancer therapeutics.

The developed probable markers may be used in early drug development or for further academic research to elucidate the mechanism of action of a drug and provide preliminary evidence of its effect. As the relationship between a drug or class of drugs and the marker becomes better understood, there is hope that clinical assays can be developed and move to clinical for further validation and to identify patients most likely to benefit from the drug.

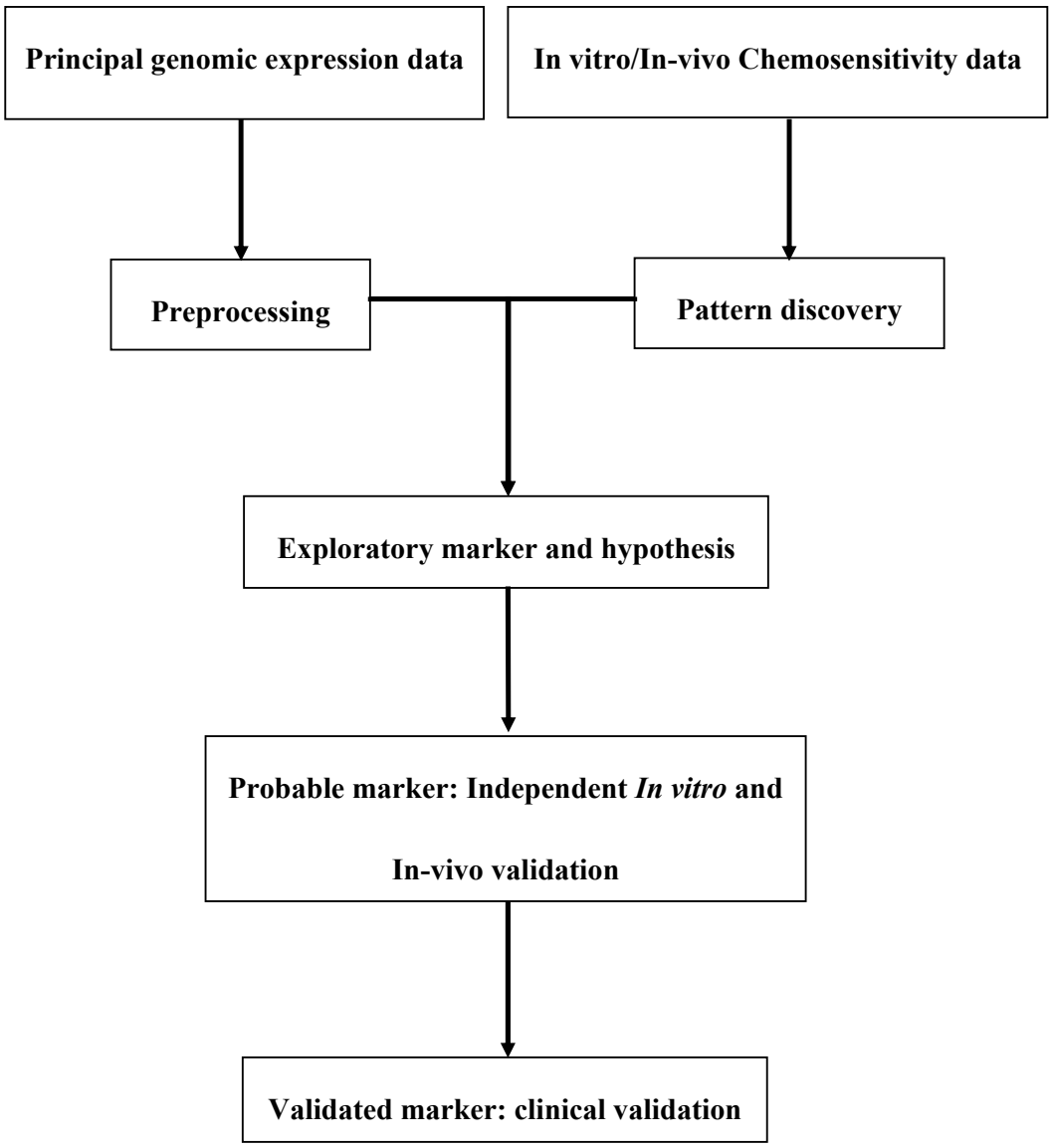



**Figure 1-1:** Framework to develop an exploratory marker, a probable validated marker and a validated marker for anticancer agents.

## 1.6 Combination chemotherapy

Combination chemotherapy is the use of more than one drug or therapy for cancer treatment in a patient. For some cancers, the best treatment strategy is a combination of surgery, radiation therapy, and chemotherapy. Sometimes

combination chemotherapy is used. Anticancer agents like cytotoxic drugs that are used to treat different cancers have variable outcomes. For example, only about 20% of patients with breast cancer respond therapeutically to the widely used drug Taxane (Paclitaxel or Docetaxel). Chemotherapy agents, especially cytotoxic drugs, often cause unwanted side-effects. Cytotoxic drugs work by killing cells which are dividing, and so some normal cells are consequently damaged too. Since the response rate is low and toxicity is high, combinations of a cytotoxic drug and other chemotherapeutic agents have been developed. The rationale for combination chemotherapy is to use drugs that work by different mechanisms of action, thereby decreasing the likelihood of developing resistant cancer cells. When drugs with different effects are combined, each drug can be used at its optimal dose without intolerable side effects. There are three types of measureable effects in combinational chemotherapy: synergistic, antagonistic and additive. If the combined agents show beneficial effect(s) to each combined agent, the combinational effect is considered "synergistic"; if the combined agents do not show favorable effects to one other, the combinational effect is considered "antagonistic"; and if the combined agents show similar effects as individual drugs, the combinational effect is considered "additive". Mathematically, this is often evaluated by the use of a "Combination Index (CI)"[81]. In another situation, even the synergistic drug combination is identified, the toxicity of the drug combination will also need to be investigated. Usually, the combined drugs are treated to the patients sequentially to avoid the accumulation of the toxicity.

The success of many targeted cancer therapies are based on their efficacy only when they are combined with cytotoxic agents in chemotherapeutics. For example, bevacizumab did not show any survival benefit as a monotherapy for patients with metastatic colorectal cancer, but it gave extra 2.5-month survival advantage when used in combination with the FOLFOX4 (Oxaliplatin, Leucovorin and Fluorouracil) chemotherapy regimen[82]. These success stories of the combination of targeted drugs with cytotoxic agents based chemotherapy has led to the hypothesis that efficacy of traditional chemotherapies could be enhanced by incorporating with targeted agents. However the mechanisms of such combinations are still unclear. It would be very helpful to provide scientifically based rationales for drug combinations between targeted agents and cytotoxic agents. We hope to extend the computational framework of the evaluation of single agent biomarkers into the evaluation of combinations of biomarkers. With the focus on optimizing dose regimens and understanding of developed biomarkers, it is possible to stratify patients into subpopulations who could benefit from such combinations.

## 1.7 Dissertation roadmap

In this dissertation, we will first propose a computational framework to study biomarkers to guide cancer chemotherapeutics. In particular, we will present the bioinformatics method to generate the principal markers based on gene or protein expression data and anticancer agents' *in vitro* chemo-response data. Next, we suggest

a novel computational method to test the uniqueness of the developed marker genes in the test data. Finally, we will further provide a robust method to detect the presence of the developed signature marker in the test data. By integrating mutation data, copy number variation data and methylation data, we can identify genes which are biologically meaningful and present strategies to narrow down the multiple gene signature into N(N<10) number of marker set for clinical translation. Another major part of the dissertation is to use translational research method to study novel biological problem in cancer therapeutics.

Overall, there are three major objectives for this work:

1. to develop a systematic bioinformatics methodology for studying molecular markers for cancer chemotherapeutics, and
2. to propose minimal marker set for clinical validation and practice, and
3. to study novel biology by using translational research methods

The outline of the dissertation is as follows: In this chapter, the backgrounds of genomic biology, cancer chemotherapeutics, biomarkers and their application to improve chemotherapeutics are introduced. In Chapter 2, bioinformatics methods to develop principal markers based on principle expression data will be presented and the methods to test the robustness of the developed expression marker and to detect of the presence of expression marker in test dataset will be proposed. In Chapter 3, a method to integrate different types of complex genomics data, such gene expression, mutation, copy number variation and method methylation data and develop robust single marker

will be presented and strategies to propose minimal marker set to clinical validation will be discussed. In Chapter 4, a novel biological problem, concordant chemotherapeutics will be identified, and the corresponding molecular markers will be generated to characterize the concordant response among chemotherapeutics. The developed genomic signature is then validated in both *in vitro* data and clinical data. In the last chapter, we will summarize our study and contributions to the field.

# Chapter 2 A systematic bioinformatics methodology to develop principal markers

Gene expression and protein expression are the phenotypic expression of gene products. Ideally, a group of genes or proteins with a combined expression pattern could uniquely characterize the condition of phenotype of biological species, such as disease state of cancer patients and chemo-response of anticancer agents. Therefore, gene or protein expression signature can serve as a principle surrogate marker for the studies of molecular phenotype, pathology, prognosis and diagnosis of cancers. For example, gene expression profiling using microarrays has been successfully applied for the classification of tumor types, stages of tumor progression, prediction of clinical outcomes and prediction of the response of anticancer agents[83-98].

The biological implication of the gene expression signatures is intrinsic, and the biological connection between genes identified by microarray and their phenotypic effect usually remains elusive. A signature gene set contains some false positives by nature of high throughput. Current advances in controlling the false discovery rate have overcome this problem to some extent, solidifying the status of expression profiling as the gold standard among non-biased genome wide approaches. In this chapter, we will review a few key technologies adopted to develop principle markers based on gene or protein expression arrays. The focus on of this chapter is to develop an innovative systematic methodology to discover biomarkers.

## 2.1 Introduction

In the last 10 years, microarray technology has been greatly advanced and has substantially gone from obscurity to being almost ubiquitous among biologists. Biologists today run high-throughput genomic studies by simultaneously measuring the expression levels of tens of thousands of genes in their biomedical research. One of the major applications is to use microarrays to discover differentially expressed genes between two or more groups, such as normal versus cancer patients, responders versus non-responder, control versus and drug treated. These identified differentially expressed genes may represent disease biomarkers in the diagnosis of the different types and subtypes of diseases or in the efficacy markers of anticancer agents in chemotherapy. The objective of principle marker development is to identify genes or proteins which show statistically significant up-and-down expression patterns in two groups of samples (Figure 2-1). In microarray derived gene expression parlance, it is usually named as gene expression signature.

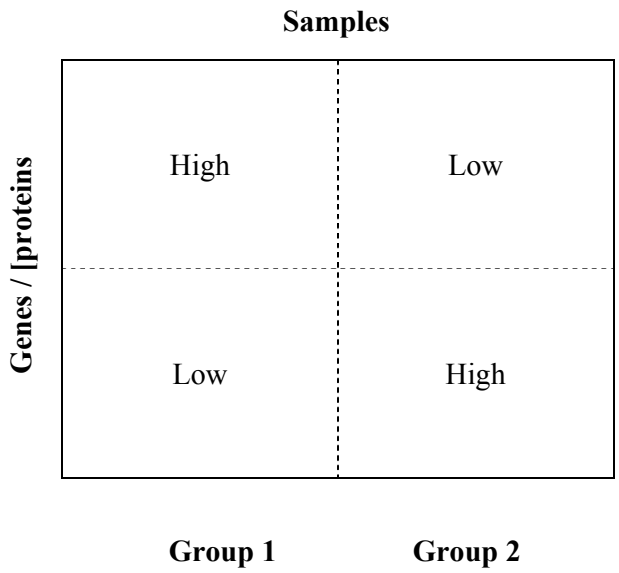



**Figure 2-1:** Schematic data matrix of the principle expression signature

The most straight forward approach to identify differentially expressed genes is known as the "fold-change" (FC) and it is calculated as

| | |
|---|---|
| $FC = \bar{u}_1 - \bar{u}_2$ | Eq 2-1 |

Where, $\bar{u}1$ and $\bar{u}2$ are typically the means of log-transformed gene expression in group 1 and group 2.

The FC method simply evaluates the average log-ratio between two groups, and considers the gene differentially expressed if the log-ratio difference is greater than a specified cut-off. The FC method represents the "up" and "down", or "high" and "low", or "over" and "under", or "on" and "off" of the gene's expression. This is the preliminary criteria for a gene to be a biomarker in clinical. However, the FC method lacks solid statistical footing because it assumes that the variance of the gene expression in two groups is equivalent. This assumption is especially problematic since variability in gene expression measurements is not uniform, even after the

variance has been stabilized by data transformation. Statistically, using this FC method alone with a fixed cutoff gives undetermined type 1 error rates.

Rather than applying a FC cutoff alone, a statistical test that incorporates variances of gene expression should be preferred. The student t-test is certainly the most popular test and has been considered as the fundamental method for differential gene analysis, especially for testing significant changes in small samples. The null hypothesis of the t-test is H0: $\overline{u}_1 = \overline{u}_2$ and the test statistic is given as

$$t = \frac{\overline{u}_1 - \overline{u}_2}{S}, \quad S = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \qquad \text{Eq 2-2}$$

Where, $n_1$, $n_2$ are the number of samples in group 1 and group 2 respectively, and $S_1^2$ and $S_2^2$ are the unbiased estimator of the variance of the expressions of the gene in group 1 samples and group 2 samples[99].

Another well known t-test based method is the Significance Analysis for Microarrays (SAM)[100, 101]. Estimating t-statistic could be problematic because the standardized variance can be skewed by low variation genes, which are false positives as their t-statistics are very large. SAM uses a modified t-statistics of the form

$$t = \frac{\overline{u}_1 - \overline{u}_2}{S + S_O} \qquad \text{Eq 2-3}$$

Where, $S$ denotes the pooled standard deviation of both group samples, and $S_O$ is a small constant for stabilizing the standard deviation.

The statistical power of t-test statistics is usually small when the sample size is small. Baldi and Long[102, 103] highlighted this problem by way of showing how

estimates of sample variances are poor when the sample size is small. Consequently, they introduced a Bayesian framework to evaluate the variance to perform standard t-test. The variance is estimated by the formula

$$\sigma^2 = \frac{w_0 \cdot \sigma_0^2 + (n-1) \cdot s^2}{w_0 + n - 2} \quad \text{Eq 2-4}$$

Where, $\sigma_0^2$ is denoted as a background variance, $w_0$ is defined as the weighted parameter and $s^2$ is the empirical sample variance. The weighted parameter $w_0$ is interpreted as a measure of confidence in the Bayesian estimate of the variance in comparison to the sample variance. The overall method is named as Cyber-T method.

Another variance modeling method has been proposed by Delmar and his colleagues[104]. The mixtures of distributions are employed to improve the estimate of the variance. The variance $\sigma^2$ is modeled as a weighted mixture of Gamma distributions, and the parameters of the mixture model are estimated from the observed data by expectation maximization (EM) approach.

Smyth[105] presented a different method that was based on general linear models, and named it as Limma. Limma is not restricted for two classes comparisons because it basically use the generalized linear model to fit the expression data for each gene

$$y = a \cdot x + b \quad \text{Eq 2-5}$$

Where, $y$ denotes the gene expression data, the log-transformed normalized gene expression array data, and $x$ represents the experiment design matrix or phenotype data, such as treatment and control, while b is the intercept vector. The subsequent

analysis is based on the fitted model parameters. Limma can not only identify significantly differential genes between two groups of samples, but can also be used for the analysis of more groups, factorial designs and time course experiments.

The statistical significance of the differential expression analyzed by testing each gene, multiple hypothesis testing is then an immediate concern. When multiple hypotheses are tested, the probability that a type I error is dedicated increases sharply with the number of hypotheses. Considering that thousands of genes can be analyzed in a single experiment, this may dramatically intensify the problem. (While controlling the family-wise type I error rate (FWE), which is the probability of one error in the family of hypotheses is needed. Benjamini and Hochberg[106] pointed out that the estimate proportion of the errors among the identified differentially expressed genes may be more appropriate. They proposed a concept “false discovery rate (FDR)” which is actually the expectation of FWE of the identified significantly differentially expressed genes. FDR criterion in the simultaneous testing of gene expression has been shown to be more powerful procedures[107-109].

All the above procedures of differential gene expression analysis can be used to analyze protein array data. Our ultimate goal in developing gene or protein expression markers is to translate these microarrays or protein arrays profiles into clinical practice and use them to guide the treatment of cancer patients or stratify patients for specific chemotherapeutics. In clinical practice, chemotherapeutics used against cancer may get a range of responses, such as complete response, partial response, stable disease response and progression disease response. Therefore, we

cannot restrict our study of differentiation markers analysis to one or two class samples. Hence, we will need to extend the analysis from two classes to three classes and more.

## 2.2 Fuzzy classification of biological data

Biological data is inherently uncertain and noisy; when handling uncertain biological data, it is difficult to separate measurement errors from inherent variability. A Boolean Network (BN) model, which simply considers gene expression as “on” or “off”, has been used to model gene regulatory networks and discover the gene expression patterns[110-114]. In the Boolean network formalism, a gene is considered to be either expressed or unexpressed, so intermediate expression levels are neglected. In reality, a gene can be expressed at intermediate levels, and to model these cases we need an alternative model to cover them. A “Fuzzy” clustering method can deal with such situations[115, 116]. A Fuzzy approach provides a systematic and unbiased way to imitate human intelligence by using qualitative descriptors such as “high” or “hot” to reduce the complexity of the natural characteristics of the data. In this study, we cluster the chemo-response data, such as GI50s and IC50s, into three fuzzy classes: “sensitive”, “medium” and “resistant”. The clustering method uses formal data discretization algorithms. For example, Figure 2-2 describes the classification of NCI60 cell lines based on the chemo-response data NLogGI50 (Negative log-transformed GI50) of Paclitaxel. 52 solid tumor cell lines are classified into three fuzzy clusters (sensitive, medium and resistant) using fuzzy c-means (FCM) method

with the three centers: Sensitive is 8.5; Medium is 7.5 and Resistant is 6.5. Samples with NLogGI50 greater than 8 are classified as “sensitive”; samples with NLogGI50 less than 7 are classified as “resistant”; and samples with NLogGI50 between 7 and 8 are classified as “medium”.

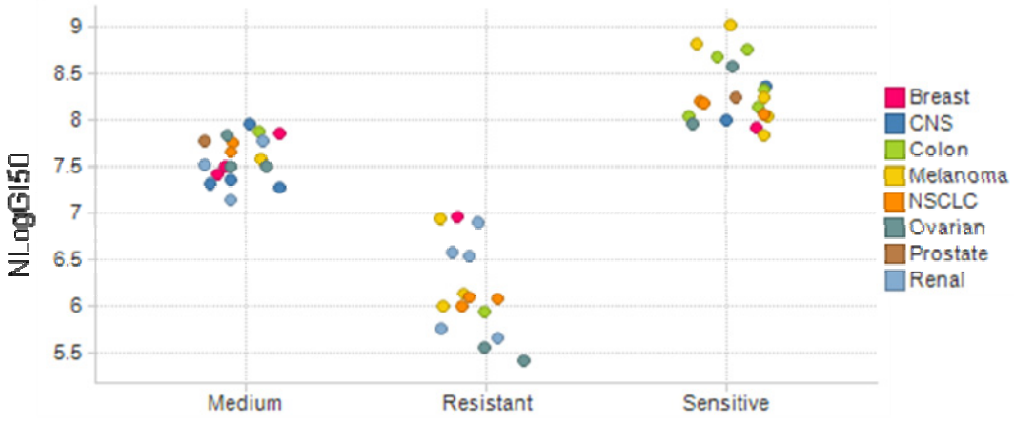



**Figure 2-2:** Classification of 52 NCI60 solid tumor cell lines based on NLogGI50 readouts of Paclitaxel.

## 2.3 Signature markers development

Fuzzy classification methods may cluster samples into three or more than three classes, such as “low”, “medium” and “high” for gene or protein expression; and “sensitive”, “medium” and “resistant” for chemo-response data. The statistical methods to develop significantly differentiated signature markers between two classes, such as “sensitive” and “resistant” samples, are state-of-the-art and have been well studied, but few methods have been presented to develop signature markers for three or more classes. Therefore, in this dissertation, one of the major tasks is to propose a

computational methodology that develops principal genomic markers, including the development of signature markers for samples with three or more classes.

Since this study is focused on developing biomarkers for chemotherapeutic agents, we have chosen to develop chemo-response signature markers for a start. Figure 2-3 depicts the typical expression pattern of signature markers developed based on two classes samples: markers(sensitive) show high expression in sensitive samples but low expression in resistant samples; markers(resistant) show low expression in sensitive samples but high expression in resistant samples. Many methods reviewed above, like SAM and Limma procedures, are very powerful in developing significantly differentially expressed markers. Table 2-1 lists the implementation of R scripts using Limma procedure to identify the top 300 significantly differentially expressed gene markers in microarray data. This implementation also includes the pre-processing steps of microarray data genes that expressed background level expression and do not vary significantly across all samples are excluded in the study. The R packages utilized in the following R scripts are “Limma”(version 3.6.9) and “genefilter”(version 1.32).

```
// ExprArray: microarray expression data, matrix format
// ResponseClass: response class information, numeric vector
library(genefilter)
## filter out the probesets with noisy level expression
f1 = pOverA(0.25, log2(100))
## filter out the probesets with insignificant variation expression
f2 = function(x) (IQR(x) > 0.5)
Fun.Filter = filterfun(f1, f2)
selected = genefilter(ExprArray, Fun.Filter)
ExprArray.PreProcessed = ExprArray[selected, ]
library(Limma)
design = cbind(mean = 1, diff = as.numeric(ResponseClass))
fit   = lmFit(ExprArray.PreProcessed, design)
fit2 = eBayes(fit)
topGenes = topTable(fit2,coef="diff",number=300,adjust.method="BH", sort.by="logFC")
Expr.SigDiff   = ExprArray.PreProcessed[topGenes$ID,]
Feature.Expr = cbind(topGenes[,c("logFC","adj.P.Val")], Expr.SigDiff)
rownames(Feature.Expr) =   topGenes[,"ID"]
```

**Table 2-1:** R scripts using Limma procedure to identify the most significantly differentiated genes between two class samples

However, this Table 2-1 only identifies the gene markers that have just two levels of expression: high and low in this pattern. When we extend the current two classes to three classes by considering “medium” samples, we obtain two types of expression patterns: type-I and type-II. In type-I, the expression of the marker has three categorical levels: high, medium and low. As shown in Figure 2-4, both markers(sensitive) and markers(resistant) keep the same pattern in both sensitive and resistant samples, but they show “medium” expression in the newly added “medium” samples. In type-II, the expression of the marker has two levels: high and low, therefore, a new group of markers, markers(medium), is identified to show high expression in medium samples and low expression in both sensitive and resistant samples (Figure 2-5). In this case, each class samples, either sensitive, medium or

resistant, have the corresponding representative markers, which are named as "markers(class name)". These markers only show "high" or "active" expression in these class samples, but show "low" or "silent" expression in all other classes samples. Therefore, the generalized expression patterns of the signature markers for "N classes" is defined as: a) type-I markers' expression is significantly correlated with the rank of chemo-response data; b) type-II markers is the combination of over expression markers or under expression markers which are only significantly "over" or "under" expressed in the specific class i (i=1, 2, …, N) samples but significantly "under" or "over" expressed in all other classes (1, 2, i-1, i+1, …, N) samples. For our interests, we have focused on the over expression markers.

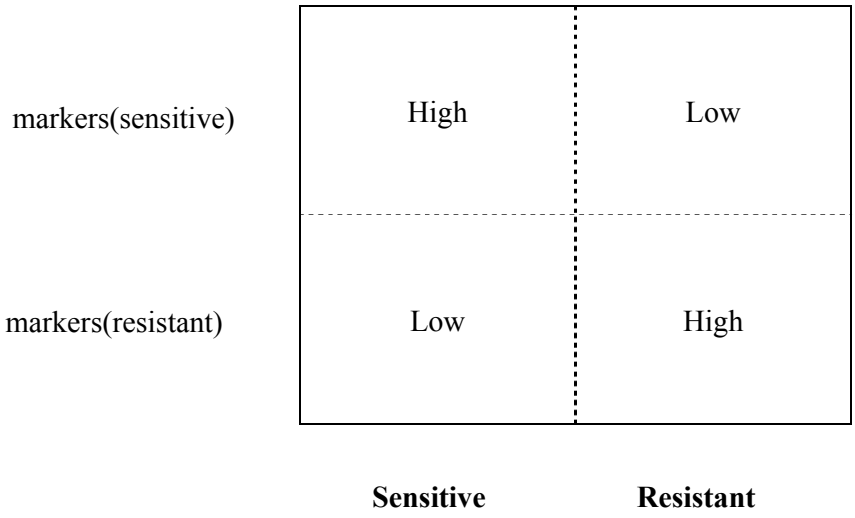



**Figure 2-3:** The expression pattern of the chemo-response signature markers for sensitive and resistant samples

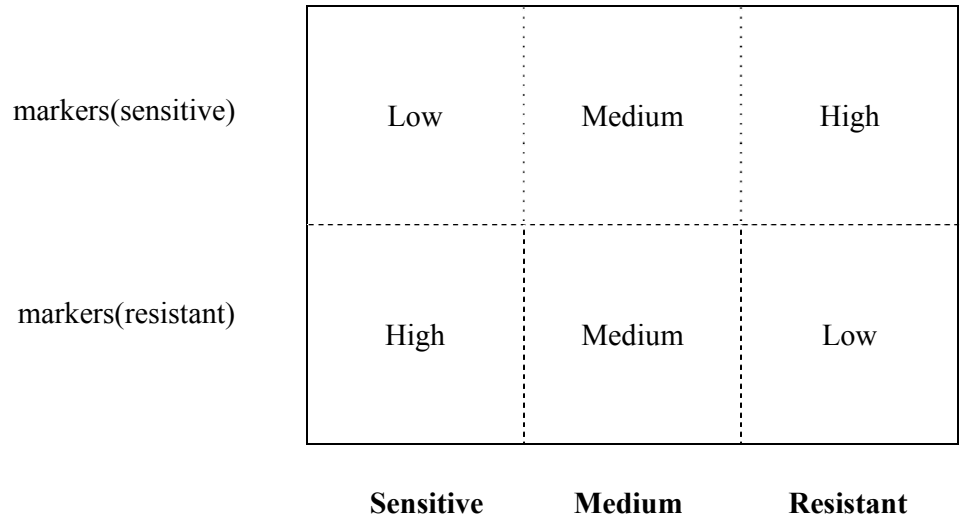



**Figure 2-4**: The expression pattern of type-I chemo-response makers for three classes: sensitive, medium and resistant samples

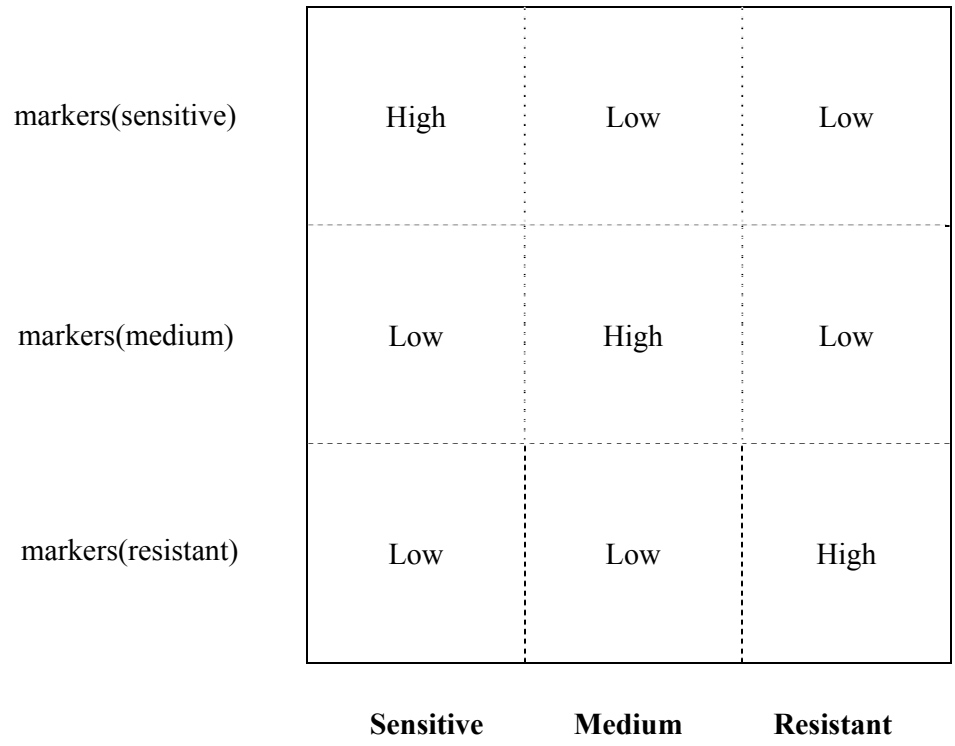



**Figure 2-5:** The expression pattern of type-II chemo-response (over expression) markers for three classes: sensitive, medium and resistant samples

Our next task is to build up a mathematical algorithm and implement it to extract both type-I and type-II signature markers for "N class" samples. The

expression of the type-I markers are significantly correlated, either positively or negatively, with the rank of response class(from 1 to N). Hence, we can use the correlation analysis method to develop type-I markers. Pearson product-moment correlation coefficient (denoted by *r* here) is a measure of the correlation or linear dependence between two variables (Eq 2-6). Another statistic metric to measure correlation is Spearman's rank correlation coefficient (denoted by ρ here), which is a non-parametric measure of statistical dependence between two variables. It estimates the monotonic trends of two variables. The Spearman correlation coefficient is defined as the Pearson moment correlation coefficient between the ranked variables. We employ the correlation coefficient method to measure the correlation between the expression of each marker and the rank of response class. In the meanwhile, the fold change method is used as the preliminary criteria to ensure that the expression of the signature marker between two adjacent class samples, such as class j samples and class j+1 samples, is significantly different. After calculating the correlation coefficient, we performed the statistical test to select the markers which show significant correlation with the rank of response class. Thereafter, p-values reported from the statistical test and the markers with the corresponding p-values are smaller than the p-value cut-off, such as 0.01 or 0.05, are selected as the significant correlated markers. Table 2-2 lists the implementation of R scripts to calculate the correlation coefficient using Pearson moment correlation coefficient and Spearman rank correlation coefficient. Table 2-3 lists the implementation of R scripts to identify Type-I markers which are significantly correlated with the rank of response class for

three classes. As we extend the number of fuzzy clusters of the chemo-response data from three classes to "N(N>3)" classes, the R code should be easily modified to generate the significantly correlated markers. On the other side, when we degenerate the number of the classes from three classes to two classes, the significantly correlated markers should give similar results as other procedures, like t-test, SAM and Limma, with same FC threshold and p-value cutoff.

$$r(E,Y)=\frac{1}{n-1}\sum_{j}^{n}\left(\frac{E_j-\overline{E}}{s_E}\right)\left(\frac{Y_j-\overline{Y}}{s_Y}\right) \qquad \text{Eq 2-6}$$

E: expression of gene i in n samples; Y: the rank of response class from fuzzy classification of chemo-response data; $\overline{E}$ and $\overline{Y}$, $s_E$ and $s_Y$ are the mean and standard deviation of E and Y respectively.

```
### Estimate the correlation coefficient between the expression of the potential marker and the
rank of class from fuzzy classification of chemo-response data
Correlation.ExprArray.ResponseClass = function(ExprArray,ResponseClass,Method)
{
    CorrelationCoefficient = vector('numeric',length(nrow(ExprArray)))
    PValue.test            = vector('numeric',length(nrow(ExprArray)))
    for (i in 1:nrow(ExprArray))
    {
        if (Method=="pearson" )
        {
            CorrelationCoefficient.ithMarker =
            cor.test(ExprArray[i,],as.numeric(ResponseClass),method="pearson")
        }
        if (Method=="spearman")
        {
             CorrelationCoefficient.ithMarker =
             cor.test(ExprArray[i,],as.numeric(ResponseClass),method="spearman")
        }
        CorrelationCoefficient[i] = CorrelationCoefficient.ithMarker$estimate
        PValue.test[i]            = CorrelationCoefficient.ithMarker$p.value
    }
    return(list(CorrelationCoefficient,PValue.test))
}
```

**Table 2-2**: The implementation of R scripts to calculate Pearson moment correlation coefficient and Spearman rank correlation coefficient

```
Class   = Response.FuzzyClassification
Class[which(Response.FuzzyClassification =="Sensitive")] = 2
Class[which(Response.FuzzyClassification =="Medium")]   = 1
Class[which(Response.FuzzyClassification =="Resistant")] = 0
Class = matrix(as.numeric(Class),nrow=1)
rownames(Class) = "Classes"
colnames(Class) = toupper(SamplesNames)
### Filtering with Fold Change metric
Min.LogFC   = FC.Threshold
Q.Expr   = vector(length=nrow(ExprArray.PreProcessed))
u2 = apply(ExprArray.PreProcessed[,which(Class==2)],1,mean)
u1 = apply(ExprArray.PreProcessed[,which(Class==1)],1,mean)
u0 = apply(ExprArray.PreProcessed[,which(Class==0)],1,mean)
fc2_1 = u2-u1
fc1_0 = u1-u0
ExprArray = ExprArray.PreProcessed [which((abs(fc2_1) >= Min.LogFC)&(abs(fc1_0) >=
Min.LogFC)),]
### Extracting the significantly correlated genes
PValue.Cutoff = 0.05
Correlation = Correlation.ExprArray.ResponseClass(ExprArray,t(Class),"pearson")
SigGene.ExprArray = ExprArray[Cor[[2]] < PValue.Cutoff,]
Sig.Correlation = matrix(Cor[[1]][Cor[[2]]< PValue.Cutoff],ncol=1)
colnames(Sig.Correlation) = "CorrelationCoefficient"
GXP.Signature = cbind(Sig.Correlation, SigGene.ExprArray)
GXP.Signature = GXP.Signature[order(Sig.Correlation,decreasing=TRUE),]
UpDown =
matrix(c(rep("Up",length(which(Sig.Correlation >0))),rep("Down",length(which(Sig.Correlation
<0)))),ncol=1)
colnames(UpDown) = "StatusOfExpression"
GXP.Signature = cbind(UpDown,GXP.Signature)
```

**Table 2-3**: The implementation of R scripts to identify Type-I markers which are significantly correlated with the rank of class from fuzzy classification of the chemo-response data (three fuzzy classes in this instance)

The type-II markers are basically the combination of N sets of significantly differentiated markers. For three classes, the type-II signature markers are the combination of markers(sensitive), markers(medium) and markers(resistant). Table 2-4 lists the R scripts to develop markers(sensitive) that are only significantly

over-expressed in sensitive samples, but are significantly under-expressed in both medium and resistant samples. Similarly, we can modify the scripts to obtain another two subsets of signature markers: markers(medium) are only significantly over-expressed in medium samples, but significantly under-expressed in both sensitive and resistant samples; markers(resistant) are only significantly over-expressed in resistant samples but significantly under-expressed in both sensitive and medium samples.

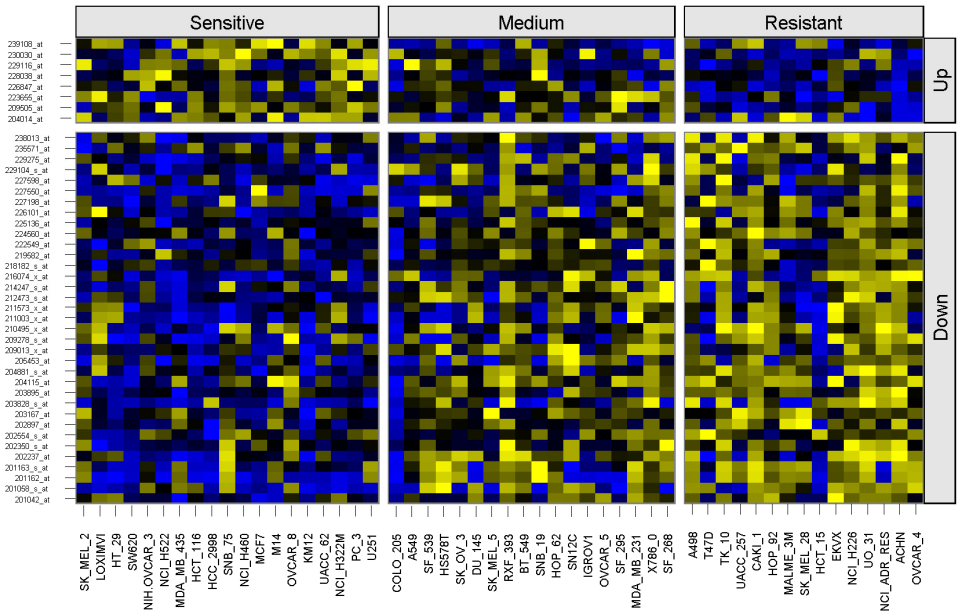
```
### Step a: Developing marker(sensitive)：markers are significantly over expressed in sensitive
sample but under expressed in both medium and resistant samples
Min.LogFC  = LogFC.Cutoff
FDR = FDR.Cutoff
Q.Expr = vector(length=nrow(ExprArray.PreProcessed))
Class.S.MR = Response.FuzzyClassification
Class.S.MR[which(Class=="Sensitive")] = 1
Class.S.MR[which((Class=="Medium")|(Class=="Resistant"))] = 0
Design = cbind(mean = 1, diff = as.numeric(Class.S.MR))
u2 = apply(ExprArray.PreProcessed[,which(Class.S.MR ==1)],1,mean)
u1 = apply(ExprArray.PreProcessed[,which(Class.S.MR ==0)],1,mean)
fc  = u2-u1
FC.Expr = ExprArray.PreProcessed[which(fc>= Min.LogFC ),]
fit1  = lmFit(FC.Expr, design)
fit2  = eBayes(fit1)
maxNum.probesets = min(nrow(FC.Expr),300)
topGenes  = topTable(fit2, coef = "diff", number=maxNum.probesets, adjust.method = "BH",
sort.by="logFC")
SigGenes.Sensitive = topGenes[which(topGenes$adj.P.Val<FDR),]
SigGenes.Expr = FC.Expr[SigGenes.Sensitive $ID,]
GeneType = matrix(rep("SenGenes",nrow(SigGenes.Expr)),ncol=1)
Feature.Expr = cbind(GeneType, SigGenes.Sensitive[,c("adj.P.Val")], SigGenes.Expr)
```

**Table 2-4:** The implementation of R scripts to identify markers(sensitive) of Type-II markers which are significantly over-expressed in sensitive samples but under-expressed in medium and resistant samples (three fuzzy classes in this instance)

Following the example of Paclitaxel, we have clustered 52 NCI60 solid tumor cell lines into three fuzzy classes. We then coupled the classified chemo-response data with microarray data (Affymetrix U133) to develop both type-I and type-II gene expression signatures. With the cutoff of LogFC=0.6 and the threshold of p-value=0.05, the developed type-I gene signature (Figure 2-6) includes 43 Affymetrix probesets (8 up, 35 under expressed in sensitive cell lines). This developed type-I gene signature show significantly correlated expression with the chemo-response of sensitive, medium and resistant samples. With the cutoff of LogFC=1 and the minimum of FDR=0.1, the type-II gene signature (Figure 2-7) includes 179 Affymetrix probesets (27 genes(sensitive) noted as “SenGenes”; 85 genes(medium) noted as “MedGenes”; 67 genes(resistant) noted as “ResGenes”), that clearly describes the over-expression pattern in the corresponding “class” samples.

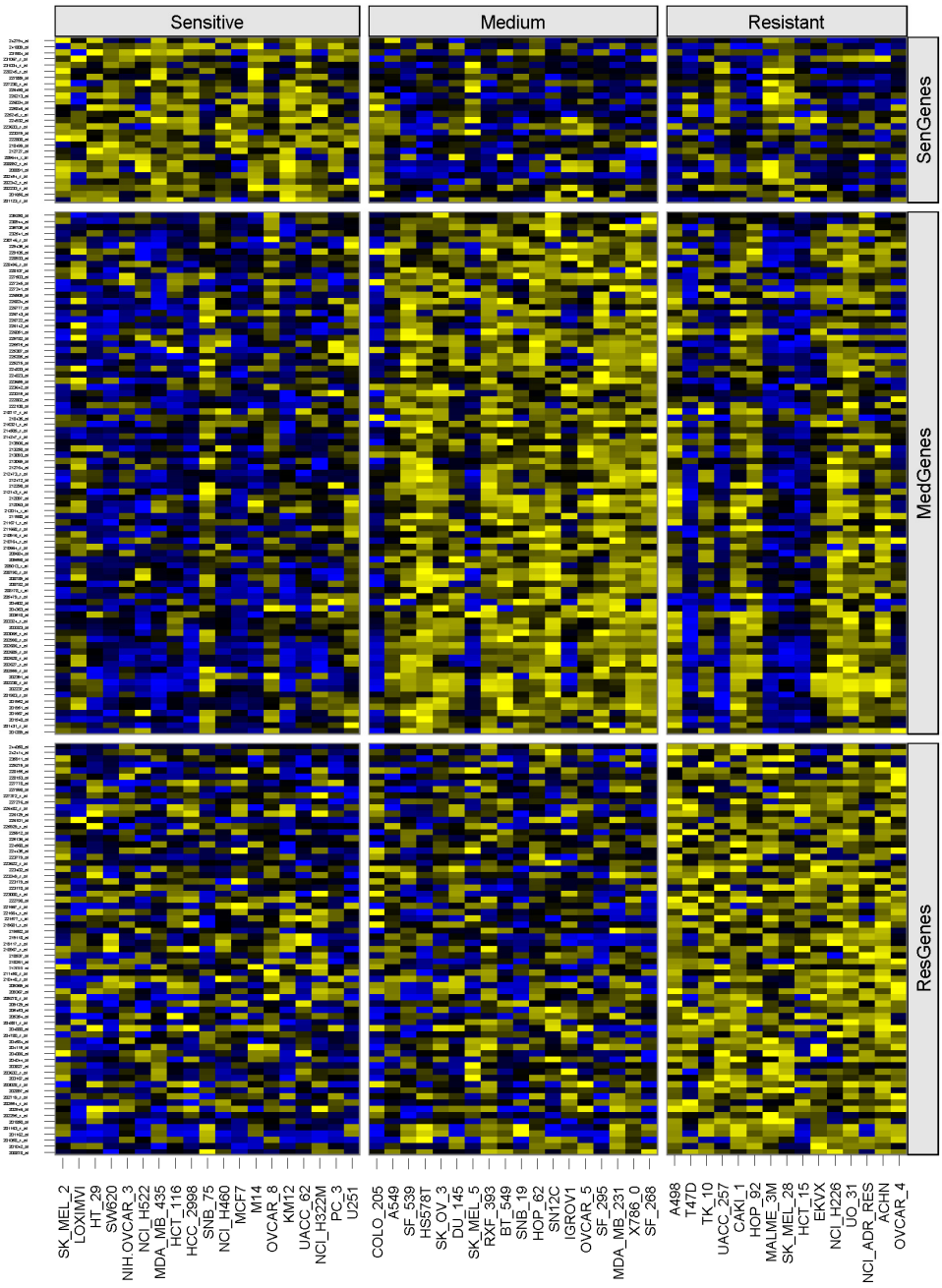
In these two developed signature marker sets, we found that 16 genes appear in both type I and type II marker sets and show significant over expression in Paclitaxel resistant cell lines: TGM2, MYL9, IGFBP7, GSTM3, SIRPA, TIMP2, GNG11, UGCG, HOXB2, TFPI2, OGFRL1, TIMP2, PLEKHA2, PRKCE, C7orf29, IGFN1. The basic hypothesis we may form is these 16 probable genes maybe related with mechanism of resistance of Paclitaxel. We then performed the gene ontology analysis using Ingenuity Pathway Analysis (IPA) tool. As shown in Figure 2-8, the analysis has indicated that the mechanism of Paclitaxel resistance may be affected by altered drug metabolism and cellular movement. Although, researchers have presented the drug metabolism could inactive cytotoxic anticancer agents, like

Paclitaxel [117], these 16 genes may have more comprehensive roles in cell proliferation. In Figure 2-9, the enriched molecular network also shows that the protein signaling pathway is seemingly active as well.




yellow: high value; blue: low value

**Figure 2-6:** Gene expression pattern of Paclitaxel(NCI60) type-I chemo-response signature for three classes: sensitive, medium and resistant samples




yellow: high value; blue: low value

**Figure 2-7:** Gene expression pattern of Paclitaxel(NCI60) type-II chemo-response signature for three classes: sensitive, medium and resistant samples

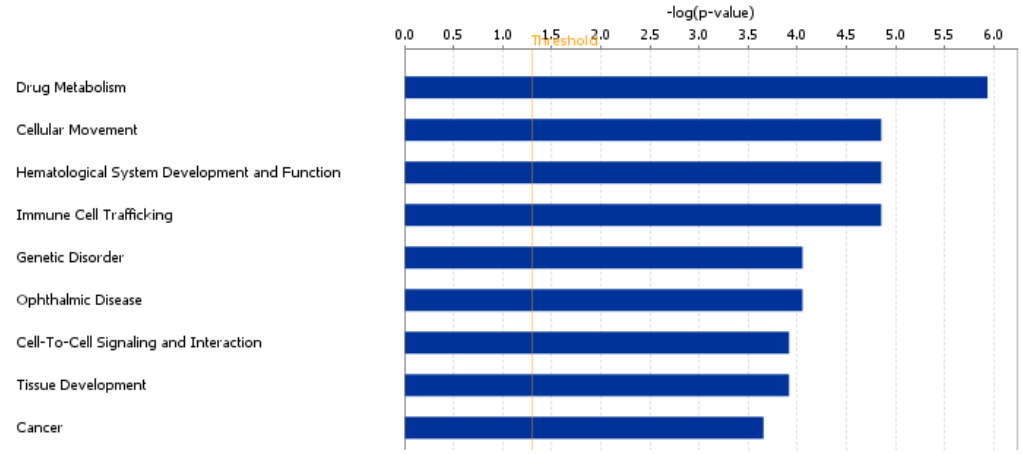



**Figure 2-8:** Gene ontology analysis of 16 Paclitaxel signature genes which are over expressed in Paclitaxel resistant cell lines (IPA content version: 11631407)

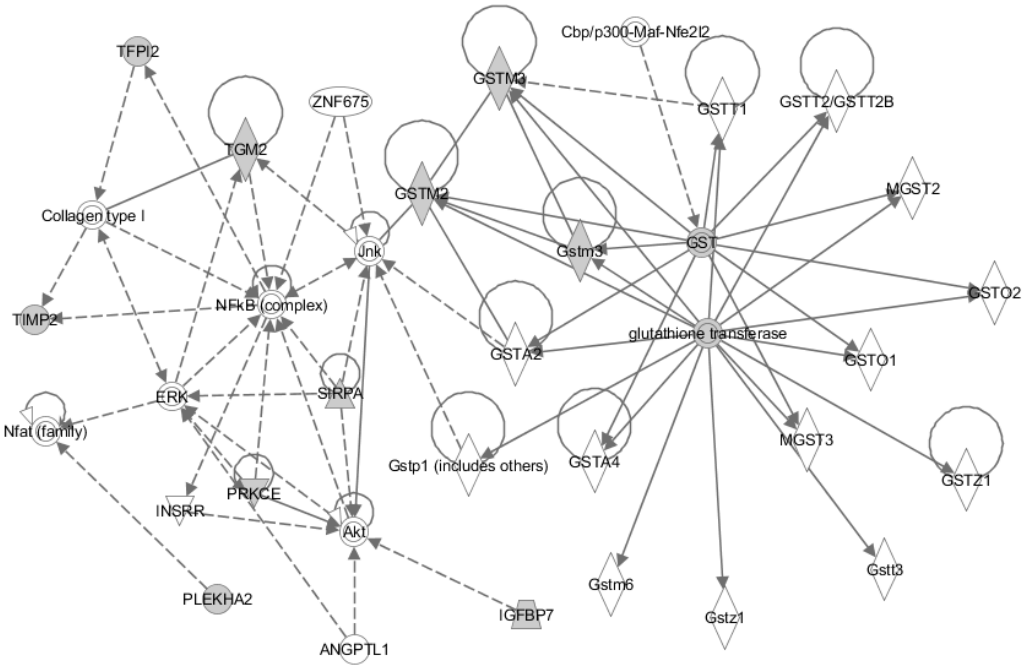



**Figure 2-9:** Core network analysis of 16 Paclitaxel signature genes which are over expressed in Paclitaxel resistant cell lines (IPA content version: 11631407)

## 2.4 Test the randomness and uniqueness of the developed signature markers

Before we apply the developed signature markers to predict the chemo-response of the anticancer agent(s), the critical questions we need to answer are: "Is the signature randomly present in the test data?" and "How is the uniqueness of the signature in the test data"? It is hard to answer these questions if we analyze the complex biology of the whole signature gene set. Instead, we could borrow a statistical metric of the correlation between two matrices known as "Mantel statistics" to evaluate the uniqueness of the developed signature markers in the test data.

The Mantel test is commonly used in ecology estimate the distance between objects such as species of organisms. In statistics, the Mantel test[118] allows linear or monotonic comparisons between the elements of two distance matrices. The null hypothesis of Mantel test is there is no relation between the two square matrices X and Y. The values within each matrix ($X_{ij}$ or $Y_{ij}$) represent a relationship between points i and j. The relationship represented by a matrix could be a geographical distance, a data distance or any other conceivable data. The basic statistics of Mantel test is using Mantel Z metric, which is simply the sum of the products of the corresponding elements of the matrices:

$$Z = \sum_{ij} X_{ij} Y_{ij} \qquad \text{Eq 2-7}$$

Where $\sum$ is the double sum over all i and all j where $i \neq j$. Because Z can take any value depending on the exact nature of X and Y, the Mantel Z metric is usually

normalized to -1 to 1, as the correlation coefficient between the pair-wise elements of X and Y. Therefore, the Mantel metric may be interpreted as a parameter which is similar to correlation coefficient. The original Mantel test[118] gives unbiased test statistics that could reveal whether two matrices are significantly similar or not by randomly permuting the rows and columns of one of the matrices many times over. Since we have only one developed signature, the global measure of similarity of the expression signature and only a set of random genes will not be enough to indicate the uniqueness of the gene expression signature in the test dataset. Hence, we randomly chose many random expression signatures to evaluate the distribution of the Mantel metric Z estimated as the similarities between the actual expression signature and random expression signatures (noted as $Z_{AR}$). Similarly, we can also create a baseline distribution of the Mantel metric Z between one random expression signature and another random expression signature (noted as $Z_{RR}$). Thus the null hypothesis of our test is:

| | |
|---|---|
| H0: $Z_{AR} = Z_{RR}$ | Eq 2-8 |

Here, we rephrase both $Z_{AR}$ and $Z_{RR}$:

$Z_{AR}$: the similarity metric between the actual expression markers set and random expression markers set

$Z_{RR}$: the similarity metric between the random expression markers set and another random expression marker set

After sampling large number of random signatures, with assuming both $Z_{AR}$ and $Z_{RR}$ follow Gaussian distribution, we test the significance of the hypothesis H0 using simple t-test statistics.

Table 2-3 lists the implementation of R scripts to evaluate the uniqueness of the developed signature makers in the test data. The Mantel Z statistics is calculated using "mantel.randtest" function in R package "ade4"(version 1.4-17). Figure 2-8 depicts the uniqueness test results of Hoeflich Mek[119] gene expression signature in Bittner breast cancer data. The Mek gene expression signature are composed of genes significantly differentially expressed between transfected HRas or Mek MCF-10A cell lines and MCF-10A cell lines[119]. The distribution of the Mantel similarity between the actual Hoeflich Ras/Mek pathway signature and the sampled random gene signatures is significantly lower than the baseline distribution which is defined as the Mantel statistics metric among random gene signatures. The unique presence of Ras/Mek pathway signature in breast cancer tumor samples is consistent with the existing preclinical and clinical studies[119-121]. However, the results from uniqueness test of Hoeflich Ras/Mek pathway signature in NCI60 suggest that the Hoeflich Ras/Mek pathway signature behaves randomly in NCI60 cell lines (Figure 2-9). This may due to the observation that basal-like breast cancer tumors cells are more dependent on Ras-Mek pathway activity compared with other cancer tumors. This is why the Hoeflich Ras/Mek pathway signature is predominantly present in breast cancer tumors[119]. Further studies have shown that the Mek inhibition is determined by elevated Ras signal and the feedback signaling of MAPK/ERK Kinase

(MEK)-Phosphoinositide 3-Kinase[122]. The predicted results of Hoeflich Ras/Mek gene signature in both datasets have shown that the pattern of the signature in Bittner Breast cancer data is clear: about 50% of the breast cancer samples have shown active Ras-Mek pathway when compared with the rest of the samples. However, the pattern of the signature in NCI60 cell lines is relatively very weak as only about 10-20% of cell lines show that the Ras-Mek pathway is active.

```
#################################################################
###
### test the uniqueness of gene signature in test data
obs_vector_Actual_Rnd = vector("numeric",length=0)
obs_vector_Rnd_Rnd     = vector("numeric",length=0)
# The actual signature markers in test data
exprData.bioMarker.Actual = exprData.bioMarker
# the random signature markers in test data
exprData.bioMarker.Rnd   =
matrix(0,nrow=dim(exprData.bioMarker)[1],ncol=dim(exprData.bioMarker)[2])
exprData.bioMarker.Rnd1 =
matrix(0,nrow=dim(exprData.bioMarker)[1],ncol=dim(exprData.bioMarker)[2])
for (iter in 1:n.permutation)
 {
     exprData.bioMarker.Rnd   =
dataset.Matrix[sample(dim(dataset.Matrix)[1])[1:dim(exprData.bioMarker)[1]],]
     exprData.bioMarker.Rnd1 =
dataset.Matrix[sample(dim(dataset.Matrix)[1])[1:dim(exprData.bioMarker)[1]],]
     # set nrepet to 10 as we are not interested in the emperical p-value of the test
     mantel.bioMarker.Act = mantel.randtest(dist(t(exprData.bioMarker.Actual),method=Metric),
                          dist(t(exprData.bioMarker.Rnd),method=Metric),nrepet=10)
     mantel.bioMarker.Rnd = mantel.randtest(dist(t(exprData.bioMarker.Rnd1),method=Metric),
                          dist(t(exprData.bioMarker.Rnd),method=Metric),nrepet=10)
     # Aggregate the Mantel metric from a random selected genes/probesets compared to
     # another randomly selected genes/probesets
     obs_vector_Rnd_Rnd     =   c(obs_vector_Rnd_Rnd, as.numeric(mantel.bioMarker.Rnd$obs))
```

**Table 2-5**: The implementation of R scripts to test the uniqueness of the signature markers in the test data using Mantel statistics metric

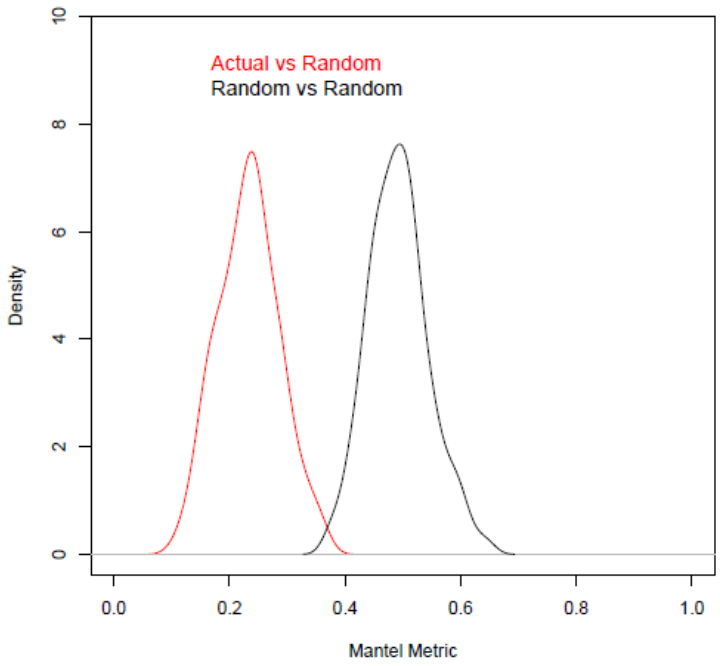



**Figure 2-10:** Hoeflich Ras/Mek pathway gene expression signature is uniquely present in Bittner breast cancer tumor samples (p.value<0.001)

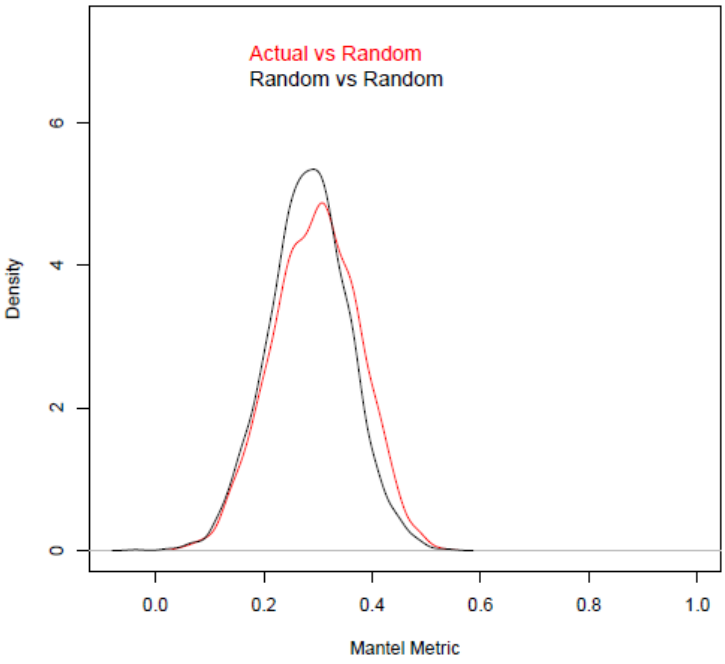



**Figure 2-11:** Hoeflich Ras/Mek pathway gene expression signature is randomly present in NCI60 cell lines data (p.value=0.953)

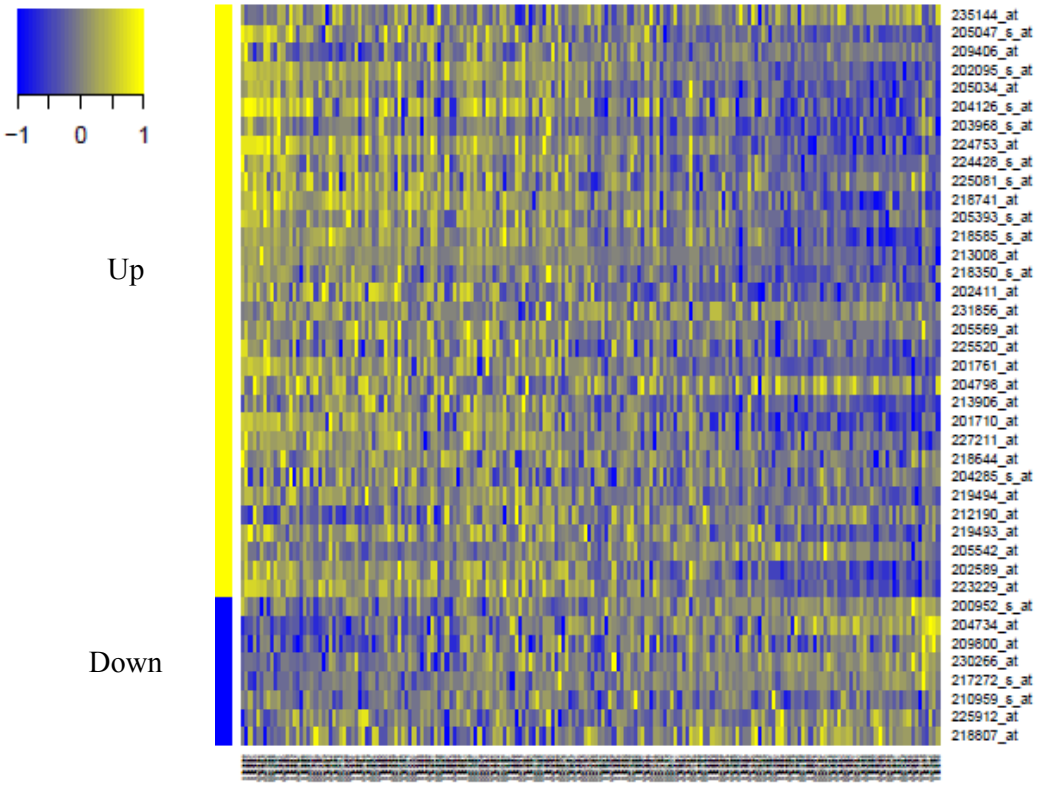



**Figure 2-12:** Hoeflich Ras/Mek pathway gene expression signature in Bittner Breast cancer datasets; samples are sorted by the predicted probability of activity using Bayesian Metagene projection methods[79].

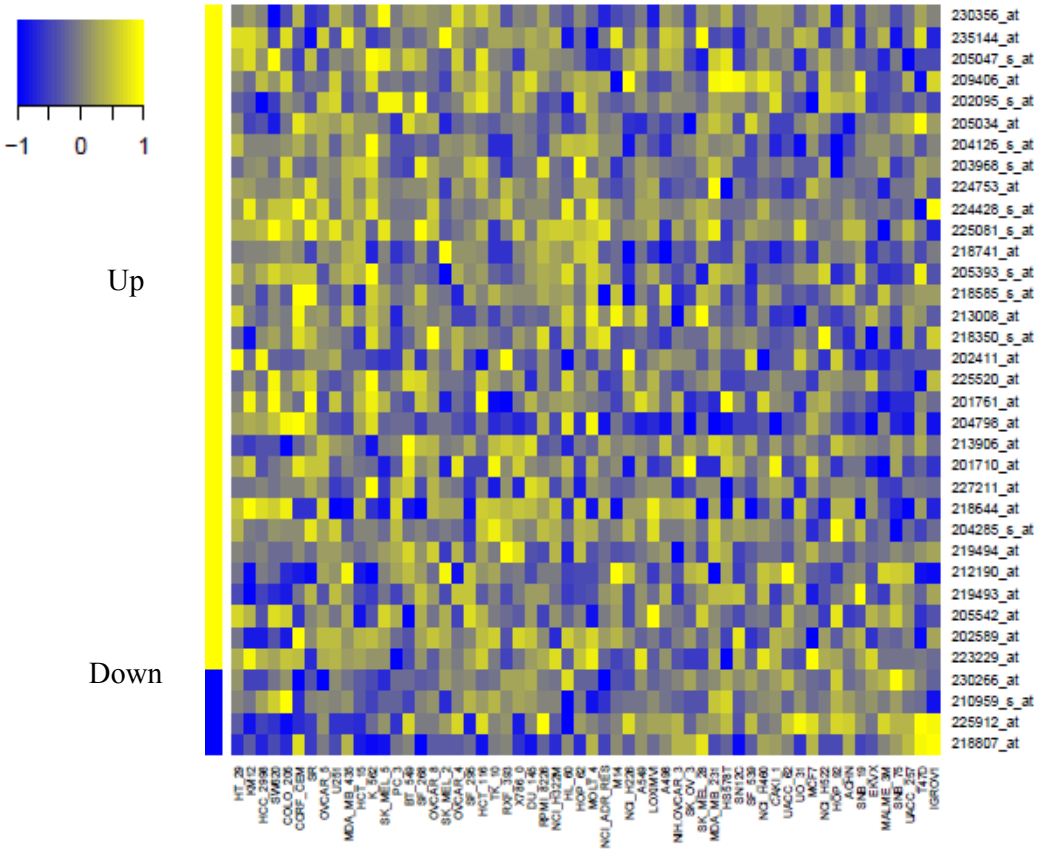



**Figure 2-13:** Hoeflich Ras/Mek pathway gene expression signature in NCI60 cell line datasets; samples are sorted by the predicted probability of activity using Bayesian Metagene projection methods[79].

## 2.5 Signature detection method

After we assessed the developed signature markers that are uniquely present in the test data, the following step is to discover the expression pattern of the signature markers in the test data. The problem with these types of analyses is that the number of signature markers usually range between tens to hundreds. This, however, does not allow us to study all the markers individually in order to detect each gene's expression pattern specifically. Bioinformatics algorithms that can manipulate the overall signature markers are needed. Generally, there are two types of methods: supervised

and unsupervised. By employing the unsupervised method, the patterns of the expression signature in the test data are inherently determined without a priori information. Typical algorithms include relevance networks[123, 124], hierarchical dendrograms and self organizing maps[125, 126], and all these algorithms require a metric, such as Euclidean distance, correlation coefficient or mutual information between the markers. The end result of unsupervised analysis of the expression signature in the test data is the agglomerative clusters information of samples and signature markers. Another type of the analysis is supervised learning. By using the context of signature development in the training data, parameters in the models, such as support vector machine (SVM)[127, 128], neural network[129] and decision trees[130], can be determined and then used to predict samples in the test data.

Although the characterization of the signature expression in the test data is interesting, it is more important to address the issue on how much is the biological pattern(s) of the expression signature in the training data remained in the test data. Therefore, a statistical method is necessary to quantitatively measure the extent of how the pattern remained in the test data. The metagene methodologies have proven to be capable in capturing the phenotypic patterns of the expression data[131-134]. The concept of metagene projection was originally presented by West et al.[79], and it has demonstrated to be useful in predicting the expression pattern that remained in cross platforms[135]. West et al. also proposed a statistical method which integrated standard binary regression models, singular value decomposition(SVD) and metagene projection to measure the extent of the expression pattern of training data remained in

the test data. Here for simplicity, we name it as “Bayesian-SVD” method. The developed signature data is constructed with samples in columns and markers in rows. Principal components of the training data are used to compute the metagene and metasample values, and then form metagene signature. The expression signature in the test data is then projected into this metagene signature. The projected metagene signature retains the most information of the signature expression pattern in the test data. West et al.[79] used Bayesian binary probit regression model to generate the probability of the remaining signature pattern for two classes problem. We name the predicted probability of the chemo-response signature as the probability of sensitivity (POS) and the predicted probability of pathway or oncogene activity signature as the probability of activity (POA). The model was fitted to the metagene signature and relative probability of sensitivity (POS) or probability of activity (POA) is estimated to the projected metagene expression signature using the Bayesian binary probit regression parameters fitted from the metagene signature. When studying chemo-response predictor, the classes are defined as ‘0’ for resistant and ‘1’ for sensitive for training. Low POS scores would be suggestive of a sample being resistant and vice versa. When studying pathway or oncogene predictors, the classes are defined as ‘0’ for hypoactive and ‘1’ for hyperactive. The critical issue of Bayesian-SVD method is we could only use the major singular value instead of all non-zero singular value in the application of microarray data analysis due to the large scale of microarray data. The Bayesian-SVD method may lose some information in

the training data. Therefore, we introduce another metagene projection method as follows and which is named as Bayesian-NMF method.

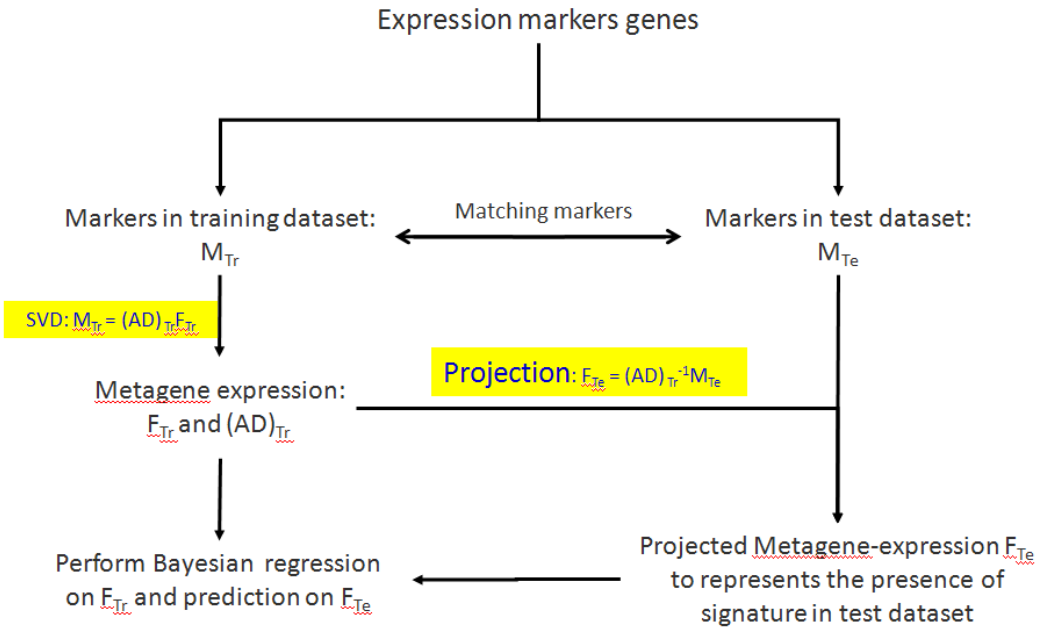



**Figure 2-14:** The schematic flow chart of detecting the expression signature in the test data using Bayesian-SVD metagene projection method. $M_{Tr}$: the marker genes in the original training dataset, $M_{Te}$: the marker genes in the test dataset, $F_{Tr}$: metagene expression, $F_{Te}$: projected metagene expression.

In this dissertation, we present a novel method to detect the expression signatures in the test datasets by assigning the predicted probability to each sample. The novel computational method is based on Non-Negative Matrix Factorization (NMF). NMF is basically a type of an Independent Components Analysis (ICA) variant with the restriction(s) to positive values[136]. The feature of non-negativity of the decomposed matrix elements facilitates the NMF method such that it can be widely used in genomic data analysis[137-140]. In contrast to Principal Component Analysis (PCA) or Singular Value Decomposition (SVD) and other ICA-based methods, the negative metagene expression is inherently close to the interpretation of

gene or protein expression. The NMF method permits the decomposed matrix to be sparse and it also permits localized features[136, 139], which means that the NMF may capture the critical expression information of the genes, such as oncogene genes or tumor suppressor genes in cancer research. This feature suggests the NMF method is superior to the PCA and other ICA based approaches. The NMF algorithm usually gives local optimizations, and this seems more suitable for the biological systems which essentially have multiple stable points after perturbations. This framework combines the Bayesian regression method, the NMF method and the metagene projection method.

Consider a gene expression data set consisting of the expression levels of N features (genes or probesets) in S samples, which represents different types of experiments or experiment time points. For gene expression signature, the number N features is typically in the hundreds, and the number S experiments is typically less than 100. We refer the gene expression signature datasets by an expression matrix X of size N by S, whose rows contain the expression levels of the N features in the S samples. The Non-negative Matrix Factorization decomposes the matrix X into the multiplication of two small sized matrices, and each of them with non-negative elements. The NMF method is the approximation of the positively linear combinations of metagene expressions, which is defined as "factor matrix". Formally, the standard NMF is described as follows:

| | |
|---|---|
| $NMF: X \approx W \cdot H$ | Eq 2-8 |

X corresponds to the gene expression signature with size N by S, W is the factoring matrix with non-negative entities, it has size N by k, with each of the k columns defining a metagene or factor; entry $W_{ij}$ is the coefficient of gene i in metagene j. The loading of matrix H has size k by S, with each of the M columns representing the metagene expression pattern of the corresponding sample; entry $H_{ij}$ represents the expression level of metagene i in sample j.

The NMF decomposition is usually done by an iterative updating method. The multiplication updating rule was proposed by Lee and Seung (1999).

$$H_{q,j} = H_{q,j} \frac{\left(W^T X\right)_{q,j}}{\left(W^T WH\right)_{q,j}}, \quad 1 \le q \le k, 1 \le j \le S$$
$$W_{i,q} = W_{i,q} \frac{\left(XH^T\right)_{i,q}}{\left(WHH^T\right)_{i,q}}, \quad 1 \le i \le N, 1 \le q \le k$$

Eq 2-9

To enhance the sparse of the metagene expressions, Pascual-Montano et al. (2006) proposed a non-smooth NMF method by adding sparseness constraints to the Lee and Seung's NMF procedure. Mathematically, Pascual-Montano et al. defined a "smoothing" matrix S, which is a positive symmetric matrix with the size q by q.

$$S = (1-\theta)I + \frac{\theta}{q} V_1 V_1^T$$

Eq 2-10

Where I is the identity matrix, $V_1$ is a vector of ones and the parameter $\theta$ represent the sparseness, which satisfies $0 \le \theta \le 1$. Therefore, the Non-smooth Nonnegative Matrix Factorization model was defined as:

$$nsNMF: X \approx W \cdot S \cdot H$$

Eq 2-11

The nsNMF method can be quickly implemented by following Lee and Seung's updating rule. Correspondingly, in the update equation for H, substitute W with WS and in the update equation for W, substitute H with SH. Table 2-4 lists the detailed implementation of NMF algorithm with Lee and Seung's multiplicative updating rule.

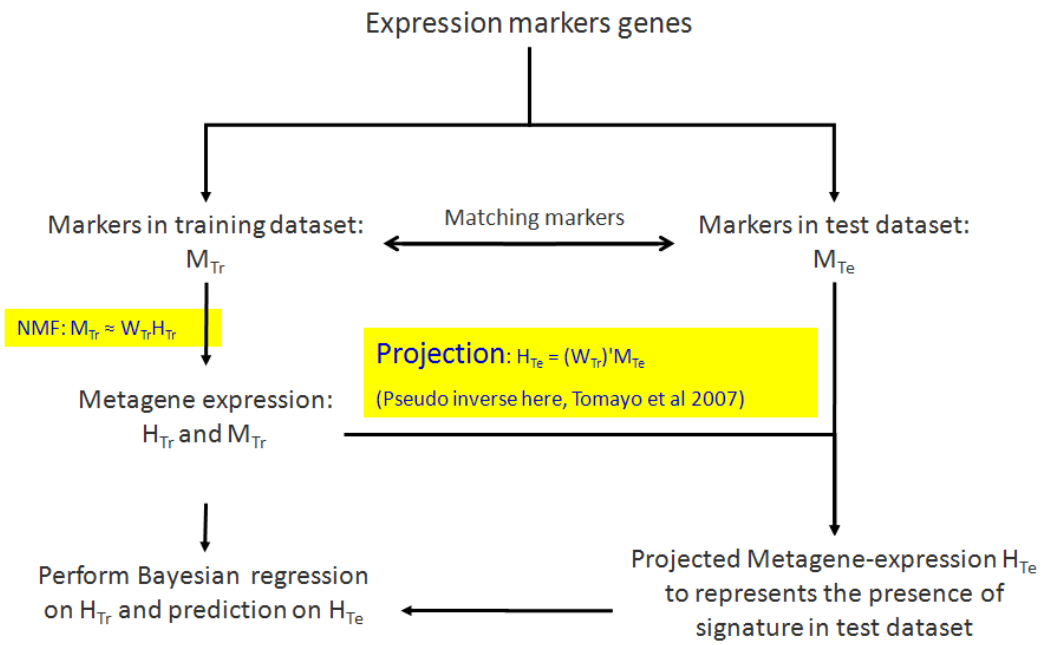


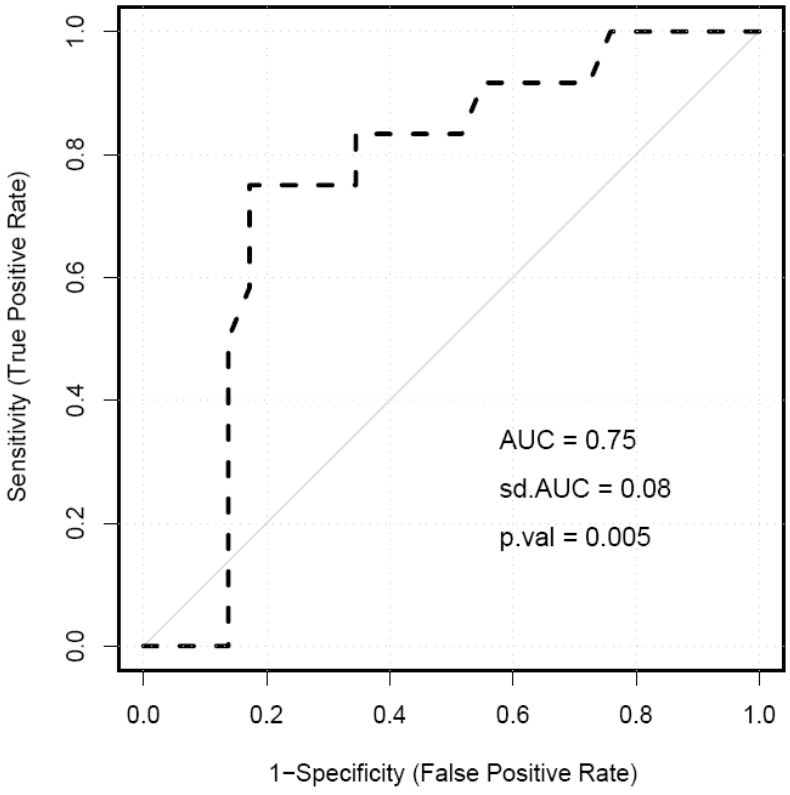

**Figure 2-15:** The schematic flow chart of detecting the expression signature in the test data using Bayesian-NMF metagene projection method. $M_{Tr}$: the marker genes in the original training dataset, $M_{Te}$: the marker genes in the test dataset, $H_{Tr}$: metagene expression, $H_{Te}$: projected metagene expression.

Figure 2-15 depicts the flow chart of using Bayesian-NMF metagene projection method to detect the expression signature in the test data. Figure 2-16 and Figure 2-17 describe the detection results of Bild Ras[141] oncogene activity signature in Ding lung[142] primary tumor and Bhattercharjee Lung[143] primary tumor data respectively. Both shows AUC is significantly greater than 0.5 when comparing the predicted activity of Ras activity in K-Ras mutants and wild type samples, which indicate that the presented Bayesian-NMF metagene projection method is able to detect the expression signatures in an unbiased way. In order to compare the performance of the Bayesian –NMF method, Figure 2-18 depicts the detection results of Bild Ras[141] oncogene activity signature in Bhattercharjee Lungprimary tumor data using Bayesian-SVD method. The Bayesian-NMF method shows very similar detection results with Bayesian-SVD method in terms of AUC metric. Although the example shown here reflects only two class problems, the dimension of the NMF decomposed matrix can be three or more, based on the

response class information and the optimal solution from NMF algorithm. Therefore, this method can be applied to multiple class problems as well.

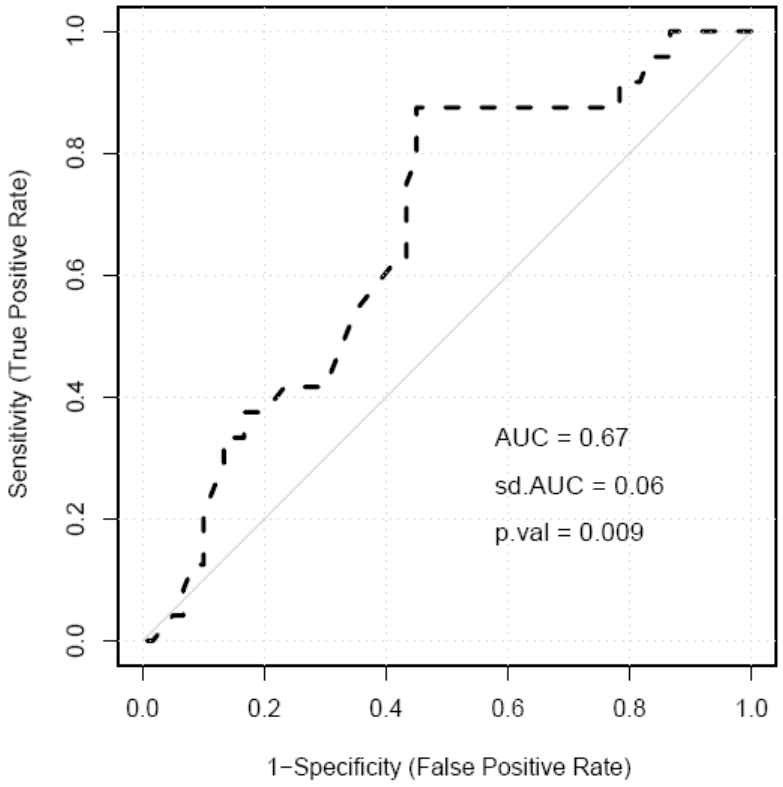
```
### NMF algorithm: Lee and Seung's updating rule
NMF = function(M,r)
{    # M is the original matrix for factorization;   r is the lower dimension and r <
min(nrow(M),ncol(M))
    connect    = min(50,nrow(M))   # converge condition 1
    tol.converge = 1.0e-6              # converge condition 2
    tol = 1.0e-8                       # small positive number
    # Initialize factorization matrices W,H
    n.row = nrow(M);   n.col = ncol(M);   r = min(r,n.row,n.col)
    mean.M =sqrt(mean(M));    sd.M = mean(sd(M))
    nmf.M = list()
    W0 = matrix(runif(n.row*r),n.row,r);       H0 = matrix(runif(r*n.col),r,n.col)
    Flag = 1;   Iter = 1;   Obj = NULL
    while(flag==1)
    {     H0[H0<0] = tol;   W0[W0<0] = tol
        # Compute new W and H (Lee and Seung)
        W1 =W0*(M%*%t(H0)/(W0%*%H0%*%t(H0)+tol))
        H1 = H0*(t(W1)%*%M/(t(W1)%*%W1%*%H0+tol))
        # Renormalize so rows of H have constant energy
        norms = sqrt(apply(H1^2,1,sum))
        H1    =   H1/matrix(rep(norms,n.col),ncol=n.col);    W1   =
W1*t(matrix(rep(norms,n.row),nrow=r))
        order.W1 = order(W1,decreasing=T);         order.H1 = order(H1,decreasing=T)
        order.W0 = order(W0,decreasing=T);         order.H0 = order(H0,decreasing=T)
        sum.obj0 = sum(abs(M-W0%*%H0));          sum.obj1 = sum(abs(M-W1%*%H1))
        if
(((sum(abs(order.W1[1:connect]-order.W0[1:connect])))==0)&((sum(abs(order.H1-order.H0)))=
=0) &(abs(sum.obj0-sum.obj1)<1.0e-3))
        {     abs.H.dif    = abs(H1-H0)
            order.H.dif = order(abs.H.dif,decreasing=T)
            if (((abs.H.dif[order.H.dif[1]]/H1[order.H.dif[1]])<tol.converge)&
              ((abs.H.dif[order.H.dif[2]]/H1[order.H.dif[2]])<tol.converge)&
              ((abs.H.dif[order.H.dif[3]]/H1[order.H.dif[3]])<tol.converge))
              {     flag = 0;          print('NMF is converged!')      }       }
        W0   = W1;           H0 =   H1
        Iter   =   Iter+1
        Obj   = c(Obj,sum.obj1)    }
        SSE = sum((M-W1%*%H1)^2)
        nmf.M$W = W1;      nmf.M$H   = H1
        nmf.M$SSE = Obj;   nmf.M$Iteration = Iter
        return(nmf.M)    }
```

**Table 2-6**: R scripts of the implementation of NMF algorithm (Lee and Seung's updating rule)



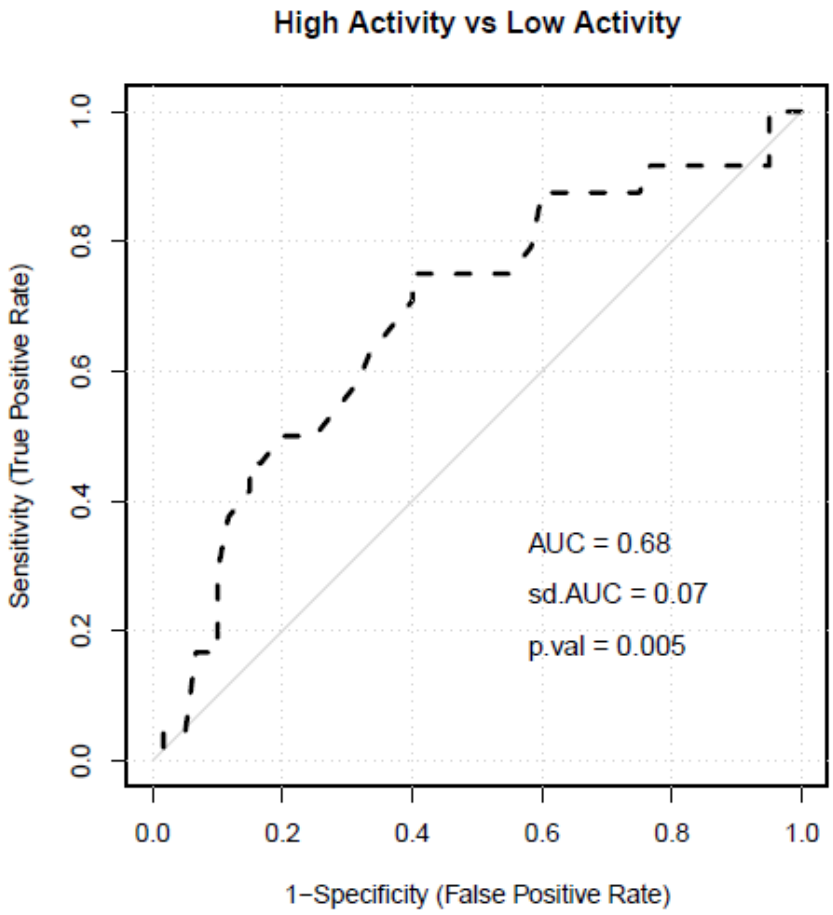

sd.AUC: Standard deviation of AUC estimate, p.val: Probability (null hypothesis: the expression signature is non-informative)

**Figure 2-16:** Receiver Operating Characteristic (ROC) curve of detecting Ras pathway expression signature in Ding Lung[142] primary tumor data using Bayesian-NMF method. In ROC plot, samples with K-Ras mutation and K-Ras wild type were compared, AUC=0.75±0.08.




sd.AUC: Standard deviation of AUC estimate, p.val: Probability (null hypothesis: the expression signature is non-informative)

**Figure 2-17:** Receiver Operating Characteristic (ROC) curve of detecting Ras pathway expression signature in Bhattercharjee Lung[143] primary tumor data using Bayesian-NMF method. In ROC plot, samples with K-Ras mutation and K-Ras wild type were compared, AUC=0.67±0.06.




sd.AUC: Standard deviation of AUC estimate, p.val: Probability (null hypothesis: the expression signature is non-informative)

**Figure 2-18:** Receiver Operating Characteristic (ROC) curve of detecting Ras pathway expression signature in Bhattercharjee Lung[143] primary tumor data using Bayesian-SVD method. In ROC plot, samples with K-Ras mutation and K-Ras wild type were compared, AUC=0.68±0.07.

## 2.5 Summary

We have reviewed a couple of methods to develop expression signature markers for two class problems. By suggesting a concept of fuzzy classification, it is very necessary to develop expression signature markers for three and more classes. Hence, two types of expression signature markers are proposed, and the corresponding development methods are also presented. The expression signature may be randomly and no uniquely present in the test data; therefore, we suggested a

computational method by borrowing a statistic metric of measuring a similarity of two matrices to evaluate the uniqueness of the signature in the test data. Lastly, in order to detect the signature in a more robust way, we proposed a novel framework that we call the Bayesian-NMF metagene projection method. The overall methodology of expression signature development, uniqueness test, and detection is systematic and robust.

# Chapter 3 Identifying minimal marker sets for clinical translation

## 3.1 Introduction

In chapter 2, we proposed a systematic bioinformatics methodology to develop the principle expression chemo-response markers, by which the markers, based on say, gene expression, microRNA etc., are significantly differentiated between two class samples, or are correlated with multiple class samples that can be identified and utilized to predict agents' chemo-response in independent cancer data. These signature markers that correlate with a phenotype of interest have shown to play very important roles in cancer research and clinical prognosis[144-168]. However, it is still very challenging to move multiple-gene signature into clinical application, the efforts for assay development, optimization and further in-vitro, in-vivo, and clinical validations many take a lot of time and resources. Therefore, narrowing down the multiple gene/protein signature from tens to hundreds of markers to a minimal gene marker set (<10) or even a single gene marker is usually requested and welcomed by clinicians.

Since cancer is driven by the alterations of various cell signaling, cellular and physiological processes, it is essential to systematically understand the corresponding heterogeneous genomic biology of the expression markers, such as mutations, copy number variation or methylation, and regulations between transcripts or proteins and

miRNAs. One type of integrated analysis of genomic data is the meta-analysis, which is an analysis often applied to multiple similar datasets in clinical practice to classify tumor types and identify robust prognosis markers[169-172]. In order to narrow down the multiple-gene signature rigorously, it is essential to decode the over or under expression for each expression marker, such as transcripts or proteins, using the corresponding sequence genomics data, copy number variation, methylation data and miRNA data. Recently, there have been a number of integrated genomic analyses between gene expression profiling and DNA copy number variation, gene mutation, DNA methylations, microRNA expression and protein expression have been reported[173-181].

Beyond the various types of integrated analysis that have been applied in cancer data, a range of computational models have also been suggested. The most popular method is the correlation coefficients, which is applied to integrate DNA copy number and gene expression[182, 183], and methylation and gene expression[184] etc. The correlation based method captures simple pair wise relations of one type of genomic aberrations and its effect on the expression of the marker, but this may limit the understanding of the whole picture. Relating multiple genetic aberrations information with phenotypic expression helps to better comprehend the mechanism of cancer phenotypes. Therefore, a coherent model that combines different types of genetic aberration data is inevitably needed. The coherent model is expected to optimize the number of signature markers by identifying the most representative marker, enabling it to shift quickly into clinical applications.

## 3.2 Integrated genomic analysis using linear model

We propose a mathematical model to describe how genetic aberrations change the expression of the each single signature marker. Let $y_i$ the expression of the signature marker *i*, such as gene expression or protein expression, and its value is numeric (log2 transformed signal); $Mu_i$ is the mutation of marker gene i and its value is integer with -1 for frame shift deletion or homozygous deletion, 0 for wild type and 1 for point mutation or other types of activation mutation; $CNV_i$ is the DNA copy number of gene i and its value is numeric with or without normalized by the HapMap reference value; $Me_i$ is the methylation of gene (or promoter) i and its value is between -1 to 0. Recall the Eq 1-1 introduced in chapter 1, we obtain

$$y_i = f(CNV_i, Mu_i, Me_i) \quad \text{Eq 3-1}$$

Consider the combinational effects of copy number variation, mutation and methylation is linear, then

$$y_i = a_0 + a_1 \cdot Mu_i + a_2 \cdot CNV_i + a_3 \cdot Me_i, a_0, a_1, a_2, a_3 > 0 \quad \text{Eq 3-2}$$

By using linear programming techniques, we could identify the genes whose expressions are significantly driven by genomic aberration. For example, here we chose CDKN2A as the example to study this model. CDKN2A is one of the marker genes from our generated Paclitaxel chemo-sensitivity gene signature and it has known role as a biomarker for G2M anticancer agents[185]. Therefore, CDKN2A could be one of the most representative single gene maker to reflect the chemo-response of Paclitaxel in NCI60 data. Figure 3-1 describes CDKN2A gene

expression relates to the mutation and methylation. According to the plot, CDKN2A homozygous mutation samples show a very low expression of CDKN2A and wild type samples with low methylation show a very high expression. Figure 3-2 depicts the CDKN2A mRNA expression and the coded protein p16 expression; both of them are highly correlated except the wild type samples with high methylation, which tend to show significant low expression. Table 3-1 lists the detailed gene expression data of CDKN2A (Affymetrix U133(A&B) array, log2-transformed MAS5), and the genetic information which include mutation (COSMIC database v52), DNA copy number variation and CpG island methylation data (CellMiner v1.0). Table 3-2 lists the estimated parameters for CDKN2A gene fitted in model Eq 3-2. According to the statistical analysis, it is obvious that both CDKN2A mutation and CpG island methylation significantly affect the gene expression of CDKN2A gene.

| Cell Line | Classes | mRNA | Mut | CNV | Me.CpG | AA Change |
|---|---|---|---|---|---|---|
| SK_MEL_2 | Sensitive | 10.13 | 0 | -0.5 | 0.4 | WT |
| LOXIMVI | Sensitive | 8.17 | -1 | -1.03 | 0 | p.M1_ *157del |
| SW620 | Sensitive | 11.89 | 0 | -0.09 | 0.89 | WT |
| NIH.OVCAR_3 | Sensitive | 12.96 | 0 | 0.69 | 0 | WT |
| NCI_H522 | Sensitive | 12.78 | 0 | 0.24 | 0 | WT |
| MDA_MB_435 | Sensitive | 11.04 | 1 | 0.61 | 0 | p.? |
| HCT_116 | Sensitive | 10.54 | 1 | -0.13 | 0.57 | p.R24fsX20 |
| HCC_2998 | Sensitive | 12.04 | 0 | 0.25 | 0 | WT |
| SNB_75 | Sensitive | 11.74 | 0 | -0.49 | 0 | WT |
| NCI_H460 | Sensitive | 7.04 | -1 | -0.83 | 1 | p.? |
| MCF7 | Sensitive | 9.03 | -1 | -0.04 | 0 | p.M1_ *157del |
| M14 | Sensitive | 10.22 | 1 | 0.88 | 0 | p.? |
| OVCAR_8 | Sensitive | 11.88 | 0 | 0.62 | 0 | WT |
| KM12 | Sensitive | 12.62 | 0 | 0.58 | 0.84 | WT |

| | | | | | | |
|---|---|---|---|---|---|---|
| UACC_62 | Sensitive | 7.52 | -1 | -0.57 | 0 | p.M1_ *157del |
| NCI_H322M | Sensitive | 6.61 | 0 | -1.22 | 1 | WT |
| PC_3 | Sensitive | 11.84 | 0 | -0.39 | 0.89 | WT |
| U251 | Sensitive | 7.37 | -1 | -1.17 | 1 | p.M1_ *157del |
| COLO_205 | Medium | 10.56 | 0 | 0.12 | 0.96 | WT |
| A549 | Medium | 8.24 | -1 | -0.7 | 1 | p.M1_ *157del |
| SF_539 | Medium | 12.08 | 0 | 0.69 | 0 | WT |
| HS578T | Medium | 5.13 | -1 | -0.19 | 1 | p.M1_ *157del |
| SK_OV_3 | Medium | 8.34 | -1 | -1.64 | 0 | p.? |
| DU_145 | Medium | 13.41 | 1 | 0.06 | 0 | p.D84Y |
| SK_MEL_5 | Medium | 7.93 | -1 | -1.48 | 0 | p.M1_ *157del |
| RXF_393 | Medium | 7.17 | -1 | -1.65 | 0 | p.M1_ *157del |
| BT_549 | Medium | 13.58 | 0 | 0.53 | 0 | WT |
| SNB_19 | Medium | 7.26 | -1 | -0.32 | 1 | p.M1_ *157del |
| HOP_62 | Medium | 6.11 | -1 | -0.11 | 1 | p.M1_ *157del |
| SN12C | Medium | 12.93 | 0 | -0.2 | 0 | WT |
| IGROV1 | Medium | 9.27 | 0 | -0.32 | 0 | WT |
| OVCAR_5 | Medium | 7.09 | -1 | -1.45 | 0 | p.M1_ *157del |
| SF_295 | Medium | 6.88 | -1 | -0.44 | 1 | p.M1_ *157del |
| MDA_MB_231 | Medium | 5.98 | -1 | -0.31 | 1 | p.M1_ *157del |
| X786_0 | Medium | 5.34 | -1 | -0.98 | 1 | p.? |
| SF_268 | Medium | 4.56 | -1 | -1 | 1 | p.M1_ *157del |
| A498 | Resistant | 7.58 | -1 | -1.36 | 1 | p.M1_ *157del |
| T47D | Resistant | 10.25 | 0 | 0.14 | 0.9 | WT |
| TK_10 | Resistant | 12.26 | 0 | -0.02 | 0.81 | WT |
| UACC_257 | Resistant | 10.80 | 0 | -0.27 | 0 | WT |
| CAKI_1 | Resistant | 6.82 | -1 | -0.71 | 0 | p.M1_ *157del |
| HOP_92 | Resistant | 9.94 | -1 | -0.1 | 0 | p.M1_ *157del |
| MALME_3M | Resistant | 6.38 | -1 | -0.79 | 0 | p.M1_ *157del |
| SK_MEL_28 | Resistant | 10.97 | 0 | -0.64 | 0 | WT |
| HCT_15 | Resistant | 6.65 | 0 | 0.64 | 1 | WT |
| EKVX | Resistant | 10.81 | 0 | 0.11 | 0.85 | WT |
| NCI_H226 | Resistant | 7.36 | -1 | 0.85 | 0 | p.? |

| | | | | | | |
|---|---|---|---|---|---|---|
| UO_31 | Resistant | 8.16 | -1 | 0.6 | 0 | p.M1_ *157del |
| NCI_ADR_RES | Resistant | 13.62 | 0 | 0.58 | 0 | WT |
| ACHN | Resistant | 6.71 | -1 | -1.61 | 0 | p.M1_ *157del |
| OVCAR_4 | Resistant | 11.55 | 0 | -0.14 | 0 | WT |

**Table 3-1:** Gene expression, mutation, DNA copy number variation and CpG methylation of CDKN2A.

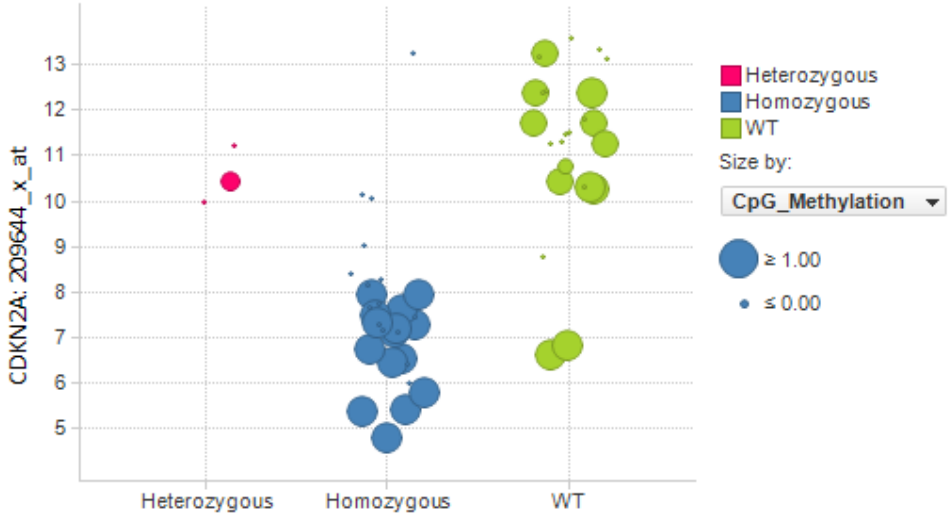



**Figure 3-1**: Jitter plot of the expression of CDKN2A gene (Affymetrix U133A&B array) and the genomic aberrations

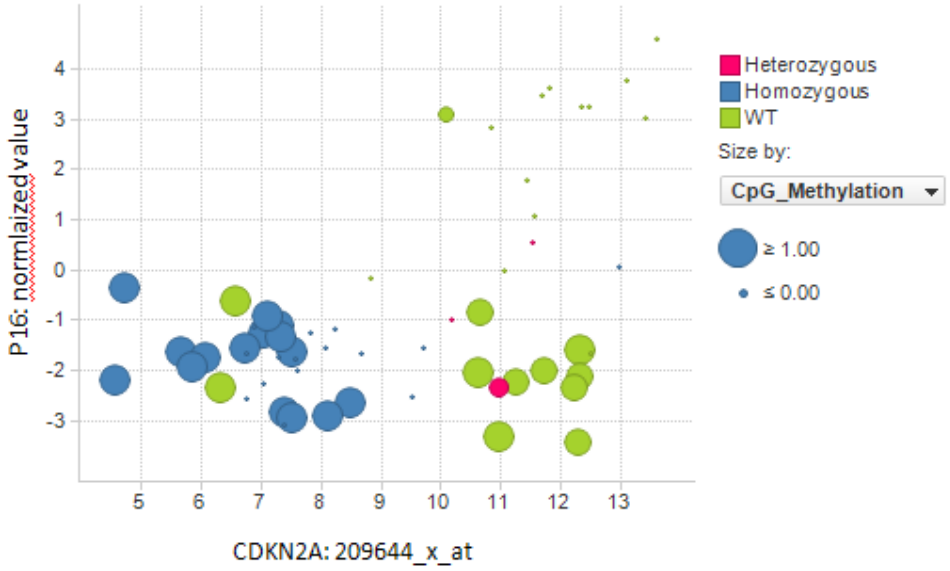



**Figure 3-2**: CDKN2A gene expression (Affymetrix U133A&B array) and CDKN2A coded protein expression, p16 with the genomic aberrations

| Parameters | Estimate | SE | t-value | P-value |
|---|---|---|---|---|
| Intercept | 11.05 | 0.31 | 35.40 | <2e-16 |
| Mut | 2.33 | 0.43 | 5.39 | 2.22e-06 |
| CNV | 0.76 | 0.38 | 1.99 | 0.053 |
| Me.CpG | 1.43 | 0.49 | 2.92 | 0.0054 |

**Table 3-2**: Model parameter estimated for CDKN2A gene in the Paclitaxel type-II gene signature

## 3.3 CDKN2A as the single gene marker for Paclitaxel

We have identified that CDKN2A has sufficient preliminary criteria to be a single gene marker via integrated genomic analysis of gene expression, mutation, DNA copy number variation and methylation. We therefore can evaluate the actual performance of CDKN2A as a single gene marker of Paclitaxel. Firstly, we plotted the jitter plot (Figure 3-1) for CDKN2A expression in 51 NCI60 solid tumor cell lines.

As expected, the Paclitaxel sensitive samples show an over-expression of CDKN2A. Interestingly, the expression of resistant samples is higher than the medium samples. This is understandable, since the original signature is the combination of three sets of significantly differentiated genes. Since the GI50 of Paclitaxel sensitive samples is 10-fold less than the GI50 of Paclitaxel resistant samples, it would be more meaningful to assess the predicted ability of CDKN2A to Paclitaxel sensitive and resistance. Figure 3-3 depicts the AUC plot to use CDKN2A gene expression to predict Paclitaxel sensitive and resistant chemo-response pattern. AUC=0.64 seems indicate the CDKN2A is only a weak single gene predictor of Paclitaxel chemo-response.

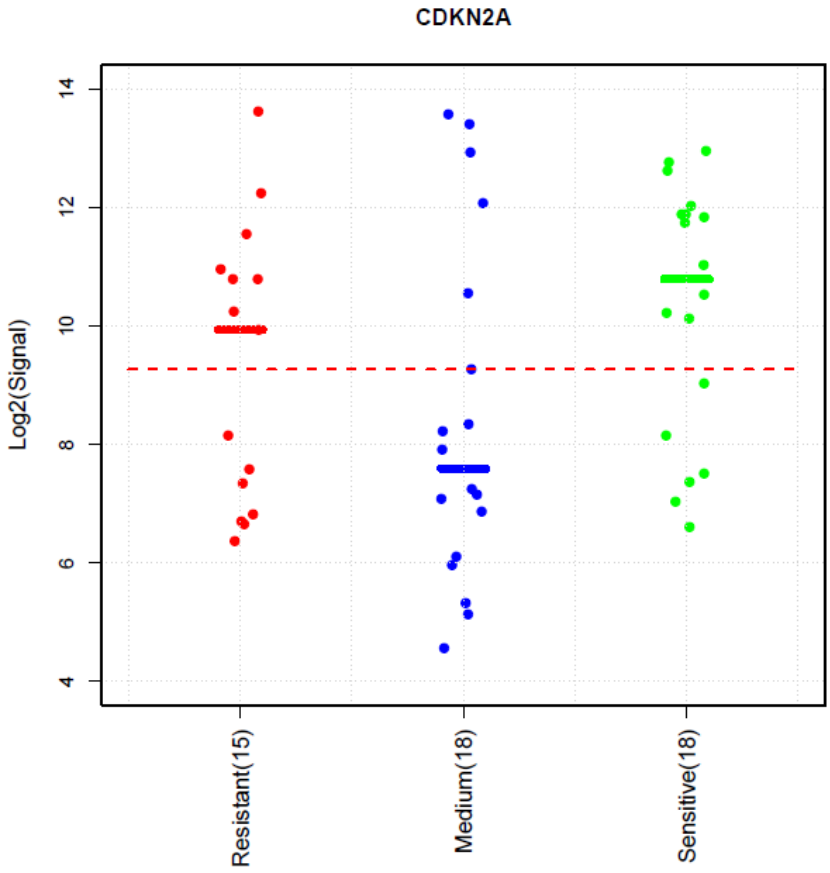



**Figure 3-3**: Jitter plot of the expression of CDKN2A gene (Affymetrix U133A&B array) in the corresponding Paclitaxel chemo-response classes

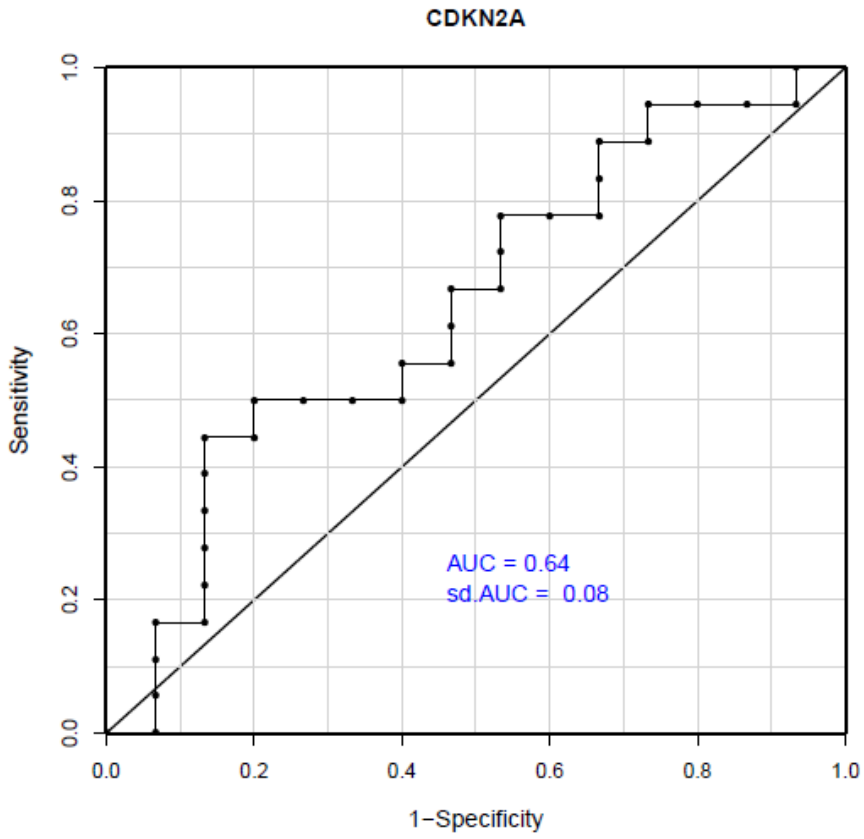



**Figure 3-4**: AUC plot of the expression pattern of CDKN2A in Paclitaxel sensitive and Paclitaxel resistant samples

## 3.3 General strategies to propose minimal marker sets for clinical translation

The integrated analysis identified the principle expression markers may potentially applicable in clinical translation since the expression pattern of each marker can be well explained by the related genomic aberrations. Specifically, N-gene marker may be requested in the clinical applications. Therefore, we suggest three general strategies to narrow down multiple gene signature markers to N-gene marker.

- Strategy-I: Use the gene markers with the gene or protein expression are well explained by the heterogeneous genomic biology using integrated genomic data analysis

- Strategy-II: Use the combination of N-gene marker set which gives minimal misclassification rate
- Strategy-III: Strategy-I + Strategy-II

Strategy-I is the aggregation of N genes, such as CDKN2A, whose expression are significantly affected by the corresponding gene mutation, CpG island methylation and DNA copy number of variation.

Strategy-II is using a searching algorithm to identify the best combination of N genes which gives the minimal misclassification rate. When handling chemo-response gene expression signature, a multinomial probit regression model is usually employed to assign a predicted probability of each response class for each sample. Consequently, the predicted accuracy and error rate is estimated. Table 3-3 depicts the possible 3-gene marker for Paclitaxel type-II gene signature. According to the table, the best combination of 3-genes gives the minimal error rate 0.18 using multinomial probit regression. This is done using R package "vbmp" version 1.20.

| **Gene 1** | **Gene 2** | **Gene 3** | **Error Rate** |
|---|---|---|---|
| 231067_s_at | 202685_s_at | 204780_s_at | 0.18 |
| 203323_at | 202685_s_at | 204584_at | 0.20 |
| 211980_at | 228107_at | 202897_at | 0.22 |
| 214247_s_at | 213906_at | 204086_at | 0.269 |
| 222802_at | 202685_s_at | 218537_at | 0.288 |
| … | … | … | … |

Table 3-3: The combination of 3 genes marker (Strategy-II) narrowed down from the Paclitaxel type-II gene signature

Strategy-III is the combination of Strategy-I and Strategy-II, which means a part of the marker genes are selected from Strategy-I and another part of the marker

genes are selected using Strategy-II. For example, if P (P<N) genes are identified by integrated genomic analysis and are requested to move into clinical for further validation, but the prediction performance is not good, strategy-II can be then followed to identify another Q (Q=N-P) genes and the combined N (P and Q) genes are expected to give a improved prediction accuracy than P genes. Table 3-4 describes the identified 3-gene marker included CDKN2A(209644_x_at) gene for Paclitaxel type-II gene signature. The best 3-gene marker (Include CDKN2A) gives error rate 0.228 that is slightly lower than the error rate reported by 3-gene marker using strategy-II.

| Gene 1 | Gene 2 | Gene 3 | Error Rate |
|---|---|---|---|
| 209644_x_at | 222108_at | 223622_s_at | 0.228 |
| 209644_x_at | 222108_at | 201850_at | 0.268 |
| 209644_x_at | 202686_s_at | 201850_at | 0.268 |
| 209644_x_at | 202685_s_at | 204780_s_at | 0.327 |
| 209644_x_at | 201951_at | 201850_at | 0.327 |
| … | … | … | … |

Table 3-4: The combination of 3 genes marker (Strategy-III) narrowed down from the Paclitaxel type-II gene signature

## 3.4 Summary and discussions

In order to narrow-down the size of multiple-marker signature to a small number(<10) of signature markers, we have suggested an integrated genomic analysis mathematical model to understand how the genetic aberrations regulate the phenotypic expression of the marker gene. CDKN2A is selected as the single gene marker of Paclitaxel, as it shows that its gene expression is significantly affected by mutation and CpG island methylation. However, the AUC analysis has indicated that

this single gene predictor is only a weak predictor. Although the single marker predictor can be quickly adopted into clinical practices, it may show poor predictability when compared with multiple-marker predictors. Alternatively, we suggested three general strategies to identify N-gene (N<10) marker for clinical translations.

# Chapter 4 A genomic signature to characterize concordant response among chemotherapeutics

## 4.1 Introduction

Cancers are complex and heterogeneous diseases characterized by the uncontrolled growth and spread of abnormal cells[186]. One of the most commonly used treatment option for cancer patients is the treatment with an appropriate chemotherapeutic agent or agents. However, one of the biggest challenges associated with chemotherapy is that patients with similar histopathology do not consistently respond the same way to the same agent. Optimizing the choices of anticancer therapy for individual patients using translational research methods is an key challenge to clinical practice[187-189]. For our usage, we adopt the definitions for "translational science" and "translational medicine" mean "new knowledge, mechanisms and techniques generated in basic science research and clinical research that are effectively translated into new approaches for prevention and diagnosis, or new treatment methods for better healthcare"[190].

Molecular markers are well recognized as powerful translational tools that provide guidance for chemotherapy treatment in a clinical setting. Molecular markers may include: a) the gene targets themselves in the case of targeted agents; b) the activity of the targeted pathways; and c) genes only indirectly related with the agents

or targeted pathway. The drug targets themselves are the most important markers for targeted anticancer agents. For example, patients with Her2 amplification show 20% higher response rate to Herceptin than patients with HER2 normal copy number[72, 73]. In addition, the activity of a targeted pathway may also serve as a marker for a drug's clinical response. For example, patients whose tumors contain activating mutations in KRas show worse response than patients with wild type KRas to Cetuximab and Panitumumab, which are fully human monoclonal antibodies specific to the epidermal growth factor receptor (EGFR). This is because an activating KRas mutation will lead to constitutively active downstream signaling of the pathway, rendering EGFR inhibition ineffective[74, 191]. In addition to drug targeted genes and targeted pathways, there are also some molecular markers that are not directly related with the targeted pathway. For example, ABCB1, TOP2A, and BCL2 are candidates of predictive markers used to predict the response to cytotoxic chemotherapy in breast cancer[76].

Many anticancer agents, such as cytotoxic drugs that are used to treat different cancers, have variable outcomes in patients. For example, only about 20% of patients with breast cancer respond therapeutically to the widely used drug taxane (Paclitaxel or Docetaxel) [192]. Chemotherapy agents – especially cytotoxic drugs – also cause unwanted side effects. Cytotoxic drugs work by killing cells which are dividing, and so some non-cancerous cells can be damaged by their action. Since the response rate is often low and toxicity is often high, combinations of a cytotoxic drug and other chemotherapeutic agents have been developed. Some combinational treatments have

become the standard first line treatment in clinical use. Typical examples include patients with breast cancer who may be given the "TFAC" (Paclitaxel(T), 5-Fluorouracil(F), Adriamycin(A) and Cyclophosphamide(C)) treatment, DC (Docetaxel and Cyclophosphamide) treatment, or FAC (5-Fluorouracil, Adriamycin and Cyclophosphamide) treatment in the neo-adjuvant setting. However, there are two challenges which may limit the application of multidrug combinations. First, combinations of multiple drugs have increased toxicity over individual agents. The administration of ineffective chemotherapy agents in combination regiments with relatively higher efficacy rates may increase the probability of side effects and decrease the quality of life for many cancer patients. Second, many tumors develop multidrug resistance (MDR) to multiple chemotherapy agents. This affects patients with a wide variety of blood cancers and solid tumors, including breast, ovarian, lung, and lower gastrointestinal tract cancers. Chemotherapy kills drug-sensitive cells, but leaves behind a higher proportion of drug-resistant cells. As the tumor begins to grow again, chemotherapy may fail because the remaining tumor cells are now resistant. MDR has been found to be correlated to the presence of ATP-binding cassette transmembrane transporter superfamily proteins, like P-glycoprotein, which may expel chemotherapy drugs from the interior of the cell[193].

What is needed, therefore, is a way of predicting the response of the cancer using molecular markers to a treatment (single agent or a multidrug combination) before administering chemotherapy. Such a rational approach to chemotherapy would prevent patients from having to undergo chemotherapy treatments that will not have a

clinically beneficial outcome. Over the last decade, the use of gene expression profiling has changed our understanding of cancer biology and raised the prospect of stratifying patients by predicting response to chemotherapy based on gene expression signatures. Recent research work has shown that the response in pre-clinical materials such as immortalized cell lines can be used to generate genomic signatures (from microarray gene expression data) that are predictive of response to a single drug[194].

The ultimate purpose of our study is to use preclinical material data to establish a systematic framework of anticancer combinations which may be effectively administrated to cancer patients. Using a translational genomic method similar to previous work[195], we develop genomic signatures for predicting response to cytotoxic agents and assess the prediction results of the signature in cell lines and patient tumor samples. The preclinical material includes cell lines and assay data from NCI-DTP (http://dtnci.nih.gov/) as well as primary human tumor derived explants from Oncotest (Germany)[7, 196], that has established a large collection of primary human tumor xenografts growing subcutaneously in nude mice. The primary xenografts retain many of the characteristics of the parental patient tumors including histology and sensitivity to anticancer drugs and to a high extent recapitulate the response of the donor patient to standard anticancer drugs[7]. Furthermore, an *in vitro* tumor clonogenic assay (TCA; inhibition of colony formation ability of cells that show anchorage independent growth in soft agar) performed on explant material derived from the *in vivo* models is highly predictive of *in vivo* response)[7, 196].

Predictive biomarkers of response to a particular agent may be difficult to distinguish from prognostic biomarkers that predict the outcome of the general treatment in a clinical setting, especially when that agent is used in combination. Using preclinical assays of sensitivity and translating genomic signatures to predict clinical response may partly address this challenge, while raising additional challenges of imperfect preclinical models. We hypothesized that the ability to predict response with high specificity for individual chemotherapeutic cytotoxic agents may be hampered by concordant chemo-sensitivity or chemoresistance in preclinical models. In this study, through an integrated analysis of basal microarray expression data and preclinical response, we identified a 168 gene expression signature for concordance of chemotherapeutic sensitivity. We first define the concordance of chemotherapeutics among 14 selected cytotoxic Standard of Care (SOC) anticancer agents. We then characterize the concordance of chemotherapeutics using a gene expression signature with robust *in vitro* validation using an independent data set. We next detect the developed signature to be present in the clinical patients who were treated with (TFAC) combination. This signature of concordant chemotherapeutics also shows a prognostic value to identify long survival patient groups in breast cancer and lung cancer.

## 4.2 Results

### 4.2.1 Concordant chemotherapeutics across 14 cytotoxic agents

The starting point of this study is the observation of chemotherapeutic response in the established *in vitro* tumor models, such as breast, ovarian, colon, renal, lung, prostate, and melanoma cancers. We focused on two independent *in vitro* drug sensitivity datasets and explored the patterns of concordant sensitivity that arise within them. The NCI-DTP has evaluated more than 10,000 compounds for evidence of the ability to inhibit cancer cell growth using a sulforhodamine B (SRB) assay in 60 human tumor cell lines. Oncotest has investigated the inhibition of anchorage independent growth activity of over 100 anticancer agents using an in vitro Tumor Clonogenic Assay (TCA) in more than 200 tumor explant models. In this study, we analyze in vitro NCI's SRB assay drug sensitivity data in 54 cancer cell lines (we exclude leukemia cell lines) and Oncotest's TCA drug sensitivity data in 52 solid tumor explant models on 14 anticancer cytotoxic agents (Table 4-1). The selected tumor types in two datasets are similar – both include breast, colon, lung, melanoma, ovarian and renal cancers – but none of the individual models appears in both data sets. Leukemia cell lines and explant models are excluded in the study as they have previously demonstrated elsewhere to show high sensitivity to multiple agents in NCI60 data[197], whereas other tumor types are seen to show divergent response to multiple agents in both datasets.

| Cytotoxic Agents | Types |
|---|---|
| CCNU | Alkylating/alkylating-like |
| Cisplatin | Alkylating/alkylating-like |
| Oxaliplatin | Alkylating/alkylating-like |
| FU | Antimetabolites |
| Gemcitabine | Antimetabolites |
| Doxorubicin | Cytotoxic/antitumor antibiotic |
| Mitoxantrone | Cytotoxic/antitumor antibiotic |
| Docetaxel | Spindle poisons/mitotic inhibitors |
| Paclitaxel | Spindle poisons/mitotic inhibitors |
| Vinblastine | Spindle poisons/mitotic inhibitors |
| Vincristine | Spindle poisons/mitotic inhibitors |
| Etoposide | Topoisomerase inhibitors |
| Irinotecan | Topoisomerase inhibitors |
| Topotecan | Topoisomerase inhibitors |

**Table 4-1**: 5 types of standard of care chemotherapy agents included in the study are: alkylating/alkylating-like(3), antimetabolites(2), antitumor antibiotic(2), spindle poison/mitotic inhibitor(4) and topoisomerase inhibitor(3)

In order to identify the pattern of *in vitro* chemo-response for 14 anticancer cytotoxic agents, we classified the samples into three classes for each agent: Sensitive, Medium and Resistant. The chemo-response data (GI50 and IC50) is negatively log-transformed (NLog) and discretized into three levels using an agglomerative clustering method. This "discretization level coalescence" method incrementally bins the number of discretization levels for the chemo-response data of each agent while minimizing the loss of total mutual information between the agents[198]. In this case, only 5% of mutual information was lost. We then define a Concordance Rate to

represent the concordance of chemotherapeutics within a class for each sample across all 14 cytotoxic agents. The Concordance Rate (Sensitive) is defined as:

$$\text{Concordance Rate}(\text{Sensitive}) = \frac{\sum_{i=1}^{N} I(V_i)}{N}, \quad I(V_i) = 1, \ if\ V_i = \text{Sensitive}, \ else\ 0 \qquad \text{Eq 4-1}$$

Similarly, we define Concordance Rates for the Medium and Resistant classes.

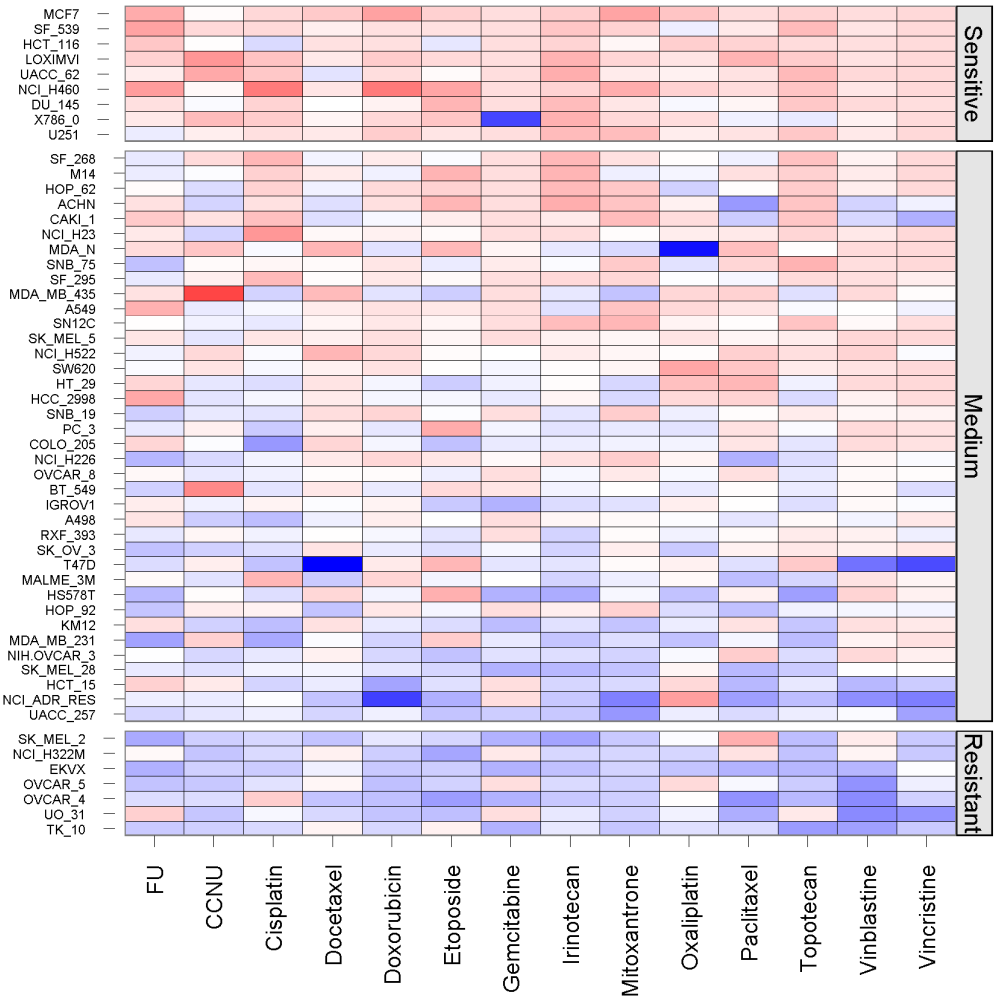
We next performed the analysis of estimated concordance rate for both NCT-DTP and Oncotest drug sensitivity data in the selected anticancer cytotoxic agents. The analysis of concordance rates based on 54 NCI human cell lines for 14 cytotoxic agents reveals that 8/54(15%) of cell lines show a Concordance Rate(Sensitive) of 67% or higher, including 1 breast (MCF7), 1 CNS (SF_539), 1 colon (HCT116), 2 melanoma (LOXIMVI, UACC_62), 1 non-small cell lung cancer (NCH_H460) and 1 prostate (DU_145). 7/54 (13%) of cell lines show a Concordance Rate(Resistant) of 67% or higher, including include 1 melanoma (SK_MEL_2), 2 non-small cell lung cancer (NCI_H322M, EKVX), 2 ovarian ( OVCAR_4, OVCAR_5) and 2 renal (UO_31, TK_10). The difference of NLogGI50 between concordant sensitive samples and concordant resistant samples represent more than 10-fold differences in sensitivity. For example, the median of NLogGI50s of concordant sensitive cell line NCI_H460 among 14 cytotoxic agents is 8.07, while the median of NLogGI50s of concordant resistant cell line TK_10 is only 5.89.

The analysis of concordance rates based on 52 Oncotest explant models in TCA for 14 cytotoxic agents shows that 8/52(15%) samples with a Concordance Rate (Sensitive) of 67% or higher, including 1 breast (MAXF_401), 1 multiple myeloma

(MMXF_L363), 3 non-small cell lung cancer (LXFL_1121, LXFA_983, LXFE_1422), 1 prostate (PRXF_22RV1LX), 1 bladder (BXF_1218) and 1 uterine (UXF_1138LX). 8/52 (15%) samples show a Concordance Rate (Resistant) of 67% or higher, including 1 colon (CXF_975), 3 non-small cell lung cancer (LXFA_289, LXFA_1041, LXFA_297), 3 ovarian (OVXF_899, OVXF_550, OVXF_1023) and 1 renal (RXF_423). The most sensitive sample, UXF_1138LX, shows a median NLogIC50 of 8, which is almost 1,000 fold over the most resistant sample LXFA_289 with a median NLogIC50 of only 5. Figure 4-1 and Figure 4-2 depict the normalized chemo-response data of NCI-DTP and Oncotest respectively with the concordant sensitive and concordant resistant samples to 14 cytotoxic agents.

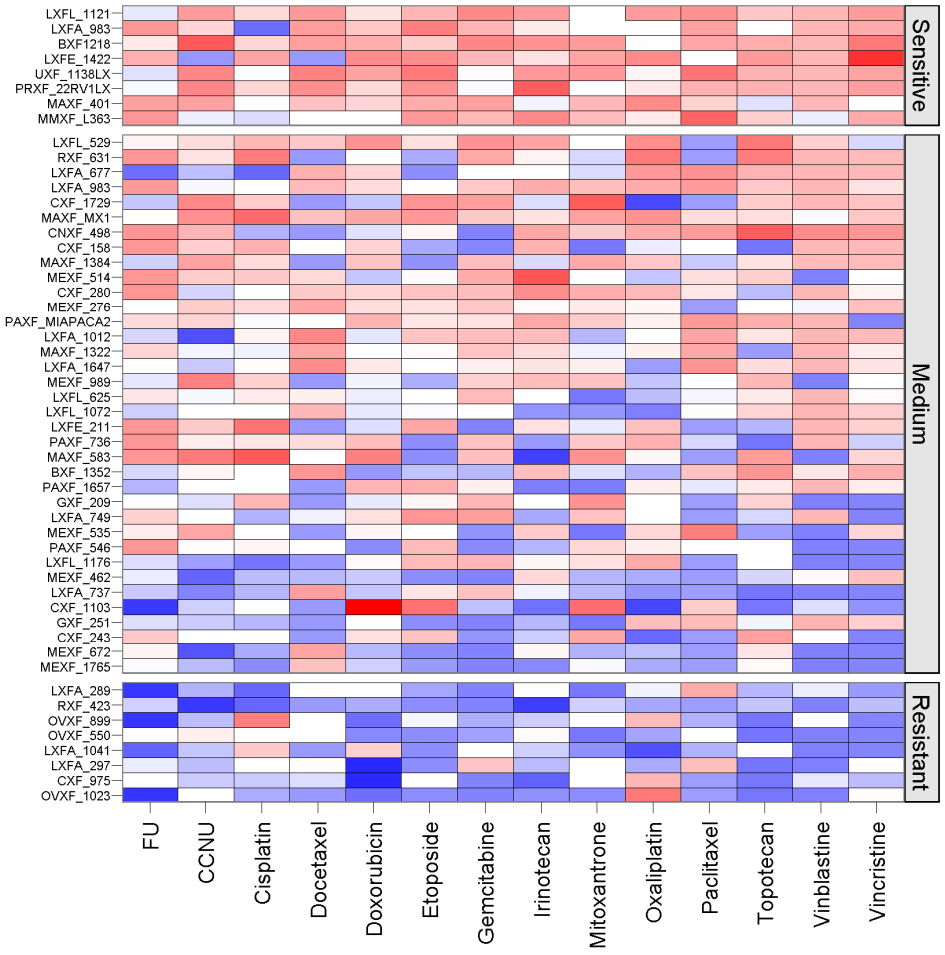
| NCI60 | Tissue | Concordance Rate (%) | Oncotest | Tissue | Concordance Rate (%) | Classes |
|---|---|---|---|---|---|---|
| MCF7 | Breast | 100 | LXFL_1121 | Lung | 85 | Sensitive |
| SF_539 | CNS | 86 | LXFA_983 | Lung | 82 | Sensitive |
| HCT_116 | Colon | 71 | BXF_1218 | Bladder | 79 | Sensitive |
| LOXIMVI | Melanoma | 93 | LXFE_1422 | Lung | 77 | Sensitive |
| UACC_62 | Melanoma | 78 | UXF_1138LX | Uterine | 77 | Sensitive |
| NCI_H460 | Lung | 93 | PRXF_22RV1LX | Prostate | 71 | Sensitive |
| DU_145 | Prostate | 92 | MAXF_401 | Breast | 67 | Sensitive |
| X786_0 | Renal | 69 | MMXF_L363 | Multi-Myeloma | 67 | Sensitive |
| SK_MEL_2 | Melanoma | 71 | LXFA_289 | Lung | 67 | Resistant |
| NCI_H322M | Lung | 71 | RXF_423 | Renal | 71 | Resistant |
| EKVX | Lung | 93 | OVXF_899 | Ovarian | 73 | Resistant |
| OVCAR_5 | Ovarian | 71 | OVXF_550 | Ovarian | 73 | Resistant |
| OVCAR_4 | Ovarian | 71 | LXFA_1041 | Lung | 75 | Resistant |
| UO_31 | Renal | 71 | LXFA_297 | Lung | 75 | Resistant |
| TK_10 | Renal | 93 | CXF_975 | Colon | 82 | Resistant |
| | | | OVXF_1023 | Ovarian | 92 | Resistant |

**Table 4-2**: Summary of concordant sensitive and concordant resistant cell lines/explants in NCI60 and Oncotest for 14 anticancer agents. 15 NCI-DTP cell lines show >67% concordance rates and 16 Oncotest models show >67% of concordance rates.




red: positive, white: zero, blue: negative

**Figure 4-1**: Concordant chemotherapeutics is observed in NCI-DTP 60 cell lines' screening data in SRB assay for 14 cytotoxic agents




red: positive, white: zero, blue: negative

**Figure 4-2**: Concordant chemo-sensitivity is observed in Oncotest explants' screening data in TCA for 14 cytotoxic agents

### 4.2.2 A novel gene expression signature to characterize concordance of chemotherapeutics

From the NCI-DTP drug sensitivity data, we used the 15 cell lines with ≥67% Concordance Rate (8 Sensitive + 7 Resistant) as evidence for high concordance among chemotherapeutics, and explore to see if there was any molecular distinction between them that is visible in genomic expression data. We tagged these 15 cell lines

as members of a “CS” (concordant sensitive) group and a “CR” (concordant resistant) group, respectively. Differential gene expression analysis (see methods and materials) was then employed to develop a gene signature to characterize the concordant groupings as a phenotype, predicted using basal gene expression profiles available from GEO (GSEA5720)[199]. The gene signature consists of 168 genes (176 probesets) that are significantly differentially expressed between concordant sensitive samples and concordant resistant samples (Figure 4-3, Appendix-1). Using a metagene method with a Bayesian binary regression procedure (see materials and methods), each sample is assigned a probability of sensitivity (POS) between 0 and 1 by the signature predictor. A box plot of the probability of sensitivity clearly shows the pattern of concordant sensitivity and concordant resistance among 15 samples (Figure 4-4).

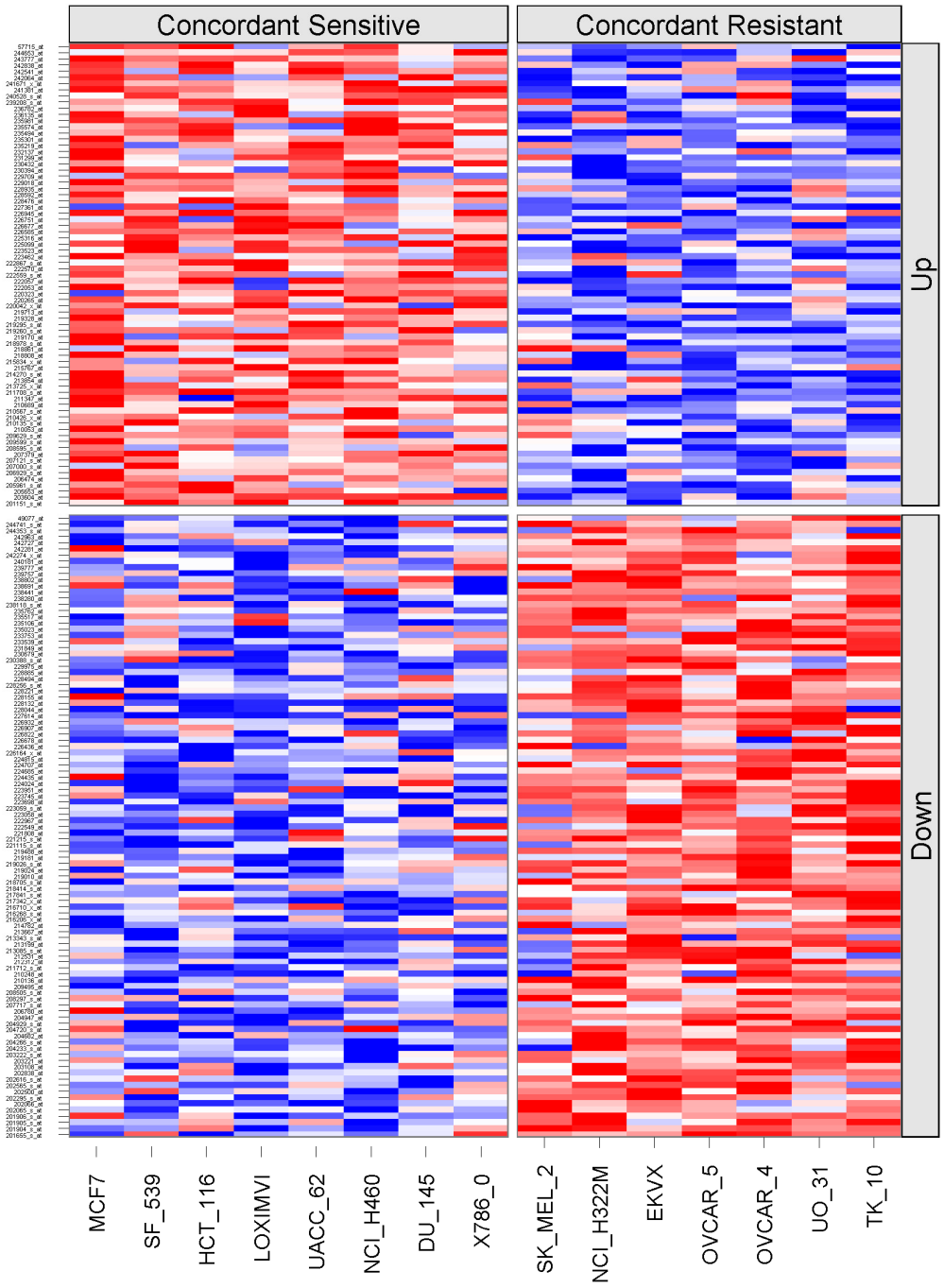
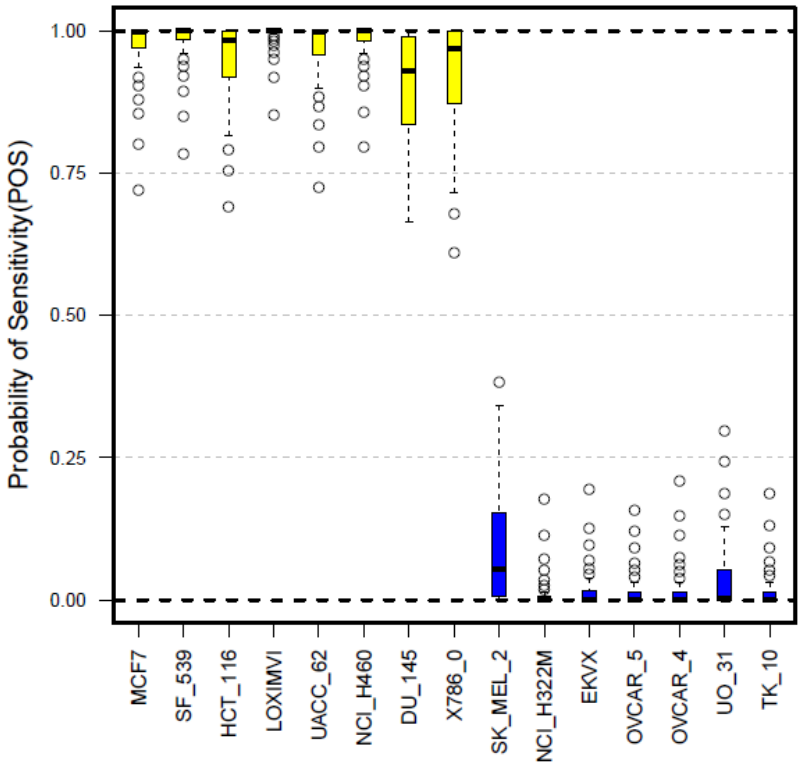



**Figure 4-3**: Heatmap of gene expression signature of concordant chemotherapeutics (red: high, white: medium, blue: low)




yellow: concordant sensitive cell lines; blue: concordant resistant cell lines

**Figure 4-4**: Boxplot of in-sample predicted (fitted) probability of sensitivity of the gene signature predictor of concordant chemotherapeutics in 15 NCI60 cell lines

## 4.23 Independent validation in Oncotest explants models

Subsequently, we then tested to assess if the signature predictor of 168 genes differentially expressed in concordant sensitive versus concordant resistant cell lines can predict concordant chemo-response in tumor derived explants from the Oncotest data. The hypothesis is that if the expression pattern of the signature genes remains significantly intact and non-random, then the signature predictor will be able to predict the concordance of chemotherapeutics for multiple cytotoxic anticancer agents in both cell lines and tumor model derived explants. Using the basal gene expression

data generated by Oncotest, we applied the predictor to 16 explants tumor models that include 8 concordant sensitive explants and 8 concordant resistant explants. The predicted probability of sensitivity (POS) of 16 explants tumor models is shown in Figure 4-5. The performance of the predictor is AUC=0.88±0.10 (Figure 4-6), which is significantly greater than 0.5 and suggests that the signature predictor is non-random. Figure 4-7 depicts the ECDF plot, which shows the robustness of the specified POS cutoff. For example, when the cutoff of POS moved from 0.35 to 0.94, 6/8 (75%) concordant sensitive explants models(symbol:: triangle) and 8/8 (100%) concordant resistant explants models (symbol:: plus) were identified. This broad cutoff of probability of sensitivity (POS) also maximizes the sum of true positive rate and true negative rate. Therefore, the performance of the signature predictor of concordant chemotherapeutics in the independent *in vitro* validation using the Oncotest explant materials suggests that the genomic signature is valid and robust for pre-clinical solid tumor models of human tumors.

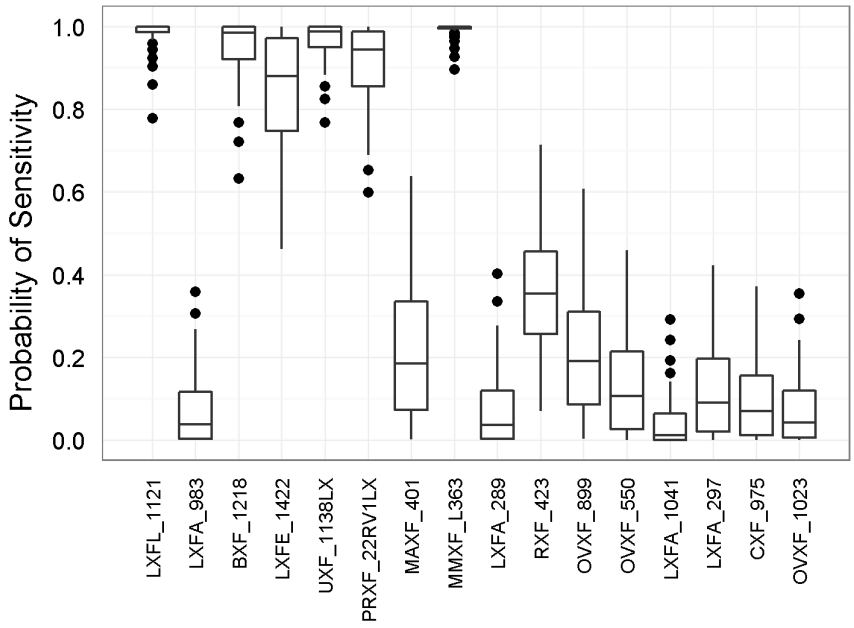



**Figure 4-5**: Boxplot of predicted probability of sensitivity of applying the gene signature predictor of concordant chemotherapeutics in 16 Oncotest explants models

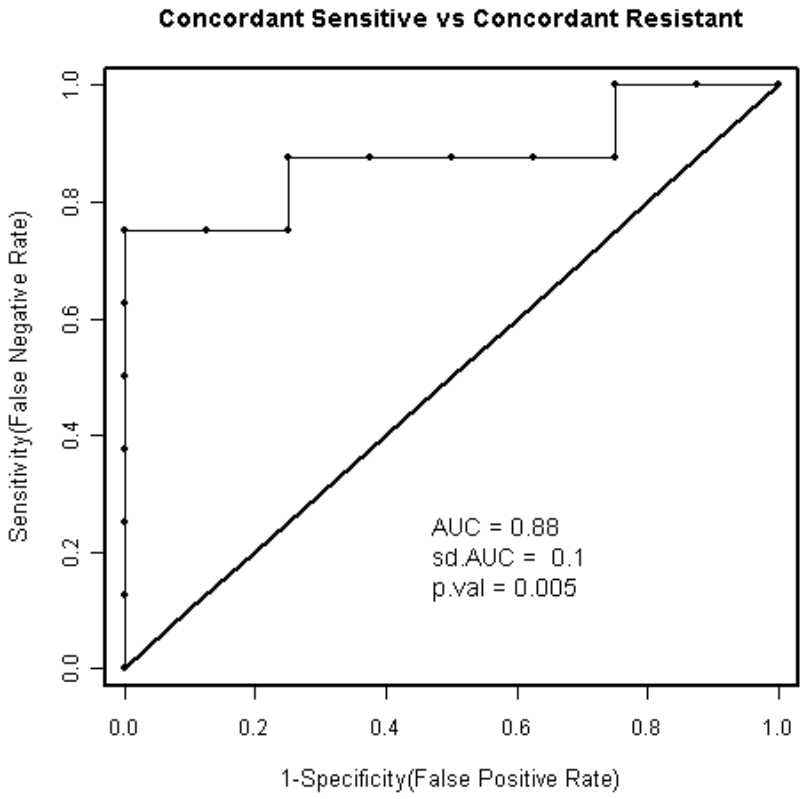
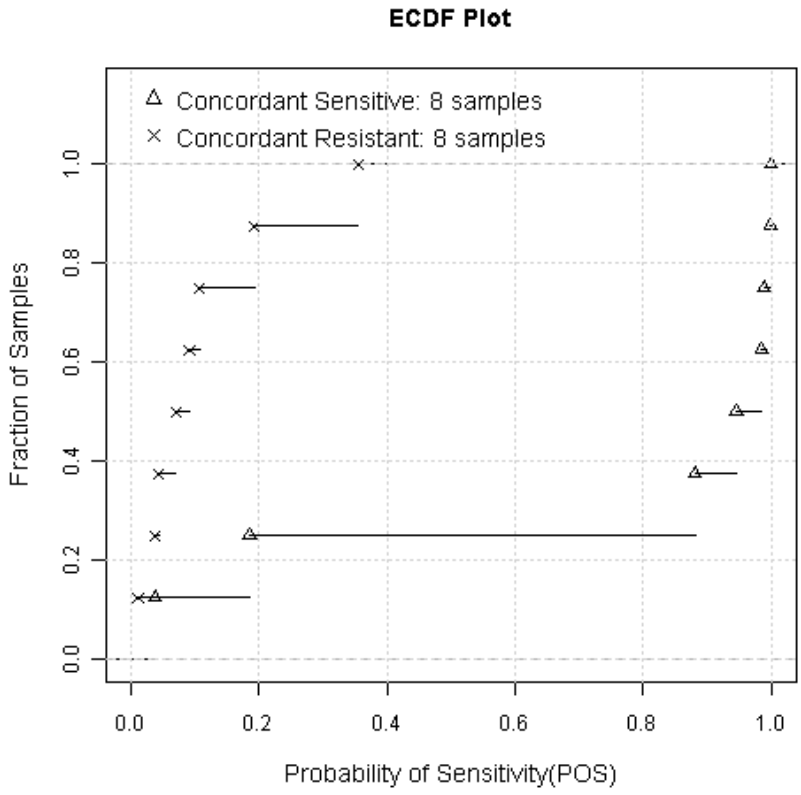



**Figure 4-6**: Receiver Operating Characteristic (ROC) curves of applying the gene signature predictor of concordant chemotherapeutics in 16 Oncotest explants models




X-Axis: probability of sensitivity (POS), Y-Axis: the fraction of samples for POS is less than the cut-off value

**Figure 4-7**: Empirical cumulative distribution of predicted probability of sensitivity (POS) of applying gene signature predictor of concordant chemotherapeutics in 16 Oncotest explants models

### 4.2.4 Independent validation in clinical patients treated with (TFAC) combination chemotherapy

To assess whether the signature of concordance was detectable in patient tumor samples and had potential clinical utility, we investigated whether the signature was able to show predictive ability for combination chemotherapy from published studies. Microarray expression data was used from a clinical cohort of 133 breast cancer patients undergoing neoadjuvant treatment with the TFAC (paclitaxel,

5-fluorouracil, adriamycin, and cyclophosphamide) chemotherapy regimen[200]. To further dissect the general resistance TFAC chemotherapeutics, we apply our gene signature predictor of concordant chemotherapeutics to this breast cancer datasets of (126) tumor samples with clinical treatment results available (32 pathological complete response (pCR) and 94 residual invasive cancer (RD) responses. To follow standard quality control practices for clinical microarray data[201-203], we preprocess the microarray data for quality standards by removing low expression, low variation and poor quality gene level annotated probesets. The original 168-gene signature is therefore reduced to a signature of 23 genes (25 probsesets) in this particular breast cancer dataset. The performance of the signature predictor is AUC=0.63±0.06 (Figure 4-8). Although the AUC value is not very high compare with the performance in Oncotest explants data, "p-value=0.017" clearly demonstrates that the signature predictor is significantly non-randomly present in TFAC treated breast cancer patients. Hence, we can conclude that the developed gene expression signature of concordant chemotherapeutics is a robust signature and it is translatable across tumor types and clinical patients.

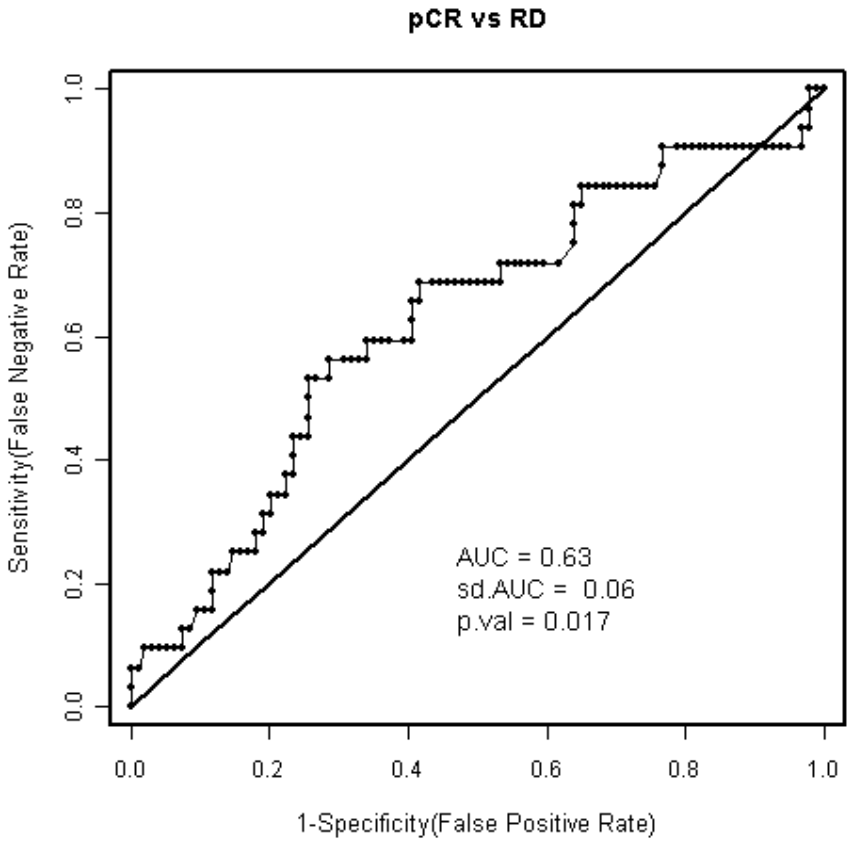
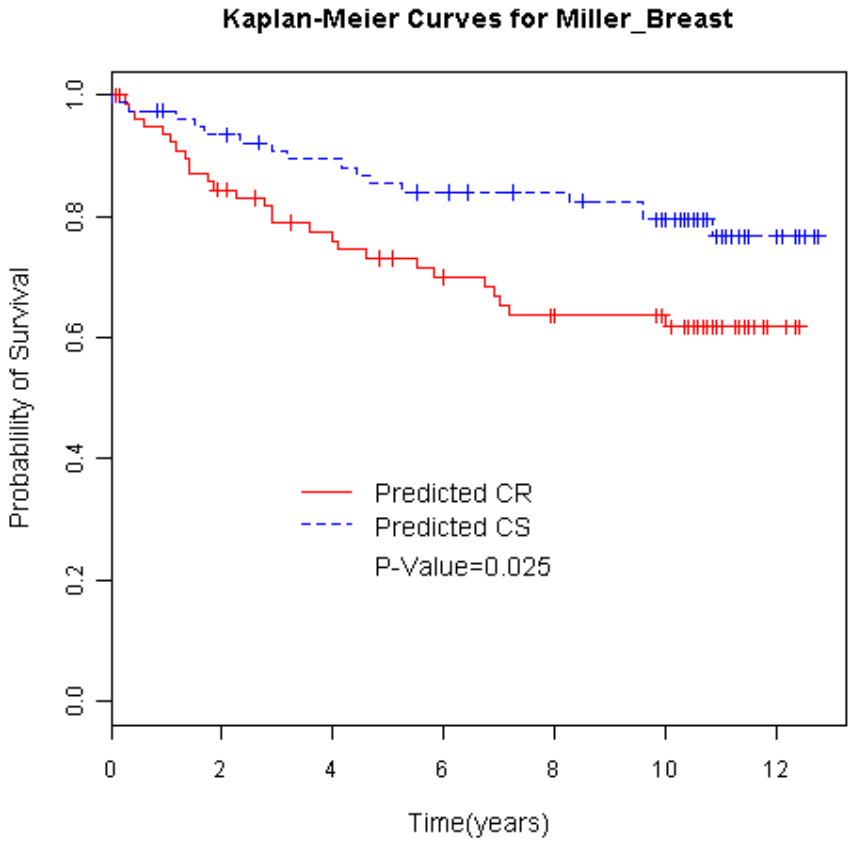



**Figure 4-8**: Receiver Operating Characteristic (ROC) curves of applying the concordant chemotherapeutics gene signature predictor in 126 breast cancer data with patients treated by (TFAC) combinational chemotherapy

### 4.2.5 Prediction of clinical outcome in cancer patients

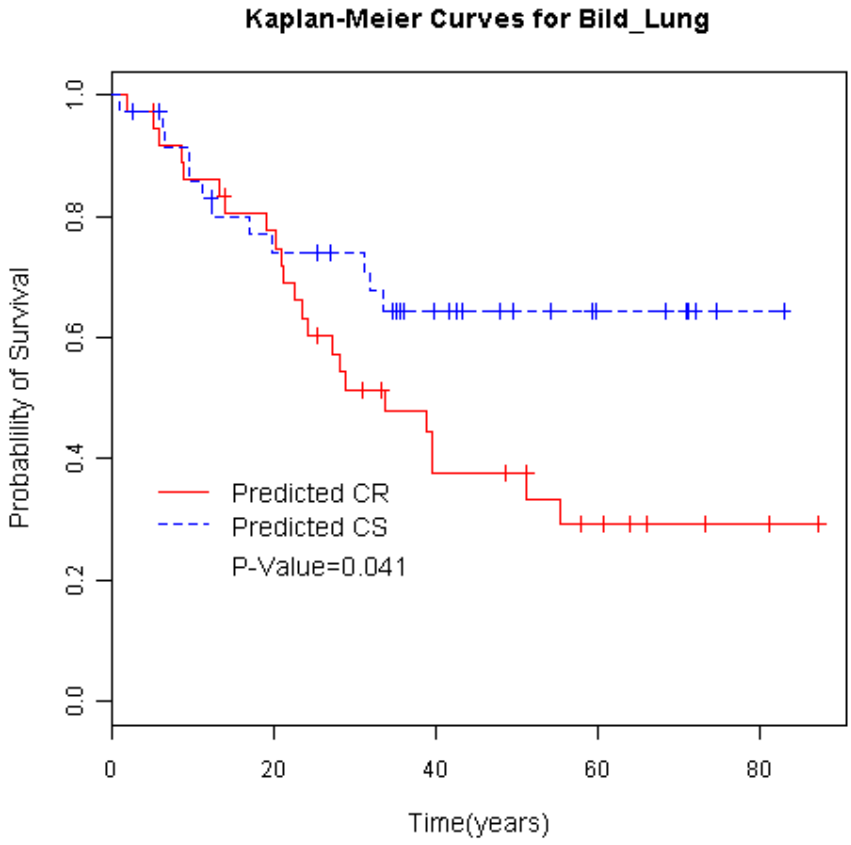
We next examined the ability of the signature predictor of concordant chemotherapeutics to predict clinical outcome in the clinical datasets: Miller breast[204] and Bild lung cancer data sets[141]. The patients' samples are stratified into three groups based on predicted probability of sensitivity of concordant chemotherapeutics: the samples with top 1/3s highest predicted probability of sensitivity of concordant chemotherapeutics are classified as "predicted CS (concordant sensitive)" group; and the samples with the lowest 1/3s of predicted

probability of sensitivity of concordant chemotherapeutics are classified as the "predicted CR (concordant resistant)" group; and the rest of the samples with medium predicted probability of sensitivity of concordant chemotherapeutics are classified as the "predicted medium" response or "divergent" response group. The Kaplan-Meier survival analysis is then performed for the two stratified patients "predicted CS" and "predicted CR" groups. Figure 4-9 shows the Kaplan–Meier survival curves for the 156 breast cancer patients. The predicted CS group patients show much better survival than predicted CR group patients with the log-rank test p-value=0.025. Figure 4-10 depicts the survival analysis for 74 lung cancer patients. The predicted CS group patients show significantly better survival than predicted CR group patients. These indicate that the presence of gene signature of concordant chemotherapeutics in both breast cancer and lung cancer patients. In clinic, the platins and taxanes are the two important standards of care cytotoxic agents, and they are widely used for the treatment of breast cancer and lung cancer patients. The strong association between the clinical outcome of stratified breast cancer and lung cancer patients by the signature of concordant chemotherapeutics signify that the gene signature predictor of concordant chemotherapeutics may be potentially useful for tailored chemotherapies in clinic for solid tumors.




Predicted CS: predicted "concordant sensitive" group; predicted CR: predicted "concordant resistant" group; P-Value: Log-Rank test probability (null hypothesis: two groups have no difference)

**Figure 4-9**: Kaplan–Meier survival curves of stratified predicted "Concordant Sensitive" and predicted "Concordant Resistant" breast cancer patients by the gene expression signature predictor of concordant chemotherapeutics. The 78 predicted CS patients showed a significantly longer disease-free survival time than the 78 predicted CR patients (P-Value<0.05).



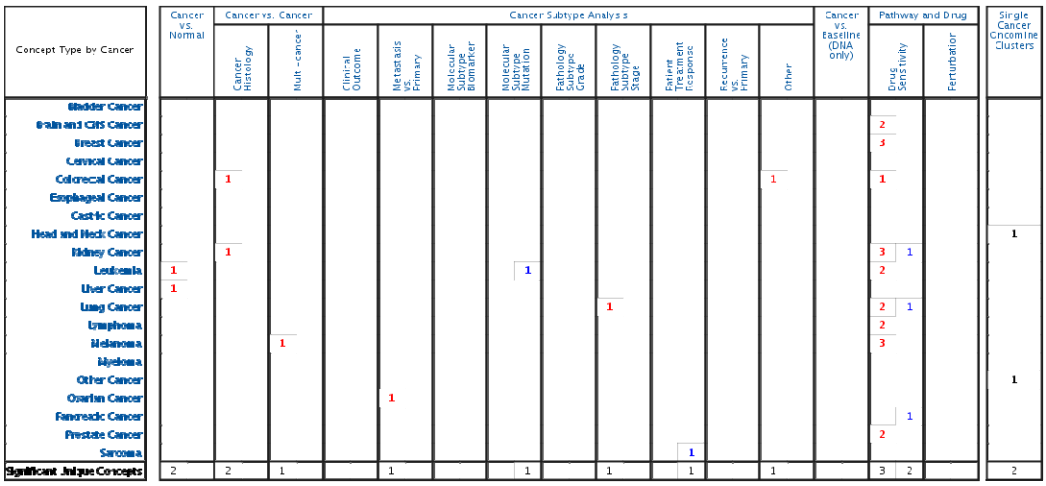

Predicted CS: predicted "concordant sensitive" group; predicted CR: predicted "concordant resistant" group; P-Value: Log-Rank test probability (null hypothesis: two groups have no difference)

**Figure 4-10**: Kaplan–Meier survival curves of stratified predicted "Concordant Sensitive" and predicted "Concordant Resistant" lung cancer patients by the gene expression signature predictor of concordant chemotherapeutics. The 37 predicted CS patients showed a significantly longer disease-free survival time than the 37 predicted CR patients (P-Value<0.05).

### 4.2.6 Meta-analysis for correlations with other drug sensitivity signatures and mechanism study of sensitivity

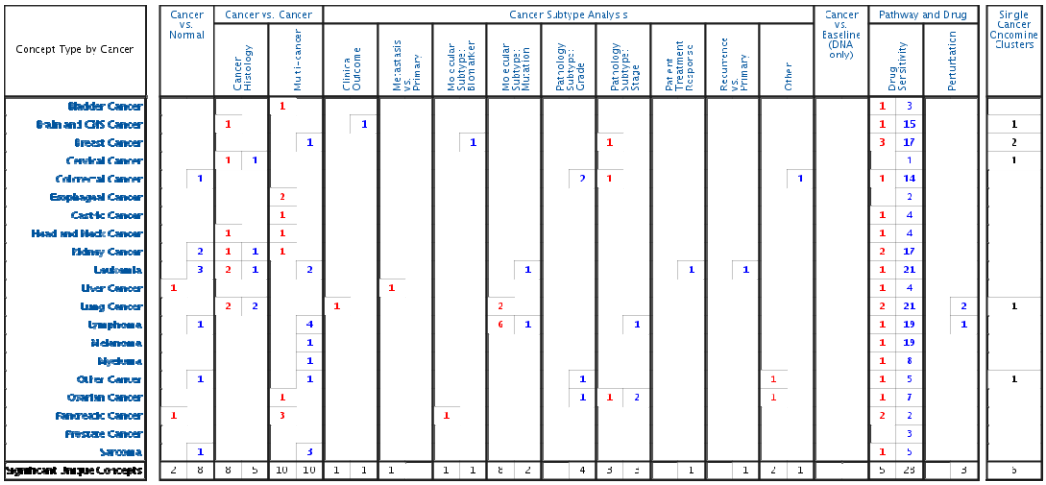
Since the concordance signature is proposed to represent concordance among multiple anticancer agents excluding the selected 14 agents, we investigated if the

signature genes for concordant sensitivity are correlated with sensitivities to other anticancer agents using Oncomine database analysis tools[205, 206]. We built two signature concepts in the Oncomine system: signature concept-a) 75 genes that are over expressed in concordant sensitive samples; signature concept-b) 93 genes that are under expressed in concordant sensitive samples. Interestingly, concept a) is only significantly associated with 17 signature concepts (Table 4-3), that include 5 unique drug sensitivity concepts (3 over expression, 2 under expression). However, concept b) is significantly associated with 109 signature concepts (Table 4-4), which include 33 unique drug sensitivity concepts (5 over expression, 28 under expression). Furthermore, the under-expression genes in the proposed signature are highly enriched in many other drug sensitivity concepts including Amsacrine, Floxuridine, Methodtrexate, Teniposide, Topotecan, Tremetrexate, Cytarabine and Temsirolimus "sensitive multi-cancer cell lines concepts" (Figure 4-11), that means the signature concept formed by under-expression genes of concordant chemotherapeutics is able to predict the chemo-response of other anticancer agents. For example, Figure 4-12 and Figure 4-13 describe the signature concept-b) formed by under-expressed genes shows consistent expression patterns in Amsacrine and Temsirolimus sensitive cell lines respectively. The NCI-DTP drug sensitivity data of Amscrine and Temsirolimus also shows apparent sensitive and resistant patterns in 15 concordant sensitive and concordant resistant cell lines (Figure 4-14 and Figure 4-15).



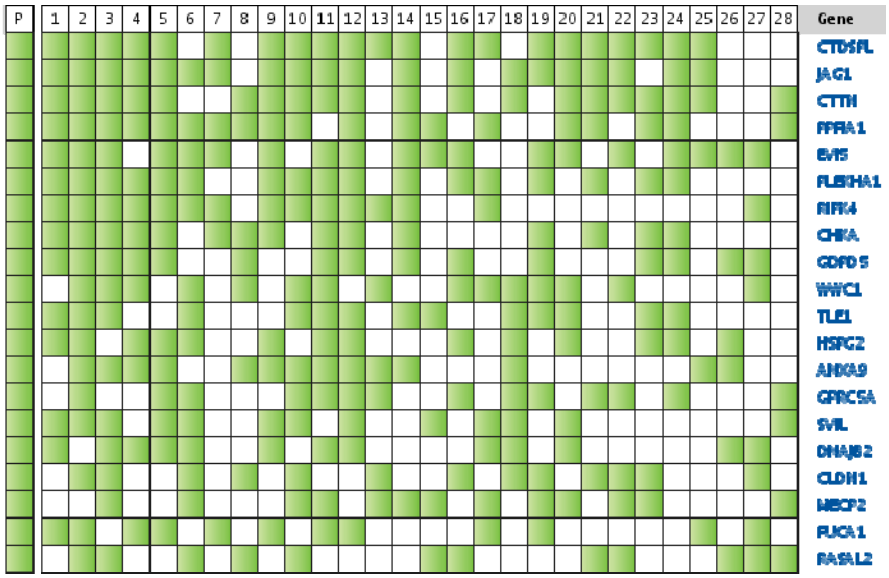

Threshold: odds ratio=2, p-value=1.0e-4, Red: over-expression, Blue: under-expression

**Table 4-3:** The associated concept summary for 75 over-expressed signature genes in concordant sensitive cell lines. Image is from Oncomine™.




Threshold: odds ratio=2, p-value=1.0e-4, Red: over-expression, Blue: under-expression

**Table 4-4:** The associated concept summary for 93 under-expressed signature genes in concordant sensitive cell lines. Image is from Oncomine™.




*P: My Concepts:* under-expressed signature genes of concordant chemotherapeutics
1. Amsacrine Sensitive - Top 1% Under-expressed (Compendia CellLine)
2. Floxuridine Sensitive - Top 10% Under-expressed (Compendia CellLine)
3. Methotrexate Sensitive - Top 10% Under-expressed (Compendia CellLine)
4. Teniposide Sensitive - Top 1% Under-expressed (Compendia CellLine)
5. Topotecan Sensitive - Top 5% Under-expressed (Compendia CellLine)
6. Trimetrexate Sensitive - Top 10% Under-expressed (Compendia CellLine)
7. Mitomycin C Sensitive - Top 1% Under-expressed (Compendia CellLine)
8. Foretinib Sensitive - Top 10% Under-expressed (Wooster CellLine)
9. Cytarabine Sensitive - Top 5% Under-expressed (Compendia CellLine)
10. Temsirolimus Sensitive - Top 10% Under-expressed (Wooster CellLine)
11. Doxorubicin Sensitive - Top 10% Under-expressed (Compendia CellLine)
12. Mitoxantrone Sensitive - Top 10% Under-expressed (Compendia CellLine)
13. Pazopanib Sensitive - Top 5% Under-expressed (Wooster CellLine)
14. Bleomycin Sensitive - Top 10% Under-expressed (Compendia CellLine)
15. Fluorouracil Sensitive - Top 5% Under-expressed (Compendia CellLine)
16. GSK1070916 Sensitive - Top 10% Under-expressed (Wooster CellLine)
17. GSK1904529 Sensitive - Top 10% Under-expressed (Wooster CellLine)
18. SN-38 Sensitive - Top 10% Under-expressed (Gemma CellLine)
19. Dihydro-5-Azacytidine Sensitive - Top 10% Under-expressed (Compendia CellLine)
20. Ormaplatin Sensitive - Top 10% Under-expressed (Compendia CellLine)
21. BEZ235 Sensitive - Top 10% Under-expressed (Wooster CellLine)
22. Mercaptopurine Sensitive - Top 10% Under-expressed (Compendia CellLine)
23. Bleomycin Sensitive - Top 10% Under-expressed (Compendia Melanoma CellLine)
24. Mechlorethamine Sensitive - Top 5% Under-expressed (Compendia CellLine)
25. Thioguanine Sensitive - Top 5% Under-expressed (Compendia CellLine)
26. SN-38 Sensitive - Top 10% Under-expressed (Shimokuni CellLine 2)
27. Paclitaxel Sensitive - Top 10% Under-expressed (Compendia CellLine)
28. GSK1070916 Sensitive - Top 10% Under-expressed (Wooster Liver CellLine)

**Figure 4-11**: Comparison of shared genes across 29 gene signature concepts (under expression genes in concordant sensitive cell lines and 28 Oncomine chemo-sensitivity signature concepts). Figure is from Oncomine™.

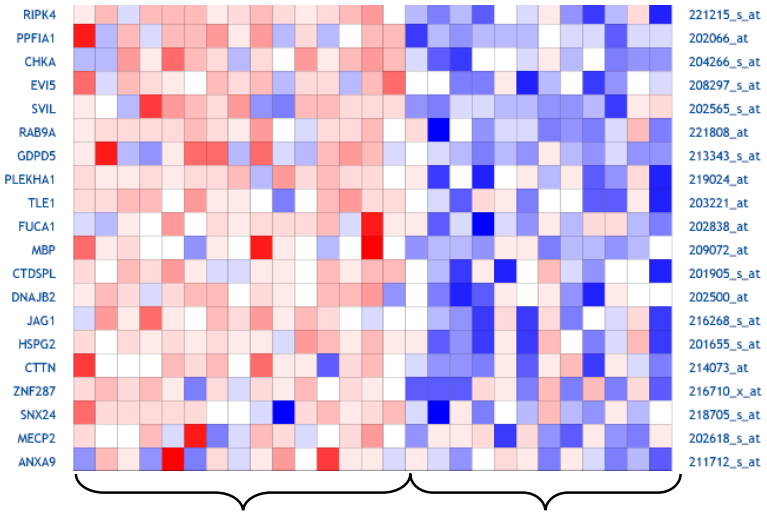



**Figure 4-12**: Signature concept of under expressed signature genes (20) in Compedia cell lines show consistent expression pattern as Amsacrine in vitro drug sensitivity profile (Oncomine$^{TM}$).

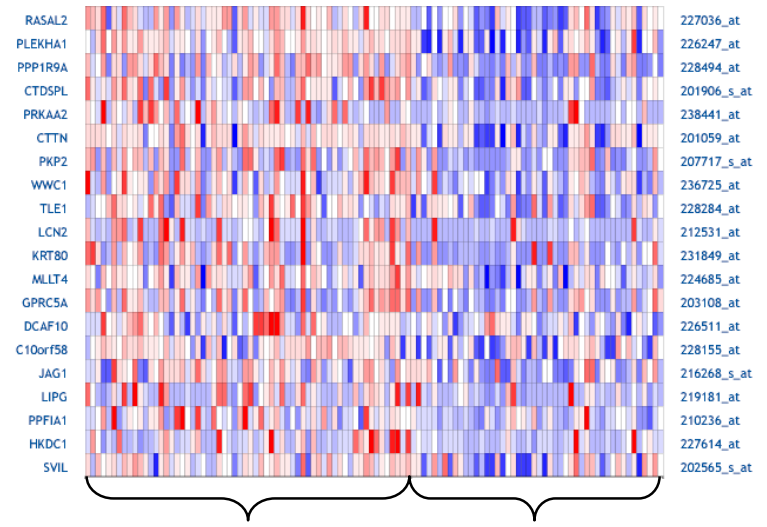



**Figure 4-13**: Signature concept of under expressed signature genes (20) in Wooster cell lines show consistent expression pattern as Temsirolimus in vitro drug sensitivity profile (Oncomine$^{TM}$).

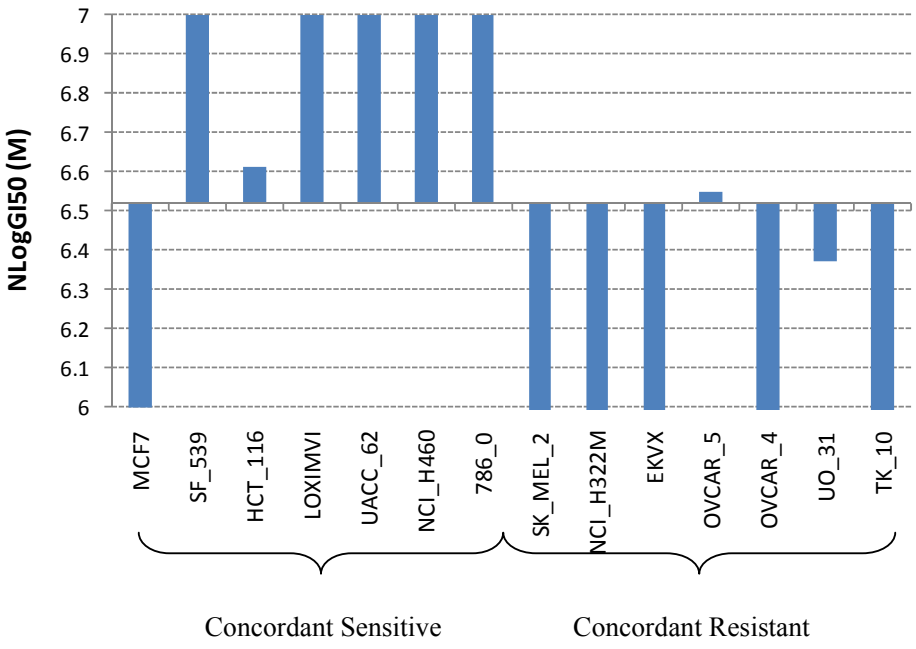



**Figure 4-14:** NLogG50 of Amsacrine for 14 cell lines which show concordant chemotherapeutics in NCI60. 6/7 concordant sensitive cell lines show sensitive to Amsacrine and 6/7 concordant resistant cell lines show resistant to Amsacrine.

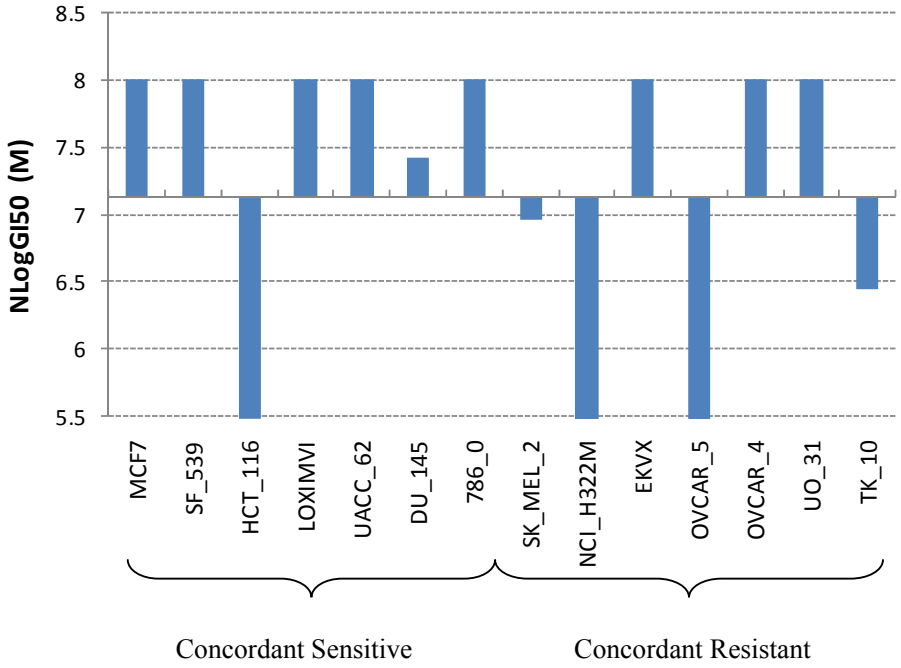



**Figure 4-15**: NLogG50 of Temsirolimus for 14 cell lines which show concordant chemotherapeutics in NCI60. 6/7 concordant sensitive cell lines show sensitive to Temsirolimus and 4/7 concordant resistant cell lines show resistant to Temsirolimus.

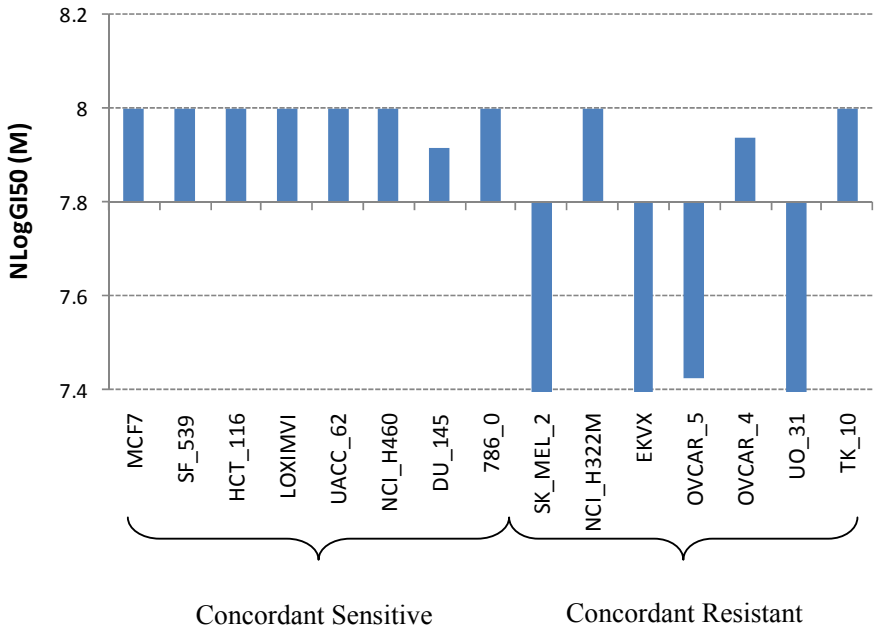



**Figure 4-16**: NLogG50 of Ridaforolimus for 15 cell lines show concordant chemotherapeutics in NCI60. 8/8 concordant sensitive cell lines show sensitive to Ridaforolimus and 4/7 concordant resistant cell lines show resistant to Ridaforolimus.

| Cell Lines | Classes | KRas | BRaf | PIK3CA | PTen | STK11 |
|---|---|---|---|---|---|---|
| MCF7 | Concordant Sensitive | WT | WT | E545K | WT | WT |
| SF_539 | Concordant Sensitive | WT | WT | WT | HD | WT |
| HCT_116 | Concordant Sensitive | G13D | WT | H1047R | WT | WT |
| LOXIMVI | Concordant Sensitive | WT | V600E | WT | WT | WT |
| UACC_62 | Concordant Sensitive | WT | V600E | WT | P248fs*5 | WT |
| NCI_H460 | Concordant Sensitive | Q61H | WT | E545K | WT | p.Q37* |
| DU_145 | Concordant Sensitive | WT | WT | WT | WT | p.K178fs*86 |
| 786_0 | Concordant Sensitive | WT | WT | WT | Q149* | WT |
| SK_MEL_2 | Concordant Resistant | WT | WT | WT | WT | WT |
| NCI_H322M | Concordant Resistant | WT | WT | WT | WT | WT |
| EKVX | Concordant Resistant | WT | WT | WT | WT | WT |
| OVCAR_5 | Concordant Resistant | G12V | WT | WT | WT | WT |
| OVCAR_4 | Concordant Resistant | WT | WT | WT | WT | WT |
| UO_31 | Concordant Resistant | WT | WT | WT | WT | WT |
| TK_10 | Concordant Resistant | WT | WT | WT | WT | WT |

WT: wild type; HD: homozygous deletion; fs: frame shift; *: deletion or substitution.

**Table 4-5**: Listed are the mutations of oncogene and tumor suppressor genes in mTOR upstream pathways. The mutation data is from COSMIC database.

The topoisomerase-II specific inhibitor Amsacrine belongs to the cytotoxic class of topoisomerase inhibitors that include Etoposide, Irinotecan and Topotecan, that are included in the 14 agents used in signature development. The strong association between signature genes and Amsacrine chemo-sensitivity further confirms the predicted ability of the signature of concordant chemotherapeutics as a chemo-response predictor to standard of care cytotoxic agents.

The signature genes are also observed to be significantly correlated with the chemo-sensitivity of the kinase inhibitor Temsirolimus (CCI-779) in the Oncomine database. Temsirolimus specifically inhibits mTOR and results in growth arrest of cells in the G1 phase of the cell cycle. mTOR is a serine/threonine protein kinase that integrates the signals of multiple upstream signaling pathways, e.g. IGF, EGF and mitogens[207, 208]. The mTOR activity plays a central role in the control of cell proliferation, survival, mobility and angiogenesis in multiple solid tumors. To follow up on this observation for Temsirolimus, we examined the drug sensitivity of another mTOR inhibitor, Ridaforolimus, for which NCI60 data is available. Ridaforolimus behaves similarly to Temsirolimus in those NCI60 cell lines (Figure 4-16) with high chemotherapeutic concordance. The presence of the signature of concordant chemotherapeutics in both Temsirolimus and Ridaforolimus sensitive cell lines may indicate that the integrated signal of upstream pathways of mTOR, like Ras/MAPK, AMPK and PI3K/Akt is high in concordant sensitive cell lines. As a preliminary evaluation of the activity of signaling pathways upstream of mTOR, we tabulated the known mutations of 5 cancer genes in mTOR upstream pathways from COSMIC

databases (Table 4-5). Based on the mutation information of KRas, BRaf, PI3KCA, PTen and STK11, it is evident that each of 8 concordant sensitive cell lines harbors known mutations in at least one pathway. However, only 1 of 7 concordant resistant cell lines, OVCAR_5, contains 1 of these mutations (KRas G12V). Although the activity of mTOR pathway in these cell lines awaits further experimental validation, the upstream pathways' information could provide an indicator that mTOR integrates multiple signals to regulate cell proliferations. As a consequence of one or more than one active upstream pathway(s), mTOR is active, and cells may proliferate very fast. We then analyzed the cell growth rate of NCI60 cell lines. Figure 4-17 depicts the doubling time of identified 15 cell lines with high concordance of chemotherapeutics: the doubling time of 8 concordant sensitive cell lines is 23.9±6.9 hours and the doubling time of 7 concordant resistant cell lines is 43.6±5.2 hours. Apparently, the concordant sensitive cell lines grow significantly faster than the concordant resistant cell lines. Some other cell lines in NCI60, like SW620 and COLO205 harbor KRas or BRaf mutation, and both grow fast in culture medium, but they show very divergent chemo-response to multiple cytotoxic agents. The active proliferation pathway is maybe one of the critical prerequisites for cells to show sensitive to multiple agents.

Hence, we hypothesize that solid tumor cells that are sensitive to multiple cytotoxic anticancer agents are growing fast with active proliferation pathway(s). We estimated the tumor doubling time in the mice of 15 Oncotest tumor models which are identified to show concordant chemotherapeutics. As shown in Figure 4-18, 6/8 concordant resistant explants with tumor doubling time in mice is equal to or more

than 8 days, which is much longer than the maximum tumor doubling time in mice of 7 concordant sensitive explants, at just 5 days.

As 93 signature genes under-expressed in concordant sensitive samples are shown to be significantly associated with multiple drug sensitivity signatures, we then performed the enrichment analysis of these genes to assess the overall functional implications of the signature of concordant chemotherapeutics and to generate additional evidence decoding the biological mechanisms of concordant chemotherapeutics. The enrichment analysis was done using MetaCore$^{TM}$ by GeneGo to identify biological functions significantly associated with concordant chemotherapeutics. The histogram (Figure 4-19) shows the top 10 enriched GeneGo pathway maps for the signature genes under-expressed in concordant sensitive cell lines. The protein signaling pathways, such as notch, WNT, late endocytic and SCF are related to the sensitivity of chemotherapeutics. This gives us a clue that the concordant resistance in chemotherapeutics is very complicated, and it is a combination of multiple biological processes instead of being exclusively driven by a single biological pathway.

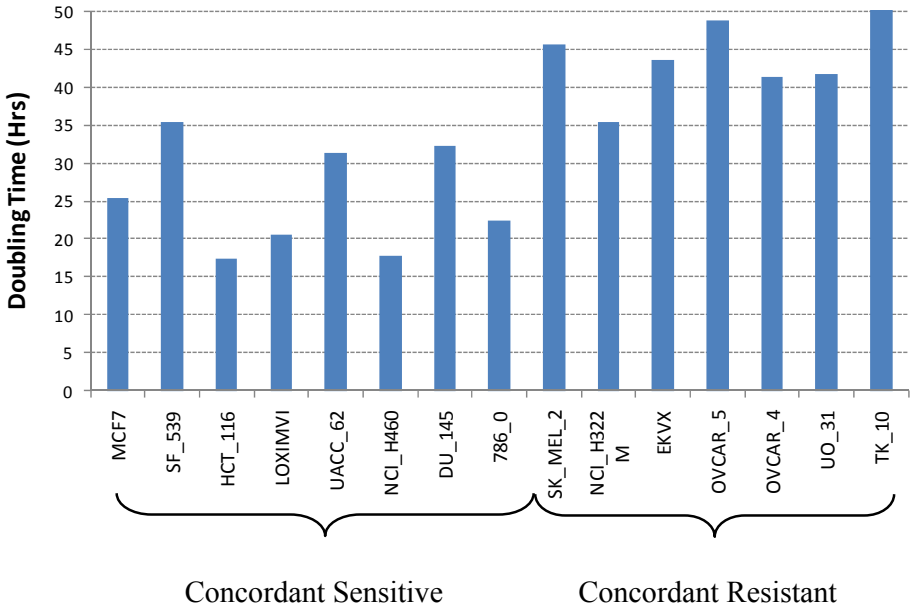



**Figure 4-17**: The doubling time of 15 NCI60 cell lines identified as high concordance of chemotherapeutics. The concordant sensitive cell lines have much shorter doubling time than concordant resistant cell lines.

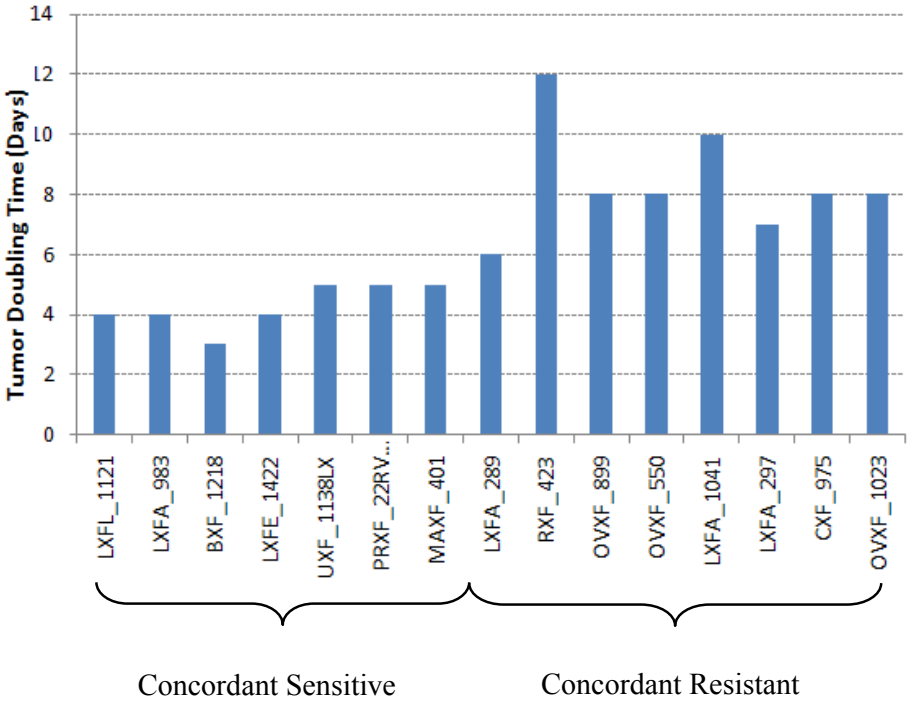



**Figure 4-18**: The tumor doubling time of 15 Oncotest explants models identified as high concordance of chemotherapeutics.

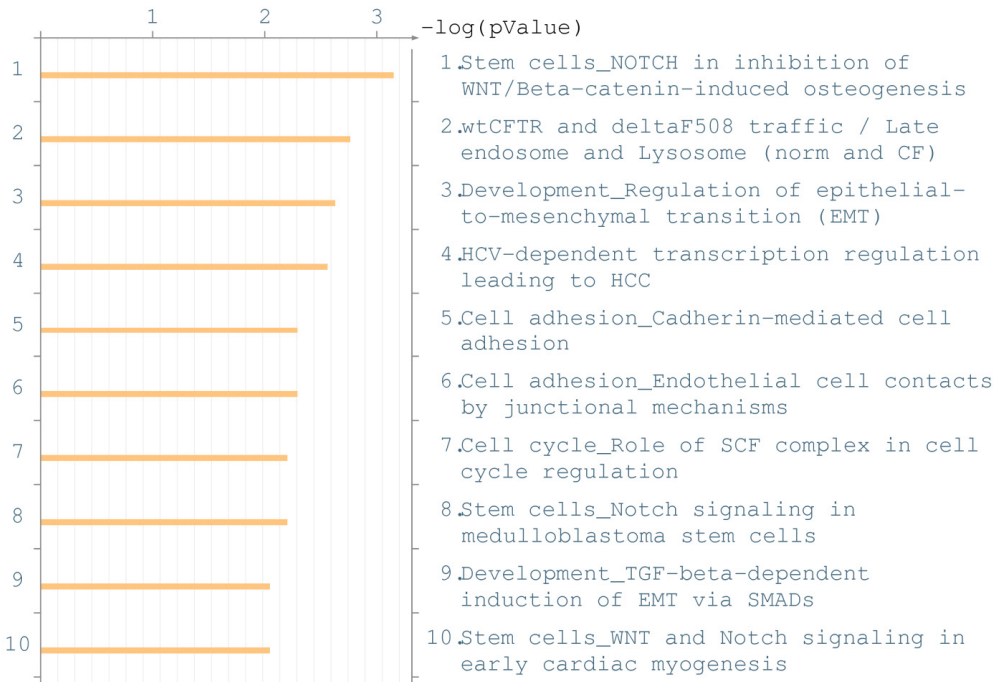



**Figure 4-19**: The top 10 most significantly enriched GeneGo pathway maps for the under expressed (in concordant sensitive cell lines) signature genes. The bars represent significance as –log(p-value) for hypergeometric distribution. All ontology enrichments were filtered at significance level 0.05 (pValue: the significance of the enrich biological process)

## 4.3 Discussions

While advances in understanding of cancer genetics, increased use of targeted agents, and genetic predictors of response to individual agents have all helped build confidence in the promise of personalized medicine, the practical implications are still limited: cancer patients are often treated with combinations of multiple agents including cytotoxic chemotherapies, whose success in combination cannot be predicted in advance for individual patients. A variety of translational strategies using preclinical drug sensitivity data and integrative genomics have been proposed to address this challenge, including an ever widening range of preclinical models across

or within tumor types [209-213]. In addition, systems-based approaches have also begun to be applied to combination treatments explicitly, through pathway models and other systems approaches still in early stages of development [119, 122, 214-216].

We utilize data from two very different sets of preclinical models – immortalized cell lines assayed for cell growth, and patient derived human tumor xenograft samples grown in 3D culture and assayed for clonogenic potential – to develop independently validated and translationally relevant predictive models of drug sensitivity. When tested across the diverse background provided by wide ranging models of solid tumors, the data can be used to identify subsets of cell lines and patient derived xenografts that show concordant sensitivity or resistance to a range of widely used chemotherapeutic agents.

This concordant chemo-response defined in cell lines has a genomic basis, as indicated by the independent predictive power of a signature of concordant response in entirely different experimental model settings. The signature is further validated by prediction of both response and clinical outcome in clinical genomic data sets from patients with different diseases and different treatments. Importantly, genes from the signature of concordant response, when used in a meta-analysis of public gene set categories or "concepts", specifically identify multiple gene sets of sensitivity for chemotherapeutic agents not in the original training set, consistent with a signature for broad chemotherapeutic concordance.

The signature is associated with, but not fully explained by, several common themes in cancer biology. While related to different proliferation rates in concordant sensitive versus concordant resistant models, the signature is a simple consequence of differential proliferation, as non-concordant models provide many counter-examples along with their divergent sensitivity profiles. The identification of the mTOR inhibitor Temsirolimus among agents picked up by the meta-analysis of expression signatures is consistent with known genetic activation upstream of mTOR in concordant resistant samples, and further supported by analysis of NCI60 sensitivity data for the mTOR inhibitor Ridaforolimus, but we do not have experimental mTOR pathway data for these cell lines.

Since most cytotoxic agents interfere with cell cycle processes, it is no surprise that these agents are most effective on rapidly growing tumors. The tumor proliferative activity and response to first-line chemotherapy (anthracycline, clophosphamide, methotrexate and 5-fluorouracil) in advanced breast carcinoma has been studied elsewhere [217]. The results show a high response rate (48%) in the group with highly proliferative tumors when compared with a response rate of 21% in the group with less proliferative tumors. However, regardless of the difference in response rate, survival analysis showed no significant difference between the two groups based on proliferation rate alone.

One response to our finding of concordant chemo-response might be for researchers designing future studies to remove concordant sensitive and concordant resistant cell lines from any training or test set for sensitivity to any agent, on the

grounds that including them would reduce the power of a predictive model. This approach may be warranted within panels of cell lines or primary models used to develop genomic predictors of non-targeted agents. But for either individual cell line study or for developing predictive models for targeted agents, we believe this is a bridge too far: models from the NCI60 panel showing concordant chemo-response have played important roles in cancer biology, and for targeted agents the details of these well-characterized models can make them particularly appropriate for target hypothesis-based experimentation.

If the signature of concordance is validated in further patient studies, there may be clinical value in identifying patients whose tumors show a signature of concordant resistance to chemotherapeutics, to ascertain whether these patients should be prioritized for earlier combinations with targeted agents, rather than front-line treatment with chemotherapeutic combinations.

We conclude that the developed gene expression signature could characterize the concordance of chemotherapeutics of standard of care agents, and it potentially may be applied as the predictor to tailor the patient's response, given by standard of care chemotherapeutics with or without combination in a range of solid tumors.

## 4.4 Methods and Materials

### 4.4.1 Anticancer Cytotoxic agents

The selected 14 Standard of Care (SOC) anticancer cytotoxic agents include 5 types of inhibitors: alkylating/alkylating-like, antimetabolites, antitumor antibiotic, spindle poison/mitotic inhibitor and topoisomerase inhibitor (Table 4-1).

### 4.4.2 In vitro tumor explants screening at Oncotest

The drug sensitivity of 14 anticancer cytotoxic agents were studied by Tumor Clonogenic Assay (TCA) in Oncotest. TCA studied the inhibition of anchorage independent growth and in vitro colony formation of tumor cells derived from human tumor xenografts of various tumor histologies in semi-solid medium. The assay is performed in a 24-well format with six replicates for untreated controls and three replicates for treatment wells. Agent effects are expressed in terms of the percentage of colony formation, obtained by comparison of the mean number of colonies in the treated wells with the mean colony count of the untreated controls (relative colony count expressed by the mean of treatment versus mean of control) [7]. IC50 is the drug concentrations necessary to inhibit colony formation by 50% (Treatment /Control = 50%). Four parameter dose response curves to determine these concentrations are fit. If an IC50 value could not be determined within the examined dose range because an agent was either too active or lacked activity, the lowest or highest concentration studied was used for the calculation.

### 4.4.3 In vitro cell line screening at NCI-DTP

The National Cancer Institute (NCI) Developmental Therapeutics Program (DTP) program, cell lines and assay related details are given at http://dtnci.nih.gov/. The 14 cytotoxic agents' chemo-response profile in NCI was reported in the DTP database. The database provides the estimates of concentration for IC50 (the concentration that cause 50% cells inhibition 100 × T/C = 50), GI50 (the concentration of test drug where 100 × (T - T0)/(C - T0) = 50, T0 is the optical density (response) of the "test well" at time zero), TGI (the concentration of test drug where 100 × (T - T0)/(C - T0) = 0, it signifies cytostatic effect) and LC50 (the concentration of drug where 100 × (T - T0)/T0 = -50), which are defined at http://dtnci.nih.gov/branches/btb/ivclshtml.

### 4.4.4 Microarray data

The tumor sample with TCA data were prepared by Oncotest and the microarray data was generated at Microarray Facility, Medical Genetics, Tuebingen, Germany (http://www.microarray-facility.com) as per the guidelines for Affymetrix gene expression microarrays (Affymetrix Inc., Santa Clara, USA). The expression profiling used RNA is from explant materials that were derived primary tumors. Following the sacrifice of mice by cervical dislocation, tumors of 400-600 mm^3 volume were excised without delay, and tumor pieces free of necrosis were flash frozen in liquid nitrogen. For gene expression profiling of human tumor xenografts, total RNA was purified using the RNeasy Mini kit (QIAGEN, Hilden, Germany).

During RNA isolation, no genomic DNA digestion was done. Prior to hybridizing to microarrays, 1.2 µg of total RNA was amplified using the One-Cycle Eukaryotic Target Labeling Assay (Affymetrix Inc., Santa Clara, USA). 15 µg of labeled complementary RNA (cRNA) was then hybridized to Affymetrix HG-U133 Plus 2.0 GeneChip expression array. The CEL files were processed using a statistical package R-project environment version 2.12.0 (http://www.r-project.com) with Bioconductor package version 2.7. The signal for probesets were condensed using MAS5.0 algorithm, normalized to 500 fluorescence units, followed by log base 2 transformation. The publicly available microarray datasets were downloaded from NCBI. Breast cancer (Hess et al. 2006) dataset (Affymetrix U133A gene expression array) was downloaded from http://www.bioinformatics.mdanderson.org /pubdata.html. NCI60 cell line expression dataset (Affymetrix U133A and U133B gene expression array) is available as accession number GSE5720 at NCBI's Gene expression omnibus. Bild lung [141] cancer and Miller breast [204] datasets (Affymetrix U133 plus2 gene expression array) are available as accession number GSE3141 and GSE3494 respectively at NCBI's Gene expression omnibus.

### 4.4.5 Statistical analysis method

Gene expression data is filtered to exclude probesets that show background level expression and that do not vary significantly across samples. Furthermore, the probesets are selected to be significantly differentiated between sensitive and resistant

samples using FDR (false discovery rate) criteria. The analysis uses R package "limma" (Linear Models for Micorarray Data[105])) (version 3.6.6).

The statistical method to generate the probability of chemotherapeutic sensitivity uses standard binary regression models combined with singular value decompositions (SVD) [79]. Here for simplicity, we name it as "Bayesian-SVD" method. First, the probeset selection from the training data is done as stated above. Training data is constructed with samples in columns and probesets (or genes) in rows. Principal components of the training data are used to compute the metagene and metasample values. Bayesian binary probit regression model is fitted to the metagene signature for assessing the relevance of each metagene and the classes of chemo-response of training samples. We assess the performance of the signature in an independent test data. Given the test data set, the gene expression data with probesets matched with signature is then projected onto the metagenes from the training data. The model was fitted and relative probability of sensitivity (POS) is predicted using the Bayesian binary probit regression parameters fitted from the metagene signature. The classes are defined as '0' for resistant and '1' for sensitive for training such that low POS scores would be suggestive of a sample being resistant and vice versa. The implementation based on methodology is done in R environment (version 2.5 and above).

We use a Receiver Operating Characteristic (ROC) curve to plot the true positive rate (sensitivity) and false positive rate (1-specificity) for different cutoff levels of the POS score estimated above. The Area Under Curve (AUC) is a measure

of the accuracy of the test. ROC curve and AUC analysis are done using R packages PresenceAbsence version 1.1. Empirical Cumulative Distribution Function (ECDF) is a cumulative distribution function that assigns probability '1/N' to each of the 'N' cases in a sample. A cumulative distribution function (CDF) gives the probability when a random variable X is less that a given value x. ($CDF(x) = Pr\{X \leq x\}$). An ECDF is a sample based estimate of this theoretical function. The ECDF plots are generated using R environment's base package. The Kaplan-Meier survival analysis is performed using the Cox-Mantel log-rank test in R "survival" package version 2.36-1. The log-rank test p-value represents the significance of the difference of the probably of survival between different patient groups.

# Chapter 5 Conclusions and discussions

As the dissertation comes to end, it is important to highlight the important findings and address a number of remaining issues. Firstly, the systematic bioinformatics methodology to develop principle genomic markers introduced in chapter 2 and chapter 3 is reviewed. In particular, miRNA expression markers, that has been largely ignored, has been highlighted for discussion. Building on this discussion, the miRNA expression markers are generated for the concordant chemotherapeutics. These interesting results may suggest some concrete direction for future research to extend the genomic marker of concordant chemotherapeutics presented in chapter 4. The conclusion of the dissertation will focus on the contributions of this work to the cancer research society.

## 5.1 Discussion of integrated genomic markers development

The starting point to develop genomic markers for chemotherapeutics is from the fuzzy classification of chemo-response data. We use fuzzy data instead of actual numeric readout of GI50 and IC50 data because of its inherent variability of the readouts from the in vitro assay screening. Classification of “sensitive”, “medium” and “resistant” can also relate to the pathological response of the treatment of the anticancer agents in clinical patients, which is usually classified as “complete response / partial response”, “stable disease” and “progressive disease”. Since the chemo-response data is classified into fuzzy classes, correspondingly, the phenotypic

expression, like gene expression and protein expression data can be considered as “high”, “medium” or “low”. Therefore, we proposed two types of genomic markers: type-I and type-II markers. Type-1 markers are significantly correlated with the rank of chemo-response classes. Type-II markers are the combination of significantly differentiated genes sets between pair-wise chemo-response classes. To move the developed the multiple-gene signature biomarker assay into clinical practice is very challenging due to poor choice of assay genes and lack substantial preclinical validations[218, 219]. The integrated genomic marker analysis is the method to select the reliable assay markers and move into further validation, by which, the genomic aberrations, such as gene mutation, DNA copy number variations and methylation, are incorporated into the analysis to explain high or low expression patterns of the specific candidate marker. The marker with its expression significantly affected by the genomic aberration(s) may be considered as potential clinical candidate biomarker, and is to be followed up with further assay validations.

### 5.1.1 MicroRNAs correlated with chemo-response

However, there are certain limitations in selecting potential clinical candidate biomarker from initial multi-gene expression signature using this integrated genomic analysis. The gene or protein expression can be modulated from DNA-RNA transcription to the post-translational modification for a protein. The genomic aberrations considered in this study are only the modification of DNAs. The gene can also be post-transcriptionally regulated by microRNAs, that binds to complementary

sequences on target messenger RNA transcripts, and usually resulting in translational repression and silence of the gene[50, 51]. It is important to include the microRNA information in the study to interpret the expression pattern of candidate markers. For example, as shown in Figure 5-1, the expression of miRNA-30 family is identified to be significantly correlated with the concordant sensitive and concordant resistant cell lines categorized in NCI60 data in chapter 4. Although it has been revealed that miRNA-30s may target TP53 protein[220], regulate B-Myc activity together with miRNA-29[221] and play very critical roles in causing familial breast cancers[222], the targets of miRNA-30 family are still ambiguous. It is difficult to interpret the biological mechanisms of miRNA-30s when integrating with prior developed gene expression biomarkers from the developed gene signature of concordant chemotherapeutics. Alternatively, it is believed miRNA-30s should be further explored, especially their roles in multiple proteins signaling pathways.

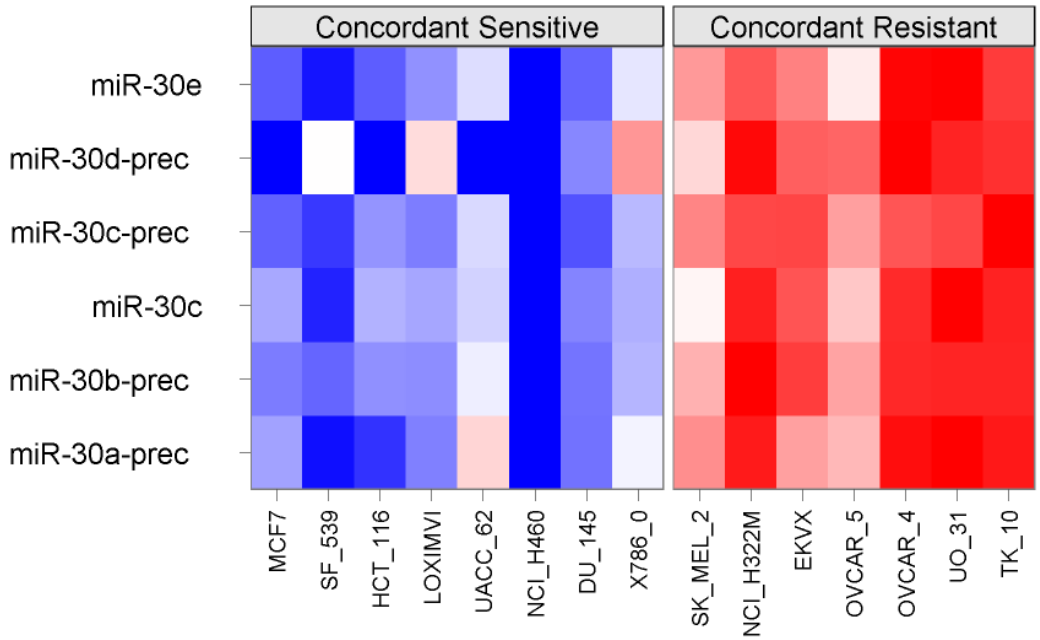



**Figure 5-1**: The expressions of miRNA-30 family are significantly correlated with concordant chemotherapeutics samples

## 5.1.2 Trans-modulation

The karyotype of a chromosomal segment affects its copy number variation. The expression of a gene is usually affected by the copy number variations, mutation and DNA methylation. The integrated genomic analysis model is proposed in chapter 3 with the focus on the gene expression and its own mutation, copy number variation and DNA methylation. However, the genetic aberrations can also modulate the expressions on the other locus. Therefore, the integrated genomic model should associate non-local genetic aberrations of chromosomal segment copy number variations with the maker's expression. Yeang presented a multiple-layer modeling framework to identify the statistical and putative causal relations of the gene expression and non-local genetic aberrations of chromosomal segment copy number

variations, mutation and methylation[223] based on known mechanistic links. Based on this reference, we can modify the integrated model proposed in chapter 3 as follows:

$$y_i = a_0 + a_1 \cdot Mu_i + a_{21} \cdot CNV_i + a_{22} \cdot CNV_j + a_3 \cdot Me_i,\ a_0, a_1, a_{21}, a_{22},\ a_3 > \quad \text{Eq 5-1}$$

Here, $CNV_j$ is the mechanistic link which is reported to trans-modulate the expression of gene *i*. If the number of known trans-modulation copy number variation is more than 1, *j* may take a series of numbers.

As we discussed in the chapter 1, the potential model, Eq 1-2, has incorporated eQTL SNP variation information, which may also consider the unknown non-local genetic aberrations. Since the objective of the integrated genomic analysis is to identify more robust expression makers and move them into clinical validation, the proposed model may be more meaningful. Thus the accuracy of our analysis may be low if unknown non-local molecular aberrations are shown to be significantly associated with the expression of the marker gene. The proposed integrated model can be used to explore the possible biological links between the gene expression and non-local genetic aberrations.

## 5.2 Proteomic expression markers

In this dissertation, we have focused on the gene expression markers. As we introduced in chapter 2, protein expression signature is another type of important principle expression markers. While gene expression can now be measured reliably and reproducibly in high throughput[202, 203, 224-227], protein expression

measurement technologies are still limited at relatively low and medium throughput. However, technologies like reverse phase protein arrays (RPPA) and reverse phase lysate arrays (RPLA) seems to provide a reliable high throughput platform to utilize the proteomic expression data for exploring the potential biomarkers for chemotherapeutics[228].

Therefore, we developed the protein expression signature to characterize the concordant chemotherapeutics. As shown in Figure 5-2, the 30 proteins (antibodies) expression signature includes 7 over expressed and 23 under expressed (in concordant sensitive cell lines) proteins. Although the expression pattern is not very concrete compared to the gene expression signature, this does give us another scope to decode the concordant chemotherapeutics. Interestingly, both tumor suppressor genes, TP53 and RB1 are under-expressed in concordant sensitive cell lines, and these cell lines have high proliferation rate than resistant cell lines. Although no critical conclusion is drawn from the protein expression signature, it does expand our understanding of the activation status of multiple protein networks, and provides a basis for further looking into the integration with genomic analysis results such as gene expression signature and miRNA markers.

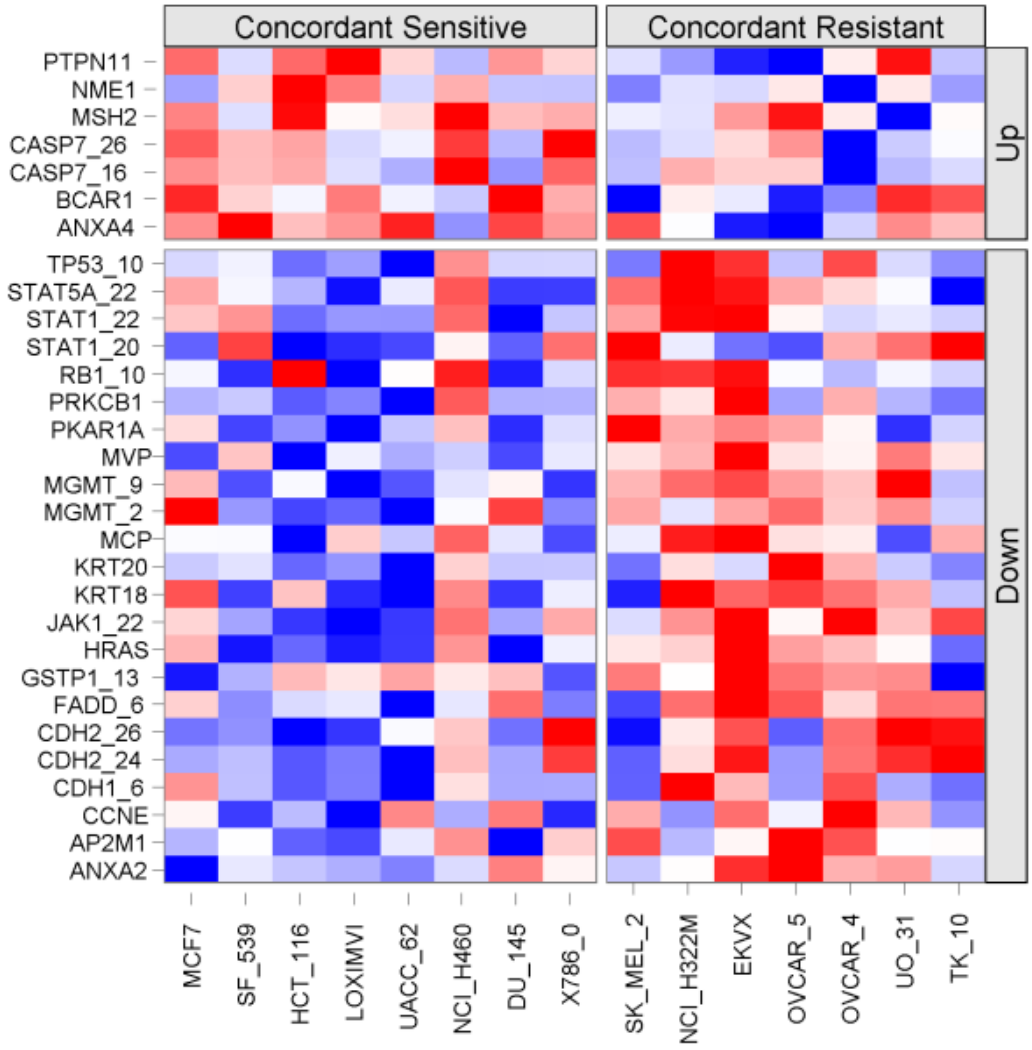



**Figure 5-2**: The protein expression signature of concordant chemotherapeutics (red: high value; blue: low value). Signature is derived with LogFC=0.6 and FDR=0.2.

## 5.2 Extension for future work

One of the most important bioinformatics analyses is to offer one-step validations for the experimental scientist. It is my special interest to suggest a few biological experiments to further discover novel biology. There are a few important and putative hypothesis considered that can be studied experimentally.

The mTOR upstream signaling pathways in concordant sensitive cell lines are more active than in concordant resistant cell lines. The phosph-mTRC1 and phosph-mTRC2 activity may be measured in the NCI60 concordant sensitive cell lines and concordant resistant cell lines.

Another interesting follow up is that the miRNA-30s family show significantly differentiation patterns in concordant sensitive and concordant resistant samples. The general understanding that miRNA-30s may play important role in protein signaling pathways[229-231] is poor: there is very few published research papers reported to date. Our results indicate that it may play a critical role in the pan-resistance of chemotherapeutics.

## 5.3 Contributions

In this dissertation, we have made a number of significant contributions to the society of biomarker research in two main areas: a) methodologies development: we introduced fuzzy classification of biological data, two types of principle expression biomarker and the development algorithms, and then an integrated genomic analysis method between gene expression and genomic aberrations is proposed to narrow down the multiple-gene signature to few(<10) gene markers or find a single gene marker; and b) novel biology： we identified an important phenomena of chemotherapeutics - cancer cells may show concordant chemotherapeutics to multiple anticancer agents. To understand the underlying biology of concordant chemotherapeutics in cancer cells, we first categorized concordant sensitive and

concordant resistant cell lines, and primary derived tumor grafts. We then developed the gene expression signature of concordant chemotherapeutics using NCI60 data, with robust in-vitro validation in Oncotest tumor graft data. Thereafter, we employed the signature to predict the TFAC treated breast cancer patients' samples and to stratify patients groups in both breast and lung cancer to evaluate the prognostic value of the signature. From the following meta-analysis in Oncomine database, we found that the signature genes are highly enriched in multiple signature concepts of anticancer agents. Furthermore, the concordant sensitive cell lines and tumor grafts are found to have higher proliferation rate than concordant resistant cell lines and tumor grafts. The mutational analysis may suggest that multiple protein signaling pathways may be hyperactive. Although this putative hypothesis needs to be tested experimentally, the developed gene signature of concordant chemotherapeutics is valuable, and should be moved into next step, which is validation.

In order to demonstrate the end-to-end application of the biomarker development framework, we introduced a number of meaningful and signature concepts with suitable classification concepts for this particular domain. In particular, we proposed to use an integrated genomic analysis model to identify clinical like biomarkers. In the contribution of novel biology, we demonstrated that the identified biological phenomena "concordant chemotherapeutics" is present in both preclinical models and clinical patients.

Finally, we hope that this dissertation helps biomarker scientists better understand the biomarker development process, and offers cancer research scientists a

deeper understanding of the resistance of chemotherapeutics to multiple cytotoxic agents.

# Appendix

**Applendix-1:** The gene signature genes and probesets (Affymetrix U133 A&B) of the concordant chemotherapeutics

| Probesets | Gene Symbol | Expression in Sensitive Cell Lines | logFC |
|---|---|---|---|
| 242541_at | ABCA9 | Up | 1.25 |
| 231299_at | AGAP3 | Up | 1.96 |
| 229709_at | ATP1B3 | Up | 1.66 |
| 229018_at | C12orf26 | Up | 2.03 |
| 219260_s_at | C17orf81 | Up | 1.07 |
| 239208_s_at | C21orf57 | Up | 1.77 |
| 235219_at | C5orf55 | Up | 1.5 |
| 235981_at | C8orf22 | Up | 1.55 |
| 57715_at | CALHM2 | Up | 1.69 |
| 211347_at | CDC14B | Up | 1.77 |
| 206474_at | CDK17 | Up | 1.52 |
| 210689_at | CLDN14 | Up | 1.74 |
| 226751_at | CNRIP1 | Up | 3.23 |
| 220323_at | CNTD2 | Up | 1.3 |
| 205653_at | CTSG | Up | 1.2 |
| 241381_at | CXorf36 | Up | 1.32 |
| 218808_at | DALRD3 | Up | 1.36 |
| 219328_at | DDX31 | Up | 1.65 |
| 207379_at | EDIL3 | Up | 1.85 |
| 240528_s_at | EXOC4 | Up | 1.59 |
| 225099_at | FBXO45 | Up | 1.03 |
| 241671_x_at | FLJ22536 | Up | 1.61 |
| 219170_at | FSD1 | Up | 2.09 |
| 235574_at | GBP4 | Up | 1.52 |
| 220265_at | GPR107 | Up | 1.17 |
| 220042_x_at | HIVEP3 | Up | 1.2 |
| 227361_at | HS3ST3B1 | Up | 2.65 |
| 235301_at | KIAA1324L | Up | 1.82 |
| 228476_at | KIAA1407 | Up | 1.48 |
| 230432_at | LOC100422737 | Up | 1.3 |
| 235494_at | LSAMP | Up | 1.45 |
| 242838_at | MAP6D1 | Up | 1.27 |
| 207121_s_at | MAPK6 | Up | 1.05 |

| | | | |
|---|---|---|---|
| 214270_s_at | MAPRE3 | Up | 2.36 |
| 208595_s_at | MBD1 | Up | 1.39 |
| 201151_s_at | MBNL1 | Up | 1.03 |
| 222867_s_at | MED31 | Up | 1.15 |
| 225316_at | MFSD2A | Up | 1.37 |
| 228592_at | MS4A1 | Up | 1.43 |
| 222570_at | NCS1 | Up | 1.25 |
| 226585_at | NEIL2 | Up | 1.94 |
| 206929_s_at | NFIC | Up | 1.71 |
| 222057_at | NOL12 | Up | 1.76 |
| 209629_s_at | NXT2 | Up | 1.91 |
| 219295_s_at | PCOLCE2 | Up | 1.67 |
| 236135_at | PNPLA7 | Up | 1.73 |
| 207000_s_at | PPP3CC | Up | 1.29 |
| 209599_s_at | PRUNE | Up | 1.44 |
| 205961_s_at | PSIP1 | Up | 1.06 |
| 243777_at | RAB7L1 | Up | 1.26 |
| 226945_at | RHBDD1 | Up | 1.24 |
| 218861_at | RNF25 | Up | 1.57 |
| 210426_x_at | RORA | Up | 1.44 |
| 222559_s_at | RPRD1A | Up | 1.28 |
| 236782_at | SAMD3 | Up | 1.33 |
| 215834_x_at | SCARB1 | Up | 1.66 |
| 211708_s_at | SCD | Up | 2.25 |
| 242064_at | SDK2 | Up | 1.23 |
| 244653_at | SETD7 | Up | 1.61 |
| 210135_s_at | SHOX2 | Up | 2.18 |
| 219713_at | SHPK | Up | 1.33 |
| 210567_s_at | SKP2 | Up | 1.11 |
| 218978_s_at | SLC25A37 | Up | 1.22 |
| 228935_at | SLC4A8 | Up | 1.55 |
| 213854_at | SYNGR1 | Up | 1.46 |
| 210053_at | TAF5 | Up | 1.32 |
| 222053_at | TAF6L | Up | 1.12 |
| 230394_at | TCP10L | Up | 1.39 |
| 223523_at | TMEM108 | Up | 2.21 |
| 223462_at | TMEM175 | Up | 1.28 |
| 213725_x_at | XYLT1 | Up | 1.74 |
| 203604_at | ZNF516 | Up | 1.39 |
| 226677_at | ZNF521 | Up | 2.08 |

| | | | |
|---|---|---|---|
| 232137_at | ZNF616 | Up | 1.44 |
| 215767_at | ZNF804A | Up | 2.05 |
| 219488_at | A4GALT | Down | -1.57 |
| 228132_at | ABLIM2 | Down | -1.54 |
| 211712_s_at | ANXA9 | Down | -1.87 |
| 242727_at | ARL5B | Down | -1.04 |
| 212312_at | BCL2L1 | Down | -1.37 |
| 229975_at | BMPR1B | Down | -3.39 |
| 224435_at | C10orf58 | Down | -2.01 |
| 228155_at | C10orf58 | Down | -2.27 |
| 239777_at | C14orf182 | Down | -2.06 |
| 219010_at | C1orf106 | Down | -1.51 |
| 223951_at | C21orf116 | Down | -1.48 |
| 213199_at | C2CD3 | Down | -1.17 |
| 224707_at | C5orf32 | Down | -1.26 |
| 209495_at | CEP250 | Down | -1.18 |
| 204233_s_at | CHKA | Down | -1.63 |
| 204266_s_at | CHKA | Down | -2.15 |
| 222549_at | CLDN1 | Down | -1.89 |
| 224815_at | COMMD7 | Down | -1.01 |
| 201906_s_at | CTDSPL | Down | -1.02 |
| 201905_s_at | CTDSPL | Down | -1.46 |
| 201904_s_at | CTDSPL | Down | -2.81 |
| 202295_s_at | CTSH | Down | -1.37 |
| 214782_at | CTTN | Down | -1.01 |
| 238280_at | CYB5RL | Down | -1.5 |
| 230679_at | DCAF10 | Down | -1.96 |
| 202500_at | DNAJB2 | Down | -1.77 |
| 204720_s_at | DNAJC6 | Down | -1.58 |
| 204947_at | E2F1 | Down | -1.24 |
| 228256_s_at | EPB41L4A | Down | -2.07 |
| 224024_at | ERGIC1 | Down | -1.73 |
| 208297_s_at | EVI5 | Down | -1.21 |
| 223058_at | FAM107B | Down | -1.57 |
| 223059_s_at | FAM107B | Down | -1.6 |
| 223745_at | FBXO31 | Down | -1.62 |
| 217342_x_at | FLJ11292 | Down | -1.22 |
| 202838_at | FUCA1 | Down | -1.73 |
| 208505_s_at | FUT2 | Down | -1.31 |
| 206780_at | GAD2 | Down | -1.64 |

| | | | |
|---|---|---|---|
| 213343_s_at | GDPD5 | Down | -1.91 |
| 242281_at | GLUL | Down | -1.38 |
| 203108_at | GPRC5A | Down | -3.17 |
| 227614_at | HKDC1 | Down | -2.83 |
| 201655_s_at | HSPG2 | Down | -2.31 |
| 216268_s_at | JAG1 | Down | -1.66 |
| 31849_at | KRT80 | Down | -2.68 |
| 12531_at | LCN2 | Down | -2.15 |
| 221115_s_at | LENEP | Down | -1.96 |
| 219181_at | LIPG | Down | -1.38 |
| 230388_s_at | LOC644246 | Down | -1.08 |
| 204682_at | LTBP2 | Down | -1.02 |
| 228885_at | MAMDC2 | Down | -1.48 |
| 235106_at | MAML2 | Down | -1.75 |
| 216206_x_at | MAP2K7 | Down | -1.19 |
| 210136_at | MBP | Down | -2.8 |
| 202616_s_at | MECP2 | Down | -1.32 |
| 244741_s_at | MGC9913 | Down | -2.34 |
| 224685_at | MLLT4 | Down | -1.01 |
| 233539_at | NAPEPLD | Down | -1.29 |
| 218414_s_at | NDE1 | Down | -1.46 |
| 235517_at | PACRGL | Down | -1.8 |
| 207717_s_at | PKP2 | Down | -2.4 |
| 219024_at | PLEKHA1 | Down | -1.28 |
| 202065_s_at | PPFIA1 | Down | -1.07 |
| 202066_at | PPFIA1 | Down | -1.33 |
| 49077_at | PPME1 | Down | -1.05 |
| 217841_s_at | PPME1 | Down | -1.39 |
| 238118_s_at | PPOX | Down | -1.72 |
| 226907_at | PPP1R14C | Down | -1.59 |
| 228494_at | PPP1R9A | Down | -1.31 |
| 238441_at | PRKAA2 | Down | -2.66 |
| 221808_at | RAB9A | Down | -1.11 |
| 219026_s_at | RASAL2 | Down | -1.58 |
| 226436_at | RASSF4 | Down | -1.49 |
| 226164_x_at | RIMKLB | Down | -1.09 |
| 221215_s_at | RIPK4 | Down | -1.68 |
| 228044_at | SERP2 | Down | -1.51 |
| 233753_at | SFRS15 | Down | -1.53 |
| 242963_at | SGMS2 | Down | -1.44 |

| | | | |
|---|---|---|---|
| 223698_at | SLC25A36 | Down | -1.37 |
| 242274_x_at | SLC25A42 | Down | -2.02 |
| 244353_s_at | SLC2A12 | Down | -1.33 |
| 228221_at | SLC44A3 | Down | -1.64 |
| 222967_at | SLC5A7 | Down | -2.05 |
| 238691_at | SNHG10 | Down | -1.67 |
| 218705_s_at | SNX24 | Down | -1.27 |
| 213667_at | SRCAP | Down | -2.08 |
| 226932_at | SSPN | Down | -1.99 |
| 226822_at | STOX2 | Down | -1.25 |
| 202565_s_at | SVIL | Down | -1 |
| 235762_at | TAS2R14 | Down | -1.81 |
| 203221_at | TLE1 | Down | -1.09 |
| 203222_s_at | TLE1 | Down | -1.9 |
| 238802_at | TYSND1 | Down | -1.68 |
| 226678_at | UNC13D | Down | -1.74 |
| 204929_s_at | VAMP5 | Down | -2.3 |
| 235023_at | VPS13C | Down | -1.58 |
| 210248_at | WNT7A | Down | -1.88 |
| 213085_s_at | WWC1 | Down | -1.41 |
| 239757_at | ZFAND6 | Down | -1.4 |
| 216710_x_at | ZNF287 | Down | -1.72 |
| 240181_at | ZSCAN12 | Down | -1.47 |

# Bibliography


1. Gottesman MM: **Mechanisms of cancer drug resistance**. *Annu Rev Med* 2002, **53**:615-627.
2. Selby P: **Acquired resistance to cancer chemotherapy**. *Br Med J (Clin Res Ed)* 1984, **288**(6426):1252-1253.
3. Jones PM, George AM: **The ABC transporter structure and mechanism: perspectives on recent research**. *Cell Mol Life Sci* 2004, **61**(6):682-699.
4. Sargeant J, Hurley KF, Duffy J, Sketris I, Sinclair D, Ducharme J: **Lost in translation or just lost?** *Ann Emerg Med* 2008, **52**(5):575-576; author reply 576-577.
5. Straus SE, Tetroe JM, Graham ID: **Knowledge translation is the use of knowledge in health care decision making**. *J Clin Epidemiol* 2011, **64**(1):6-10.
6. Alley MC, Scudiero DA, Monks A, Hursey ML, Czerwinski MJ, Fine DL, Abbott BJ, Mayo JG, Shoemaker RH, Boyd MR: **Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay**. *Cancer Res* 1988, **48**(3):589-601.
7. Fiebig HH, Maier A, Burger AM: **Clonogenic assay with established human tumour xenografts: correlation of in vitro to in vivo activity as a basis for anticancer drug discovery**. *Eur J Cancer* 2004, **40**(6):802-820.
8. Sulston J: **Society and the human genome. Sir Frederick Gowland Hopkins Memorial Lecture**. *Biochem Soc Trans* 2001, **29**(Pt 2):27-31.
9. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y *et al*: **The sequence and de novo assembly of the giant panda genome**. *Nature* 2010, **463**(7279):311-317.
10. Zhao J, Grant SF: **Advances in whole genome sequencing technology**. *Curr Pharm Biotechnol* 2011, **12**(2):293-305.
11. Chan EY: **Advances in sequencing technology**. *Mutat Res* 2005, **573**(1-2):13-40.
12. Dovichi NJ: **Advances in DNA sequencing technology**. *Hum Mutat* 1993, **2**(2):82-84.
13. Hunkapiller MW: **Advances in DNA sequencing technology**. *Curr Opin Genet Dev* 1991, **1**(1):88-92.
14. Mihaly Z, Gyorffy B: **[Next generation sequencing technologies (NGST) -- development and applications]**. *Orv Hetil* 2011, **152**(2):55-62.
15. Cullum R, Alder O, Hoodless PA: **The next generation: using new sequencing technologies to analyse gene regulation**. *Respirology* 2011, **16**(2):210-222.
16. Roukos DH: **Next-generation sequencing and epigenome technologies: potential medical applications**. *Expert Rev Med Devices* 2010, **7**(6):723-726.

17. Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, O'Toole PW: **Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions**. *Nucleic Acids Res* 2010, **38**(22):e200.
18. Metzker ML: **Sequencing technologies - the next generation**. *Nat Rev Genet* 2010, **11**(1):31-46.
19. Hert DG, Fredlake CP, Barron AE: **Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods**. *Electrophoresis* 2008, **29**(23):4618-4626.
20. Shastry BS: **Copy number variation and susceptibility to human disorders (Review)**. *Mol Med Report* 2009, **2**(2):143-147.
21. Zhang F, Gu W, Hurles ME, Lupski JR: **Copy number variation in human health, disease, and evolution**. *Annu Rev Genomics Hum Genet* 2009, **10**:451-481.
22. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W *et al*: **Global variation in copy number in the human genome**. *Nature* 2006, **444**(7118):444-454.
23. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME *et al*: **Copy number variation: new insights in genome diversity**. *Genome Res* 2006, **16**(8):949-961.
24. Hastings PJ, Lupski JR, Rosenberg SM, Ira G: **Mechanisms of change in gene copy number**. *Nat Rev Genet* 2009, **10**(8):551-564.
25. Hendrich BD, Willard HF: **Epigenetic regulation of gene expression: the effect of altered chromatin structure from yeast to mammals**. *Hum Mol Genet* 1995, **4 Spec No**:1765-1777.
26. Jaenisch R, Bird A: **Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals**. *Nat Genet* 2003, **33 Suppl**:245-254.
27. Sulewska A, Niklinska W, Kozlowski M, Minarowski L, Naumnik W, Niklinski J, Dabrowska K, Chyczewski L: **Detection of DNA methylation in eucaryotic cells**. *Folia Histochem Cytobiol* 2007, **45**(4):315-324.
28. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW: **Direct detection of DNA methylation during single-molecule, real-time sequencing**. *Nat Methods* 2010, **7**(6):461-465.
29. Levenson VV: **DNA methylation biomarkers of cancer: moving toward clinical application**. *Pharmacogenomics* 2004, **5**(6):699-707.
30. Das PM, Singal R: **DNA methylation and cancer**. *J Clin Oncol* 2004, **22**(22):4632-4642.
31. Szyf M, Pakneshan P, Rabbani SA: **DNA methylation and breast cancer**. *Biochem Pharmacol* 2004, **68**(6):1187-1197.
32. Hobert O: **Gene regulation by transcription factors and microRNAs**. *Science* 2008, **319**(5871):1785-1786.

33. Chen K, Rajewsky N: **The evolution of gene regulation by transcription factors and microRNAs**. *Nat Rev Genet* 2007, **8**(2):93-103.
34. Nasser SM: **Gene expression profiling in breast cancer**. *J Med Liban* 2009, **57**(2):83-88.
35. Turaga K, Acs G, Laronga C: **Gene expression profiling in breast cancer**. *Cancer Control* 2010, **17**(3):177-182.
36. Bao T, Davidson NE: **Gene expression profiling of breast cancer**. *Adv Surg* 2008, **42**:249-260.
37. Cheang MC, van de Rijn M, Nielsen TO: **Gene expression profiling of breast cancer**. *Annu Rev Pathol* 2008, **3**:67-97.
38. Pusztai L: **Gene expression profiling of breast cancer**. *Breast Cancer Res* 2009, **11 Suppl 3**:S11.
39. Morris SR, Carey LA: **Gene expression profiling in breast cancer**. *Curr Opin Oncol* 2007, **19**(6):547-551.
40. Petersen S, Heckert C, Rudolf J, Schluns K, Tchernitsa OI, Schafer R, Dietel M, Petersen I: **Gene expression profiling of advanced lung cancer**. *Int J Cancer* 2000, **86**(4):512-517.
41. Singhal S, Miller D, Ramalingam S, Sun SY: **Gene expression profiling of non-small cell lung cancer**. *Lung Cancer* 2008, **60**(3):313-324.
42. Lacroix L, Commo F, Soria JC: **Gene expression profiling of non-small-cell lung cancer**. *Expert Rev Mol Diagn* 2008, **8**(2):167-178.
43. Pollack JR: **DNA microarray technology. Introduction**. *Methods Mol Biol* 2009, **556**:1-6.
44. van Ruissen F, Baas F: **Serial analysis of gene expression (SAGE)**. *Methods Mol Biol* 2007, **383**:41-66.
45. Hu M, Polyak K: **Serial analysis of gene expression**. *Nat Protoc* 2006, **1**(4):1743-1760.
46. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M: **Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing**. *Biotechniques* 2008, **45**(1):81-94.
47. Hanriot L, Keime C, Gay N, Faure C, Dossat C, Wincker P, Scote-Blachon C, Peyron C, Gandrillon O: **A combination of LongSAGE with Solexa sequencing is well suited to explore the depth and the complexity of transcriptome**. *BMC Genomics* 2008, **9**:418.
48. Heller MJ: **DNA microarray technology: devices, systems, and applications**. *Annu Rev Biomed Eng* 2002, **4**:129-153.
49. Bustin SA, Benes V, Nolan T, Pfaffl MW: **Quantitative real-time RT-PCR--a perspective**. *J Mol Endocrinol* 2005, **34**(3):597-601.
50. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**(2):281-297.
51. Bartel DP: **MicroRNAs: target recognition and regulatory functions**. *Cell* 2009, **136**(2):215-233.

52. O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT: **c-Myc-regulated microRNAs modulate E2F1 expression**. *Nature* 2005, **435**(7043):839-843.
53. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA *et al*: **MicroRNA expression profiles classify human cancers**. *Nature* 2005, **435**(7043):834-838.
54. Rose GD, Fleming PJ, Banavar JR, Maritan A: **A backbone-based theory of protein folding**. *Proc Natl Acad Sci U S A* 2006, **103**(45):16623-16633.
55. Anfinsen CB: **Principles that govern the folding of protein chains**. *Science* 1973, **181**(96):223-230.
56. Brognard J, Hunter T: **Protein kinase signaling networks in cancer**. *Curr Opin Genet Dev* 2011, **21**(1):4-11.
57. Haughian JM, Reno EM, Thorne AM, Bradford AP: **Protein kinase C alpha-dependent signaling mediates endometrial cancer cell growth and tumorigenesis**. *Int J Cancer* 2009, **125**(11):2556-2564.
58. Murray NR, Kalari KR, Fields AP: **Protein kinase Ciota expression and oncogenic signaling mechanisms in cancer**. *J Cell Physiol* 2011, **226**(4):879-887.
59. Dominguez I, Sonenshein GE, Seldin DC: **Protein kinase CK2 in health and disease: CK2 and its role in Wnt and NF-kappaB signaling: linking development and cancer**. *Cell Mol Life Sci* 2009, **66**(11-12):1850-1857.
60. Klysik J, Theroux SJ, Sedivy JM, Moffit JS, Boekelheide K: **Signaling crossroads: the function of Raf kinase inhibitory protein in cancer, the central nervous system and reproduction**. *Cell Signal* 2008, **20**(1):1-9.
61. MacBeath G: **Protein microarrays and proteomics**. *Nat Genet* 2002, **32 Suppl**:526-532.
62. Ellington AA, Kullo IJ, Bailey KR, Klee GG: **Antibody-based protein multiplex platforms: technical and operational challenges**. *Clin Chem* 2010, **56**(2):186-193.
63. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S *et al*: **Repeated observation of breast tumor subtypes in independent gene expression data sets**. *Proc Natl Acad Sci U S A* 2003, **100**(14):8418-8423.
64. Lamant L, de Reynies A, Duplantier MM, Rickman DS, Sabourdy F, Giuriato S, Brugieres L, Gaulard P, Espinos E, Delsol G: **Gene-expression profiling of systemic anaplastic large-cell lymphoma reveals differences based on ALK status and two distinct morphologic ALK+ subtypes**. *Blood* 2007, **109**(5):2156-2164.
65. Oestreicher N, Ramsey SD, Linden HM, McCune JS, van't Veer LJ, Burke W, Veenstra DL: **Gene expression profiling and breast cancer care: what are the potential benefits and policy implications?** *Genet Med* 2005, **7**(6):380-389.

66. Michaelson JJ, Loguercio S, Beyer A: **Detection and interpretation of expression quantitative trait loci (eQTL)**. *Methods* 2009, **48**(3):265-276.
67. Gerrits A, Li Y, Tesson BM, Bystrykh LV, Weersing E, Ausema A, Dontje B, Wang X, Breitling R, Jansen RC *et al*: **Expression quantitative trait loci are highly sensitive to cellular differentiation state**. *PLoS Genet* 2009, **5**(10):e1000692.
68. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ *et al*: **A gene-expression signature as a predictor of survival in breast cancer**. *N Engl J Med* 2002, **347**(25):1999-2009.
69. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530-536.
70. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T *et al*: **A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer**. *N Engl J Med* 2004, **351**(27):2817-2826.
71. Riley RD, Heney D, Jones DR, Sutton AJ, Lambert PC, Abrams KR, Young B, Wailoo AJ, Burchill SA: **A systematic review of molecular and biological tumor markers in neuroblastoma**. *Clin Cancer Res* 2004, **10**(1 Pt 1):4-12.
72. Cobleigh MA, Vogel CL, Tripathy D, Robert NJ, Scholl S, Fehrenbacher L, Wolter JM, Paton V, Shak S, Lieberman G *et al*: **Multinational study of the efficacy and safety of humanized anti-HER2 monoclonal antibody in women who have HER2-overexpressing metastatic breast cancer that has progressed after chemotherapy for metastatic disease**. *J Clin Oncol* 1999, **17**(9):2639-2648.
73. Seidman AD, Fornier MN, Esteva FJ, Tan L, Kaptain S, Bach A, Panageas KS, Arroyo C, Valero V, Currie V *et al*: **Weekly trastuzumab and paclitaxel therapy for metastatic breast cancer with analysis of efficacy by HER2 immunophenotype and gene amplification**. *J Clin Oncol* 2001, **19**(10):2587-2595.
74. Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, Tebbutt NC, Simes RJ, Chalchal H, Shapiro JD, Robitaille S *et al*: **K-ras mutations and benefit from cetuximab in advanced colorectal cancer**. *N Engl J Med* 2008, **359**(17):1757-1765.
75. Tol J, Koopman M, Cats A, Rodenburg CJ, Creemers GJ, Schrama JG, Erdkamp FL, Vos AH, van Groeningen CJ, Sinnige HA *et al*: **Chemotherapy, bevacizumab, and cetuximab in metastatic colorectal cancer**. *N Engl J Med* 2009, **360**(6):563-572.
76. Sekine I, Shimizu C, Nishio K, Saijo N, Tamura T: **A literature review of molecular markers predictive of clinical response to cytotoxic**

**chemotherapy in patients with breast cancer**. *Int J Clin Oncol* 2009, **14**(2):112-119.

77. Taube SE, Clark GM, Dancey JE, McShane LM, Sigman CC, Gutman SI: **A perspective on challenges and issues in biomarker development and drug and biomarker codevelopment**. *J Natl Cancer Inst* 2009, **101**(21):1453-1463.
78. **Vemurafenib (Zelboraf) for metastatic melanoma**. *Med Lett Drugs Ther* 2011, **53**(1374):77-78.
79. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA, Jr., Marks JR, Nevins JR: **Predicting the clinical status of human breast cancer by using gene expression profiles**. *Proc Natl Acad Sci U S A* 2001, **98**(20):11462-11467.
80. Ma XJ, Patel R, Wang X, Salunga R, Murage J, Desai R, Tuggle JT, Wang W, Chu S, Stecker K *et al*: **Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay**. *Arch Pathol Lab Med* 2006, **130**(4):465-473.
81. Chou TC, Talalay P: **Generalized equations for the analysis of inhibitions of Michaelis-Menten and higher-order kinetic systems with two or more mutually exclusive and nonexclusive inhibitors**. *Eur J Biochem* 1981, **115**(1):207-216.
82. Kerbel RS: **Antiangiogenic therapy: a universal chemosensitization strategy for cancer?** *Science* 2006, **312**(5777):1171-1175.
83. de Tayrac M, Aubry M, Saikali S, Etcheverry A, Surbled C, Guenot F, Galibert MD, Hamlat A, Lesimple T, Quillien V *et al*: **A 4-gene signature associated with clinical outcome in high-grade gliomas**. *Clin Cancer Res* 2011, **17**(2):317-327.
84. Peters CJ, Rees JR, Hardwick RH, Hardwick JS, Vowler SL, Ong CA, Zhang C, Save V, O'Donovan M, Rassl D *et al*: **A 4-gene signature predicts survival of patients with resected adenocarcinoma of the esophagus, junction, and gastric cardia**. *Gastroenterology* 2010, **139**(6):1995-2004 e1915.
85. Abel F, Dalevi D, Nethander M, Jornsten R, De Preter K, Vermuelen J, Stallings R, Kogner P, Maris J, Nilsson S: **A 6-gene signature identifies four molecular subgroups of neuroblastoma**. *Cancer Cell Int* 2011, **11**(1):9.
86. Huang H, Shiffman ML, Friedman S, Venkatesh R, Bzowej N, Abar OT, Rowland CM, Catanese JJ, Leong DU, Sninsky JJ *et al*: **A 7 gene signature identifies the risk of developing cirrhosis in patients with chronic hepatitis C**. *Hepatology* 2007, **46**(2):297-306.
87. Zhang X, Yang JJ, Kim YS, Kim KY, Ahn WS, Yang S: **An 8-gene signature, including methylated and down-regulated glutathione peroxidase 3, of gastric cancer**. *Int J Oncol* 2010, **36**(2):405-414.
88. Onaitis M, D'Amico TA, Clark CP, Guinney J, Harpole DH, Rawlins EL: **A 10-gene progenitor cell signature predicts poor prognosis in lung adenocarcinoma**. *Ann Thorac Surg* 2011, **91**(4):1046-1050.

89. Mettu RK, Wan YW, Habermann JK, Ried T, Guo NL: **A 12-gene genomic instability signature predicts clinical outcomes in multiple cancer types**. *Int J Biol Markers* 2010, **25**(4):219-228.
90. Mettu RK, Wan YW, Habermann JK, Ried T, Guo NL: **A 12-gene genomic instability signature predicts clinical outcomes in multiple cancer types**. *Int J Biol Markers* 2010, **25**(4).
91. Jezequel P, Campone M, Roche H, Gouraud W, Charbonnel C, Ricolleau G, Magrangeas F, Minvielle S, Geneve J, Martin AL *et al*: **A 38-gene expression signature to predict metastasis risk in node-positive breast cancer after systemic adjuvant chemotherapy: a genomic substudy of PACS01 clinical trial**. *Breast Cancer Res Treat* 2009, **116**(3):509-520.
92. Carlucci F, Marinello E, Tommassini V, Pisano B, Rosi F, Tabucchi A: **A 57-gene expression signature in B-cell chronic lymphocytic leukemia**. *Biomed Pharmacother* 2009, **63**(9):663-671.
93. Vermeulen J, De Preter K, Laureys G, Speleman F, Vandesompele J: **59-gene prognostic signature sub-stratifies high-risk neuroblastoma patients**. *Lancet Oncol* 2009, **10**(11):1030.
94. Mook S, Schmidt MK, Weigelt B, Kreike B, Eekhout I, van de Vijver MJ, Glas AM, Floore A, Rutgers EJ, van 't Veer LJ: **The 70-gene prognosis signature predicts early metastasis in breast cancer patients between 55 and 70 years of age**. *Ann Oncol* 2010, **21**(4):717-722.
95. Mook S, Schmidt MK, Viale G, Pruneri G, Eekhout I, Floore A, Glas AM, Bogaerts J, Cardoso F, Piccart-Gebhart MJ *et al*: **The 70-gene prognosis-signature predicts disease outcome in breast cancer patients with 1-3 positive lymph nodes in an independent validation study**. *Breast Cancer Res Treat* 2009, **116**(2):295-302.
96. Straver ME, Glas AM, Hannemann J, Wesseling J, van de Vijver MJ, Rutgers EJ, Vrancken Peeters MJ, van Tinteren H, Van't Veer LJ, Rodenhuis S: **The 70-gene signature as a response predictor for neoadjuvant chemotherapy in breast cancer**. *Breast Cancer Res Treat* 2010, **119**(3):551-558.
97. Zhang Y, Sieuwerts AM, McGreevy M, Casey G, Cufer T, Paradiso A, Harbeck N, Span PN, Hicks DG, Crowe J *et al*: **The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy**. *Breast Cancer Res Treat* 2009, **116**(2):303-309.
98. Metzeler KH, Hummel M, Bloomfield CD, Spiekermann K, Braess J, Sauerland MC, Heinecke A, Radmacher M, Marcucci G, Whitman SP *et al*: **An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia**. *Blood* 2008, **112**(10):4193-4201.
99. Gusnanto A, Ploner A, Pawitan Y: **Fold-change estimation of differentially expressed genes using mixture mixed-model**. *Stat Appl Genet Mol Biol* 2005, **4**:Article26.

100. 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.
101. Zhang S: **A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance**. *BMC Bioinformatics* 2007, **8**:230.
102. 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.
103. Long AD, Mangalam HJ, Chan BY, Tolleri L, Hatfield GW, Baldi P: **Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12**. *J Biol Chem* 2001, **276**(23):19937-19944.
104. Delmar P, Robin S, Daudin JJ: **VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data**. *Bioinformatics* 2005, **21**(4):502-508.
105. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3**:Article3.
106. Benjamini Y HY: **Controlling the false discovery rate: A practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society Series B(Methodological)* 1995, **57**:289–300.
107. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A: **False discovery rate, sensitivity and sample size for microarray studies**. *Bioinformatics* 2005, **21**(13):3017-3024.
108. Qian HR, Huang S: **Comparison of false discovery rate methods in identifying genes with differential expression**. *Genomics* 2005, **86**(4):495-503.
109. Broberg P: **A comparative review of estimates of the proportion unchanged genes and the false discovery rate**. *BMC Bioinformatics* 2005, **6**:199.
110. Giacomantonio CE, Goodhill GJ: **A Boolean model of the gene regulatory network underlying Mammalian cortical area development**. *PLoS Comput Biol* 2010, **6**(9).
111. Graudenzi A, Serra R, Villani M, Colacci A, Kauffman SA: **Robustness analysis of a boolean model of gene regulatory network with memory**. *J Comput Biol* 2011, **18**(4):559-577.
112. Graudenzi A, Serra R, Villani M, Damiani C, Colacci A, Kauffman SA: **Dynamical Properties of a Boolean Model of Gene Regulatory Network with Memory**. *J Comput Biol* 2011.
113. Hickman GJ, Hodgman TC: **Inference of gene regulatory networks using boolean-network inference methods**. *J Bioinform Comput Biol* 2009, **7**(6):1013-1029.

114. Hobert O, Carrera I, Stefanakis N: **The molecular and gene regulatory signature of a neuron**. *Trends Neurosci* 2010, **33**(10):435-445.
115. Woolf PJ, Wang Y: **A fuzzy logic approach to analyzing gene expression data**. *Physiol Genomics* 2000, **3**(1):9-15.
116. Ressom H, Reynolds R, Varghese RS: **Increasing the efficiency of fuzzy logic-based gene expression data analysis**. *Physiol Genomics* 2003, **13**(2):107-117.
117. Garcia-Martin E, Pizarro RM, Martinez C, Gutierrez-Martin Y, Perez G, Jover R, Agundez JA: **Acquired resistance to the anticancer drug paclitaxel is associated with induction of cytochrome P450 2C8**. *Pharmacogenomics* 2006, **7**(4):575-585.
118. Mantel N: **The detection of disease clustering and a generalized regression approach**. *Cancer Res* 1967, **27**(2):209-220.
119. Hoeflich KP, O'Brien C, Boyd Z, Cavet G, Guerrero S, Jung K, Januario T, Savage H, Punnoose E, Truong T *et al*: **In vivo antitumor activity of MEK and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models**. *Clin Cancer Res* 2009, **15**(14):4649-4664.
120. Salh B, Marotta A, Matthewson C, Ahluwalia M, Flint J, Owen D, Pelech S: **Investigation of the Mek-MAP kinase-Rsk pathway in human breast cancer**. *Anticancer Res* 1999, **19**(1B):731-740.
121. Polo ML, Arnoni MV, Riggio M, Wargon V, Lanari C, Novaro V: **Responsiveness to PI3K and MEK inhibitors in breast cancer. Use of a 3D culture system to study pathways related to hormone independence in mice**. *PLoS One* 2010, **5**(5):e10786.
122. Mirzoeva OK, Das D, Heiser LM, Bhattacharya S, Siwak D, Gendelman R, Bayani N, Wang NJ, Neve RM, Guan Y *et al*: **Basal subtype and MAPK/ERK kinase (MEK)-phosphoinositide 3-kinase feedback signaling determine susceptibility of breast cancer cells to MEK inhibition**. *Cancer Res* 2009, **69**(2):565-572.
123. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS: **Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks**. *Proc Natl Acad Sci U S A* 2000, **97**(22):12182-12186.
124. Butte AJ, Kohane IS: **Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements**. *Pac Symp Biocomput* 2000:418-429.
125. Mahony S, McInerney JO, Smith TJ, Golden A: **Gene prediction using the Self-Organizing Map: automatic generation of multiple gene models**. *BMC Bioinformatics* 2004, **5**:23.
126. Ressom H, Wang D, Natarajan P: **Clustering gene expression data using adaptive double self-organizing map**. *Physiol Genomics* 2003, **14**(1):35-46.
127. Mao Y, Zhao X, Wang S, Cheng Y: **Urinary nucleosides based potential biomarker selection by support vector machine for bladder cancer recognition**. *Anal Chim Acta* 2007, **598**(1):34-40.

128. Tang EK, Suganthan PN, Yao X: **Gene selection algorithms for microarray data based on least squares support vector machine**. *BMC Bioinformatics* 2006, **7**:95.
129. Herrero J, Valencia A, Dopazo J: **A hierarchical unsupervised growing neural network for clustering gene expression patterns**. *Bioinformatics* 2001, **17**(2):126-136.
130. Schietgat L, Vens C, Struyf J, Blockeel H, Kocev D, Dzeroski S: **Predicting gene function using hierarchical multi-label decision tree ensembles**. *BMC Bioinformatics* 2010, **11**:2.
131. Ghazoui Z, Buffa FM, Dunbier AK, Anderson H, Dexter T, Detre S, Salter J, Smith IE, Harris AL, Dowsett M: **Close and stable relationship between proliferation and a hypoxia metagene in aromatase inhibitor treated ER-positive breast cancer**. *Clin Cancer Res* 2011.
132. Tsigelny IF, Kouznetsova VL, Sweeney DE, Wu W, Bush KT, Nigam SK: **Analysis of metagene portraits reveals distinct transitions during kidney organogenesis**. *Sci Signal* 2008, **1**(49):ra16.
133. Noguchi H, Park J, Takagi T: **MetaGene: prokaryotic gene finding from environmental genome shotgun sequences**. *Nucleic Acids Res* 2006, **34**(19):5623-5630.
134. Paananen J, Storvik M, Wong G: **CROPPER: a metagene creator resource for cross-platform and cross-species compendium studies**. *BMC Bioinformatics* 2006, **7**:418.
135. Tamayo P, Scanfeld D, Ebert BL, Gillette MA, Roberts CW, Mesirov JP: **Metagene projection for cross-platform, cross-species characterization of global transcriptional states**. *Proc Natl Acad Sci U S A* 2007, **104**(14):5959-5964.
136. Lee DD, Seung HS: **Learning the parts of objects by non-negative matrix factorization**. *Nature* 1999, **401**(6755):788-791.
137. Greene D, Cagney G, Krogan N, Cunningham P: **Ensemble non-negative matrix factorization methods for clustering protein-protein interactions**. *Bioinformatics* 2008, **24**(15):1722-1728.
138. Frigyesi A, Hoglund M: **Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes**. *Cancer Inform* 2008, **6**:275-292.
139. Carmona-Saez P, Pascual-Marqui RD, Tirado F, Carazo JM, Pascual-Montano A: **Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization**. *BMC Bioinformatics* 2006, **7**:78.
140. Gao Y, Church G: **Improving molecular cancer class discovery through sparse non-negative matrix factorization**. *Bioinformatics* 2005, **21**(21):3970-3975.
141. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A *et al*: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies**. *Nature* 2006, **439**(7074):353-357.

142. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB *et al*: **Somatic mutations affect key pathways in lung adenocarcinoma**. *Nature* 2008, **455**(7216):1069-1075.
143. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M *et al*: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses**. *Proc Natl Acad Sci U S A* 2001, **98**(24):13790-13795.
144. Loke P, Hammond SN, Leung JM, Kim CC, Batra S, Rocha C, Balmaseda A, Harris E: **Gene expression patterns of dengue virus-infected children from nicaragua reveal a distinct signature of increased metabolism**. *PLoS Negl Trop Dis* 2010, **4**(6):e710.
145. Tavolaro S, Chiaretti S, Messina M, Peragine N, Del Giudice I, Marinelli M, Santangelo S, Mauro FR, Guarini A, Foa R: **Gene expression profile of protein kinases reveals a distinctive signature in chronic lymphocytic leukemia and in vitro experiments support a role of second generation protein kinase inhibitors**. *Leuk Res* 2010, **34**(6):733-741.
146. Poynton HC, Zuzow R, Loguinov AV, Perkins EJ, Vulpe CD: **Gene expression profiling in Daphnia magna, part II: validation of a copper specific gene expression signature with effluent from two copper mines in California**. *Environ Sci Technol* 2008, **42**(16):6257-6263.
147. Badot V, Galant C, Nzeusseu Toukap A, Theate I, Maudoux AL, Van den Eynde BJ, Durez P, Houssiau FA, Lauwerys BR: **Gene expression profiling in the synovium identifies a predictive signature of absence of response to adalimumab therapy in rheumatoid arthritis**. *Arthritis Res Ther* 2009, **11**(2):R57.
148. Defamie V, Cursio R, Le Brigand K, Moreilhon C, Saint-Paul MC, Laurens M, Crenesse D, Cardinaud B, Auberger P, Gugenheim J *et al*: **Gene expression profiling of human liver transplants identifies an early transcriptional signature associated with initial poor graft function**. *Am J Transplant* 2008, **8**(6):1221-1236.
149. Nakayama R, Mitani S, Nakagawa T, Hasegawa T, Kawai A, Morioka H, Yabe H, Toyama Y, Ogose A, Toguchida J *et al*: **Gene expression profiling of synovial sarcoma: distinct signature of poorly differentiated type**. *Am J Surg Pathol* 2010, **34**(11):1599-1607.
150. Singh A, Greninger P, Rhodes D, Koopman L, Violette S, Bardeesy N, Settleman J: **A gene expression signature associated with "K-Ras addiction" reveals regulators of EMT and tumor cell survival**. *Cancer Cell* 2009, **15**(6):489-500.
151. Lafferty-Whyte K, Cairney CJ, Will MB, Serakinci N, Daidone MG, Zaffaroni N, Bilsland A, Keith WN: **A gene expression signature classifying telomerase and ALT immortalization reveals an hTERT regulatory**

**network and suggests a mesenchymal stem cell origin for ALT**. *Oncogene* 2009, **28**(43):3765-3774.

152. Watanabe T, Kobunai T, Sakamoto E, Yamamoto Y, Konishi T, Horiuchi A, Shimada R, Oka T, Nagawa H: **Gene expression signature for recurrence in stage III colorectal cancers**. *Cancer* 2009, **115**(2):283-292.
153. Rambow F, Piton G, Bouet S, Leplat JJ, Baulande S, Marrau A, Stam M, Horak V, Vincent-Naulleau S: **Gene expression signature for spontaneous cancer regression in melanoma pigs**. *Neoplasia* 2008, **10**(7):714-726, 711 p following 726.
154. Miller TW, Balko JM, Ghazoui Z, Dunbier A, Anderson H, Dowsett M, Gonzalez-Angulo AM, Mills GB, Miller WR, Wu H *et al*: **A Gene Expression Signature from Human Breast Cancer Cells with Acquired Hormone Independence Identifies MYC as a Mediator of Antiestrogen Resistance**. *Clin Cancer Res* 2011, **17**(7):2024-2034.
155. Sabatier R, Finetti P, Cervera N, Lambaudie E, Esterni B, Mamessier E, Tallet A, Chabannon C, Extra JM, Jacquemier J *et al*: **A gene expression signature identifies two prognostic subgroups of basal breast cancer**. *Breast Cancer Res Treat* 2011, **126**(2):407-420.
156. Mengual L, Burset M, Ribal MJ, Ars E, Marin-Aguilera M, Fernandez M, Ingelmo-Torres M, Villavicencio H, Alcaraz A: **Gene expression signature in urine for diagnosing and assessing aggressiveness of bladder urothelial carcinoma**. *Clin Cancer Res* 2010, **16**(9):2624-2633.
157. Horesh Y, Katsel P, Haroutunian V, Domany E: **Gene expression signature is shared by patients with Alzheimer's disease and schizophrenia at the superior temporal gyrus**. *Eur J Neurol* 2011, **18**(3):410-424.
158. Kim HK, Choi IJ, Kim CG, Kim HS, Oshima A, Michalowski A, Green JE: **A gene expression signature of acquired chemoresistance to cisplatin and fluorouracil combination chemotherapy in gastric cancer patients**. *PLoS One* 2011, **6**(2):e16694.
159. McWeeney SK, Pemberton LC, Loriaux MM, Vartanian K, Willis SG, Yochum G, Wilmot B, Turpaz Y, Pillai R, Druker BJ *et al*: **A gene expression signature of CD34+ cells to predict major cytogenetic response in chronic-phase chronic myeloid leukemia patients treated with imatinib**. *Blood* 2010, **115**(2):315-325.
160. Laffaire J, Rivals I, Dauphinot L, Pasteau F, Wehrle R, Larrat B, Vitalis T, Moldrich RX, Rossier J, Sinkus R *et al*: **Gene expression signature of cerebellar hypoplasia in a mouse model of Down syndrome during postnatal development**. *BMC Genomics* 2009, **10**:138.
161. Porpaczy E, Bilban M, Heinze G, Gruber M, Vanura K, Schwarzinger I, Stilgenbauer S, Streubel B, Fonatsch C, Jaeger U: **Gene expression signature of chronic lymphocytic leukaemia with Trisomy 12**. *Eur J Clin Invest* 2009, **39**(7):568-575.
162. Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, Mann FE, Fukuoka J, Hames M, Bergen AW *et al*: **Gene expression signature of**

**cigarette smoking and its role in lung adenocarcinoma development and survival**. *PLoS One* 2008, **3**(2):e1651.

163. Kennerly E, Ballmann A, Martin S, Wolfinger R, Gregory S, Stoskopf M, Gibson G: **A gene expression signature of confinement in peripheral blood of red wolves (Canis rufus)**. *Mol Ecol* 2008, **17**(11):2782-2791.
164. Assou S, Cerecedo D, Tondeur S, Pantesco V, Hovatta O, Klein B, Hamamah S, De Vos J: **A gene expression signature shared by human mature oocytes and embryonic stem cells**. *BMC Genomics* 2009, **10**:10.
165. Chanrion M, Negre V, Fontaine H, Salvetat N, Bibeau F, Mac Grogan G, Mauriac L, Katsaros D, Molina F, Theillet C *et al*: **A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer**. *Clin Cancer Res* 2008, **14**(6):1744-1752.
166. Ellsworth RE, Seebach J, Field LA, Heckman C, Kane J, Hooke JA, Love B, Shriver CD: **A gene expression signature that defines breast cancer metastases**. *Clin Exp Metastasis* 2009, **26**(3):205-213.
167. Lin Y, Lin S, Watson M, Trinkaus KM, Kuo S, Naughton MJ, Weilbaecher K, Fleming TP, Aft RL: **A gene expression signature that predicts the therapeutic response of the basal-like breast cancer to neoadjuvant chemotherapy**. *Breast Cancer Res Treat* 2010, **123**(3):691-699.
168. Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, Lopez-Doriga A, Santos C, Marijnen C, Westerga J *et al*: **Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer**. *J Clin Oncol* 2011, **29**(1):17-24.
169. Li YG, Geng X: **A meta-analysis on the association of HER-2 overexpression with prognosis in human osteosarcoma**. *Eur J Cancer Care (Engl)* 2010, **19**(3):313-316.
170. Zheng Z, Pan TC, Li J, Chen T, Song DW, Yi J: **[Meta-analysis of relationship between lymph node micrometastasis and prognosis in stage I non-small cell lung cancer patients]**. *Ai Zheng* 2004, **23**(2):185-188.
171. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schutz F *et al*: **Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures**. *Breast Cancer Res* 2008, **10**(4):R65.
172. Araujo SE, Bernardo WM, Habr-Gama A, Kiss DR, Cecconello I: **DNA ploidy status and prognosis in colorectal cancer: a meta-analysis of published data**. *Dis Colon Rectum* 2007, **50**(11):1800-1810.
173. Thum KE, Shin MJ, Gutierrez RA, Mukherjee I, Katari MS, Nero D, Shasha D, Coruzzi GM: **An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in Arabidopsis**. *BMC Syst Biol* 2008, **2**:31.
174. Sanai N: **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma**. *World Neurosurg* 2010, **74**(1):4-5.

175. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP *et al*: **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1**. *Cancer Cell* 2010, **17**(1):98-110.
176. Tian Q, Stepaniants SB, Mao M, Weng L, Feetham MC, Doyle MJ, Yi EC, Dai H, Thorsson V, Eng J *et al*: **Integrated genomic and proteomic analyses of gene expression in Mammalian cells**. *Mol Cell Proteomics* 2004, **3**(10):960-969.
177. Chapman EJ, Williams SV, Platt FM, Hurst CD, Chambers P, Roberts P, Knowles MA: **Integrated genomic and transcriptional analysis of the in vitro evolution of telomerase-immortalized urothelial cells (TERT-NHUC)**. *Genes Chromosomes Cancer* 2009, **48**(8):694-710.
178. Vincent-Salomon A, Lucchesi C, Gruel N, Raynal V, Pierron G, Goudefroye R, Reyal F, Radvanyi F, Salmon R, Thiery JP *et al*: **Integrated genomic and transcriptomic analysis of ductal carcinoma in situ of the breast**. *Clin Cancer Res* 2008, **14**(7):1956-1965.
179. Yang D, Jiang Y, He F: **An integrated view of the correlations between genomic and phenomic variables**. *J Genet Genomics* 2009, **36**(11):645-651.
180. Kuo KW, Yang PY, Huang YS, Shieh DZ: **Variations in gene expression and genomic stability of human hepatoma cells integrated with hepatitis B virus DNA**. *Biochem Mol Biol Int* 1998, **44**(6):1133-1140.
181. Reif DM, White BC, Moore JH: **Integrated analysis of genetic, genomic and proteomic data**. *Expert Rev Proteomics* 2004, **1**(1):67-75.
182. Salari K, Tibshirani R, Pollack JR: **DR-Integrator: a new analytic tool for integrating DNA copy number and gene expression data**. *Bioinformatics* 2010, **26**(3):414-416.
183. Bussey KJ, Chin K, Lababidi S, Reimers M, Reinhold WC, Kuo WL, Gwadry F, Ajay, Kouros-Mehr H, Fridlyand J *et al*: **Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel**. *Mol Cancer Ther* 2006, **5**(4):853-867.
184. Li M, Balch C, Montgomery JS, Jeong M, Chung JH, Yan P, Huang TH, Kim S, Nephew KP: **Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer**. *BMC Med Genomics* 2009, **2**:34.
185. Kars MD, Iseri OD, Gunduz U: **A microarray based expression profiling of paclitaxel and vincristine resistant MCF-7 cells**. *Eur J Pharmacol* 2011, **657**(1-3):4-9.
186. Hanahan D, Weinberg RA: **The hallmarks of cancer**. *Cell* 2000, **100**(1):57-70.
187. Macconaill LE, Garraway LA: **Clinical implications of the cancer genome**. *J Clin Oncol* 2010, **28**(35):5219-5228.
188. Trainer AH, Meiser B, Watts K, Mitchell G, Tucker K, Friedlander M: **Moving toward personalized medicine: treatment-focused genetic testing**

**of women newly diagnosed with ovarian cancer**. *Int J Gynecol Cancer* 2010, **20**(5):704-716.

189. Weller M, Wick W, Hegi ME, Stupp R, Tabatabai G: **Should biomarkers be used to design personalized medicine for the treatment of glioblastoma?** *Future Oncol* 2010, **6**(9):1407-1414.
190. Woolf SH: **The meaning of translational research and why it matters**. *JAMA* 2008, **299**(2):211-213.
191. van Krieken H, Tol J: **Setting future standards for KRAS testing in colorectal cancer**. *Pharmacogenomics* 2009, **10**(1):1-3.
192. Wiseman LR, Spencer CM: **Paclitaxel. An update of its use in the treatment of metastatic breast cancer and ovarian and other gynaecological cancers**. *Drugs Aging* 1998, **12**(4):305-334.
193. Gottesman MM, Fojo T, Bates SE: **Multidrug resistance in cancer: role of ATP-dependent transporters**. *Nat Rev Cancer* 2002, **2**(1):48-58.
194. Loboda A, Nebozhyn M, Klinghoffer R, Frazier J, Chastain M, Arthur W, Roberts B, Zhang T, Chenard M, Haines B *et al*: **A gene expression signature of RAS pathway dependence predicts response to PI3K and RAS pathway inhibitors and expands the population of RAS pathway activated tumors**. *BMC Med Genomics* 2010, **3**:26.
195. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN *et al*: **Chemosensitivity prediction by transcriptional profiling**. *Proc Natl Acad Sci U S A* 2001, **98**(19):10787-10792.
196. Fiebig HH, Schuler J, Bausch N, Hofmann M, Metz T, Korrat A: **Gene signatures developed from patient tumor explants grown in nude mice to predict tumor response to 11 cytotoxic drugs**. *Cancer Genomics Proteomics* 2007, **4**(3):197-209.
197. Stein WD, Litman T, Fojo T, Bates SE: **A Serial Analysis of Gene Expression (SAGE) database analysis of chemosensitivity: comparing solid tumors with cell lines and comparing solid tumors from different tissue origins**. *Cancer Res* 2004, **64**(8):2805-2816.
198. Hartemink A: **Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks**. *PhD Dissertation.* Cambridge, Massachusetts: Massachusetts Institute of Technology; 2001.
199. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Res* 2002, **30**(1):207-210.
200. Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia JA, Booser D, Theriault RL, Buzdar AU, Dempsey PJ *et al*: **Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer**. *J Clin Oncol* 2006, **24**(26):4236-4244.
201. Kauffmann A, Huber W: **Microarray data quality control improves the detection of differentially expressed genes**. *Genomics* 2010, **95**(3):138-142.

202. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu TM, Goodsaid FM, Pusztai L *et al*: **The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models**. *Nat Biotechnol* 2010, **28**(8):827-838.
203. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY *et al*: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements**. *Nat Biotechnol* 2006, **24**(9):1151-1161.
204. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET *et al*: **An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival**. *Proc Natl Acad Sci U S A* 2005, **102**(38):13550-13555.
205. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **ONCOMINE: a cancer microarray database and integrated data-mining platform**. *Neoplasia* 2004, **6**(1):1-6.
206. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P *et al*: **Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles**. *Neoplasia* 2007, **9**(2):166-180.
207. Hudson CC, Liu M, Chiang GG, Otterness DM, Loomis DC, Kaper F, Giaccia AJ, Abraham RT: **Regulation of hypoxia-inducible factor 1alpha expression and function by the mammalian target of rapamycin**. *Mol Cell Biol* 2002, **22**(20):7004-7014.
208. Rubio-Viqueira B, Hidalgo M: **Targeting mTOR for cancer treatment**. *Adv Exp Med Biol* 2006, **587**:309-327.
209. Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe JP, Tong F *et al*: **A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes**. *Cancer Cell* 2006, **10**(6):515-527.
210. Kuo WL, Das D, Ziyad S, Bhattacharya S, Gibb WJ, Heiser LM, Sadanandam A, Fontenay GV, Hu Z, Wang NJ *et al*: **A systems analysis of the chemosensitivity of breast cancer cells to the polyamine analogue PG-11047**. *BMC Med* 2009, **7**:77.
211. Heiser LM, Wang NJ, Talcott CL, Laderoute KR, Knapp M, Guan Y, Hu Z, Ziyad S, Weber BL, Laquerre S *et al*: **Integrated analysis of breast cancer cell lines reveals unique signaling pathways**. *Genome Biol* 2009, **10**(3):R31.
212. Loss LA, Sadanandam A, Durinck S, Nautiyal S, Flaucher D, Carlton VE, Moorhead M, Lu Y, Gray JW, Faham M *et al*: **Prediction of epigenetically regulated genes in breast cancer cell lines**. *BMC Bioinformatics* 2010, **11**:305.
213. Creighton CJ, Fu X, Hennessy BT, Casa AJ, Zhang Y, Gonzalez-Angulo AM, Lluch A, Gray JW, Brown PH, Hilsenbeck SG *et al*: **Proteomic and transcriptomic profiling reveals a link between the PI3K pathway and**

**lower estrogen-receptor (ER) levels and activity in ER+ breast cancer**. *Breast Cancer Res* 2010, **12**(3):R40.

214. Layek R, Datta A, Bittner M, Dougherty ER: **Cancer therapy design based on pathway logic**. *Bioinformatics* 2011, **27**(4):548-555.
215. Fitzgerald JB, Schoeberl B, Nielsen UB, Sorger PK: **Systems biology and combination therapy in the quest for clinical efficacy**. *Nat Chem Biol* 2006, **2**(9):458-466.
216. Golub TR: **Mining the genome for combination therapies**. *Nat Med* 2003, **9**(5):510-511.
217. Bonetti A, Zaninelli M, Rodella S, Molino A, Sperotto L, Piubello Q, Bonetti F, Nortilli R, Turazza M, Cetto GL: **Tumor proliferative activity and response to first-line chemotherapy in advanced breast carcinoma**. *Breast Cancer Res Treat* 1996, **38**(3):289-297.
218. Carden CP, Sarker D, Postel-Vinay S, Yap TA, Attard G, Banerji U, Garrett MD, Thomas GV, Workman P, Kaye SB *et al*: **Can molecular biomarker-based patient selection in Phase I trials accelerate anticancer drug development?** *Drug Discov Today* 2010, **15**(3-4):88-97.
219. Cummings J, Raynaud F, Jones L, Sugar R, Dive C: **Fit-for-purpose biomarker method validation for application in clinical trials of anticancer drugs**. *Br J Cancer* 2010, **103**(9):1313-1317.
220. Li J, Donath S, Li Y, Qin D, Prabhakar BS, Li P: **miR-30 regulates mitochondrial fission through targeting p53 and the dynamin-related protein-1 pathway**. *PLoS Genet* 2010, **6**(1):e1000795.
221. Martinez I, Cazalla D, Almstead LL, Steitz JA, DiMaio D: **miR-29 and miR-30 regulate B-Myb expression during cellular senescence**. *Proc Natl Acad Sci U S A* 2011, **108**(2):522-527.
222. Shen J, Ambrosone CB, Zhao H: **Novel genetic variants in microRNA genes and familial breast cancer**. *Int J Cancer* 2009, **124**(5):1178-1182.
223. Yeang CH: **An integrated analysis of molecular aberrations in NCI-60 cell lines**. *BMC Bioinformatics* 2010, **11**:495.
224. Mane SP, Evans C, Cooper KL, Crasta OR, Folkerts O, Hutchison SK, Harkins TT, Thierry-Mieg D, Thierry-Mieg J, Jensen RV: **Transcriptome sequencing of the Microarray Quality Control (MAQC) RNA reference samples using next generation sequencing**. *BMC Genomics* 2009, **10**:264.
225. Arikawa E, Sun Y, Wang J, Zhou Q, Ning B, Dial SL, Guo L, Yang J: **Cross-platform comparison of SYBR Green real-time PCR with TaqMan PCR, microarrays and other gene expression measurement technologies evaluated in the MicroArray Quality Control (MAQC) study**. *BMC Genomics* 2008, **9**:328.
226. Chen JJ, Hsueh HM, Delongchamp RR, Lin CJ, Tsai CA: **Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data**. *BMC Bioinformatics* 2007, **8**:412.
227. Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu TM, Bao W, Fang H, Kawasaki ES, Hager J, Tikhonova IR *et al*: **Performance**

**comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project**. *Nat Biotechnol* 2006, **24**(9):1140-1150.

228. Park ES, Rabinovsky R, Carey M, Hennessy BT, Agarwal R, Liu W, Ju Z, Deng W, Lu Y, Woo HG *et al*: **Integrative analysis of proteomic signatures, mutations, and drug responsiveness in the NCI 60 cancer cell line set**. *Mol Cancer Ther* 2010, **9**(2):257-267.
229. Ozcan S: **MiR-30 family and EMT in human fetal pancreatic islets**. *Islets* 2009, **1**(3):283-285.
230. Joglekar MV, Patil D, Joglekar VM, Rao GV, Reddy DN, Mitnala S, Shouche Y, Hardikar AA: **The miR-30 family microRNAs confer epithelial phenotype to human pancreatic cells**. *Islets* 2009, **1**(2):137-147.
231. Duisters RF, Tijsen AJ, Schroen B, Leenders JJ, Lentink V, van der Made I, Herias V, van Leeuwen RE, Schellings MW, Barenbrug P *et al*: **miR-133 and miR-30 regulate connective tissue growth factor: implications for a role of microRNAs in myocardial matrix remodeling**. *Circ Res* 2009, **104**(2):170-178, 176p following 178.