Abstract

Personalized identification of altered pathway using accumulated data

TaeJin Ahn
Interdisciplinary Program in Bioinformatics
The Graduate School
Seoul National University

Identifying altered pathways in an individual is important for understanding disease mechanisms and for the future application of custom therapeutic decisions. Existing pathway analysis techniques are mainly focused on discovering altered pathways between normal and cancer groups and are not suitable for identifying the pathway aberrance that may occur in an individual sample. A simple way to identify individual’s pathway aberrance is to compare normal and tumor data from the same individual. However, the matched normal data from the same individual is often unavailable in clinical situation. We therefore suggest a new approach for the personalized identification of altered pathways, making special use of accumulated normal data in cases when a patient’s matched normal data is unavailable. The philosophy behind our method is to quantify the aberrance of an individual sample's pathway by comparing it to accumulated normal samples. We propose and examine personalized extensions of pathway statistics, Over-Representation Analysis (ORA) and Functional Class Scoring (FCS), to generate individualized pathway aberrance score (iPAS).

Collected microarray data of normal tissue of lung and colon mucosa is served as reference to investigate a number of cancer individuals of lung adenocarcinoma and colon cancer, respectively. Our method concurrently captures known facts of cancer survival pathways and identifies the pathway aberrances that represent cancer differentiation status and survival. It also provides more improved validation rate of survival related pathways than when a single cancer sample is interpreted in the context of cancer-only cohort. In addition, our method is useful in classifying unknown samples into cancer or normal groups. Particularly, we identified ‘amino acid synthesis and interconversion’ pathway is a good indicator of lung adenocarcinoma (AUC 0.982 at independent validation). We also suggest a new approach for discovering rare mutations that have functional impact in the context of pathway by iteratively combining rare mutations until no more mutations with pathway impact can be added. The approach is shown to sensitively capture mutations that change pathway level gene expression at breast cancer data.

Clinical importance of the method is providing pathway interpretation of single cancer even though its matched normal data is unavailable.

Keywords : Cancer, Gene expression, Pathway analysis, Network analysis, Personalized medicine

Student number : 2008-30134

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The goals of this chapter are to i) review the existing pathway analysis methods and ii) discuss limitations of each class of methods, iii) introducing personalization challenges in pathway analysis.

1.1 Existing pathway analysis approaches (Group to group)

Khatri et al. [1] provide comprehensive review for existing pathway analysis approaches. They summarized the importance of pathway analysis to understand complex biology and introduce key features of pathway analysis. They also provide classification of existing pathway analysis techniques into three generations: Over representation analysis (ORA), Functional Class Scoring (FCS) and Pathway topology based (PT).

1.1.1 Importance of pathway analysis

Pathway analysis has become the first choice for gaining insight into the underlying biology of differentially expressed genes and proteins, as it reduces complexity and has increased explanatory power.

High-throughput sequencing and gene/protein profiling techniques have transformed biological research by enabling comprehensive monitoring of a biological system. Analysis of high-throughput data typically yields a list of differentially expressed genes or proteins. This list is extremely useful in identifying genes that may have roles in a given phenomenon or phenotype. However, for many investigators, this list often fails to provide mechanistic insights into the underlying biology of the condition being studied. In this way, the advent of high-throughput profiling technologies presents a new challenge, that of extracting meaning from a long list of differentially expressed genes and proteins.

One approach to this challenge has been to simplify analysis by grouping long lists of individual genes into smaller sets of related genes or proteins. This approach reduces the complexity of analysis. Researchers have developed a large number of knowledge bases to help with this task. The knowledge bases describe biological processes, components, or structures in which individual genes and proteins are known to be involved in, as well as how and where gene products interact with each other. One example of this idea is to identify groups of genes that function in the same pathways.

Analyzing high-throughput molecular measurements in the context of pathway is useful for two reasons. First, grouping thousands of genes, proteins, and/or other biological molecules by the pathways they are involved in reduces the complexity to just several hundred pathways for the experiment. Second, identifying active pathways that differ between two conditions can have more explanatory power than a simple list of different genes or proteins

1.1.2 Component of pathway analysis

Pathway analysis can be divided into four important components: Gene level statistics, Pathway level statistics (Gene set level statistics), Assessing the significance and Correction for multiple hypothesis [1]. Jui-Hung et al., [2] mentioned the other important component, data processing (Figure 1.1). Data processing is practically very important. Thus, we introduce components of pathway analysis according to the Jui-Hung's definition.

Data preprocessing

There are two important but frequently overlooked data preprocessing steps. Normalization allows expression values obtained from different experiments to be directly comparable [3, 4]. The expression values of a small, but different set of genes may be missing in different microarray experiments due to technical issues. Imputation of missing data is thus important for maximal data coverage when the results of multiple experiments are compared. A number of methods are available for normalization [4, 5], yet this critical step is frequently omitted [3]. The most common normalization algorithms—RMA [4] and MAS 5.0 [6]—are designed for expression levels generated with microarrays that follow a lognormal distribution. Thus it is important to log transform the raw intensity values from microarrays. Failure to do so would bias toward high expression values, reducing statistical power because of increase in variance [4]. Log transform is also applied to RNA-seq data. Expression level determined by RNA-seq is usually quantified in Reads Per Kilobase exon Model per million mapped reads (RPKM; density of reads that map to a gene normalized for the length of its mature transcript and for the sequencing depth of the experiment [7]), which after log transform correlates well with normalized intensity measured with microarray, also after log transform [7]. Missing data can be imputed using methods based on K nearest neighbors, singular value decomposition, or least square regression models. Least square regression algorithms were reported to produce lower estimation error than other methods [8, 9]. In this article we use a popular least square regression algorithm, LSimpute_gene [10], to impute missing values in all 132 experimental data sets.

Single gene statistics

The first step in pathway analysis using mRNA data is to compute a gene-level statistic of differential expression, e.g. a t statistic, a signal to noise ratio (mean to standard deviation ratio), a fold change or a Wilcoxon rank sum statistic. Because phenotype change can affect different genes in opposite directions, i.e. increase the expression levels of some genes while decrease the levels of others, and we want to be able to identify the gene sets that contain both types of genes, it is desirable to eliminate the direction of differential expression by taking the absolute or square of the statistic [11, 12]. However, data transformations that eliminate direction—such as absolute values—lead to asymmetrical distributions, and can nullify some analytical estimates of significance based on analytical background distributions such as the w^2 test [13,14] (see 'Estimating significance' and 'The validity of analytical background distributions' sections). The many-to-many correspondence between genes and probe sets on a microarray creates ambiguity in determining expression levels of genes [15,16]. A common practice is to calculate the mean or median expression levels of the probe sets that correspond to the same gene; however, doing so usually increases the number of false negatives [17]. An alternative is to perform meta analysis [17], using for example, the method proposed by Fisher [18] or by Stouffer [19, 20]. Rather than merging the expression values directly, these methods merge probe set level statistics. A similar problem exists in RNA-seq, where some sequencing reads are mapped to multiple genomic locations. Such multi-mappers originate from paralogs, segmentally duplicated regions and low sequence complexity [21]. Ignoring multimappers reduces sensitivity and undercounts some genes [7]. Strategies for assigning multi-mappers are discussed in [7, 21, 22].

Pathway level statistics (Gene set-level statistics)

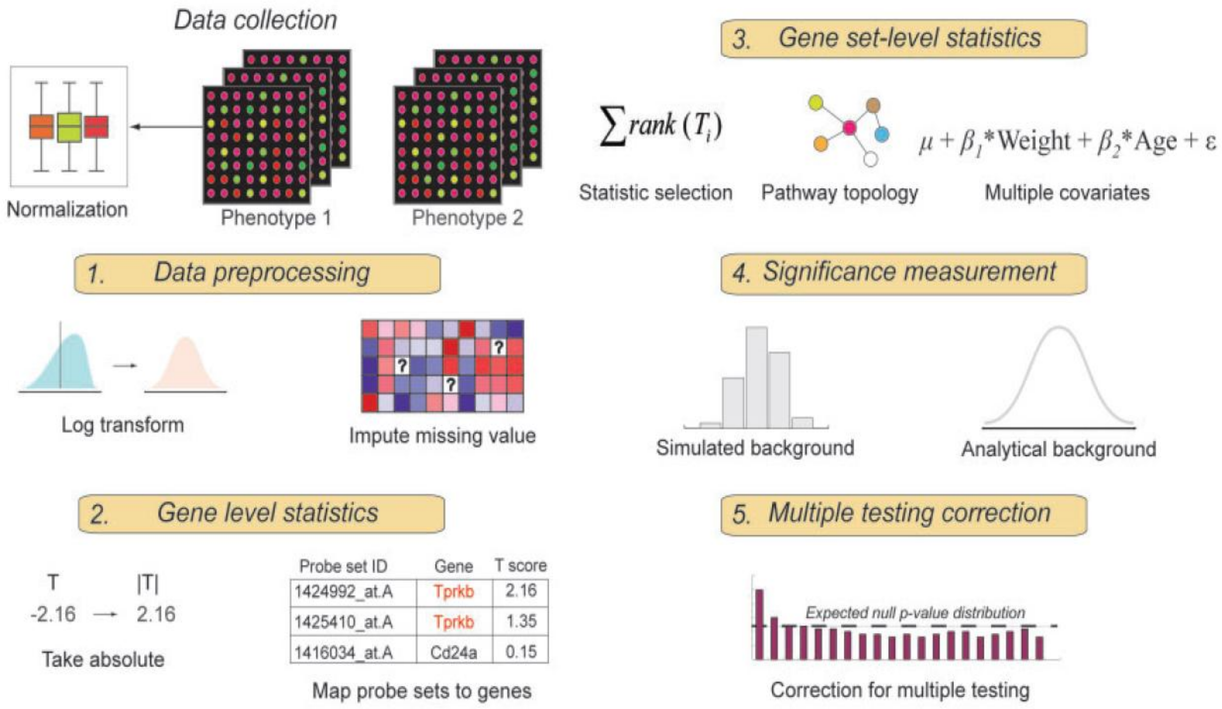
The purpose of a pathway level (it can also be called as “gene set-level”) statistic is to decide whether a gene set is distinct in some statistically significant way. A gene set statistic can be defined in terms of properties of the genes in the set, e.g. the mean, median, variance, etc. of a gene-level statistic. When a property (and its corresponding statistic) is chosen, the null hypothesis must, of course, also be specified. There are two null hypotheses as defined by Tian et al. [23]. In one case (Q1) the background distribution is obtained by shuffling genes; in the other (Q2), the background distribution is obtained by shuffling phenotypes, i.e. samples (see ‘Estimating significance’ section). The rationale for using Q1 is that a significant gene set should be distinguishable from an equal size set composed of randomly chosen genes. On the other hand,Q2 focuses on a gene set and tests whether its association with the phenotype change is distinguishable from randomly shuffled phenotype changes. Q2 is generally favored because it preserves the relationship of the genes in the set [23, 24, 25] and directly addresses the question of finding gene sets whose expression changes correlates with phenotype changes. Gene set-level statistics generally ignore clinical covariates—factors such as age, sex and weight—which can also cause differential expression, confounding the impact of phenotype changes [26]. The effect of covariates can be estimated using, for example, a linear regression model [26]. If a t statistic is used in the gene level, it can be generalized using a linear regression model for covariate correction, after accounting for the increased number of variables to avoid over fitting [26]. Most gene set-level statistics also ignore relationships among genes within the set. For example, if the gene set is a pathway, its topological information is ignored. Including topological information is important for accounting for the effect of genetic buffering [26], which deduces that if a gene fails to propagate its influence to a pathway neighbor, its biological role is buffered. Conversely, a gene that regulates many of its downstream genes may play a pivotal role in the expression changes of the pathway associated with phenotype changes. Methods for including topological information by weighted gene set-level statistics are discussed in [27, 28, 29].

Estimating significance

One can use significance in the standard way: the probability that the null hypothesis, evaluated on the background (or null) distribution, is correct. The background distribution can sometimes be written analytically, as in the case of a Gaussian distribution, and it can always be simulated by shuffling experimental data. As noted in the above section, simulated background is dictated by the choice of the null hypothesis (Q1 or Q2), which often leads to different conclusions [23]. Most frequently the gene set-level statistic, e.g. the mean of the t-statistic values of genes in the set, is assumed to follow a normal distribution when expression change has no association with phenotype change [23]. In such case the significance (P-value) ofa gene set can be computed analytically [14]. Such an assumption is in question when the expression levels of genes in a set are dependent on one another, which is common for genes in a pathway [34]. In 'The validity of analytical background distributions' section we will discuss analytical backgrounds and empirical corrections [14] to make them more useful. To be concrete in illustrating how significance is estimated using a simulated background distribution, suppose we are interested in estimating the probability that the enrichment score obtained for a particular gene set is a chance occurrence of phenotype changes. The procedure would be to shuffle the phenotype labels, calculate the differential expression of each gene, rank all genes and compute an enrichment score for the same gene set. The entire process is repeated multiple times to obtain a distribution of enrichment scores, and the P-value of the actual enrichment score is simply the fraction of shuffles that produce enrichment scores at least as great as observed. Although simulating the background distribution obviates the need of an analytical background, it can be computational demanding—at least N shuffles need to be performed to achieve a P-value resolution of 1/N [30].

Correction for multiple testing

Correction for multiple testing P-value is the appropriate measure of statistical significance when only one gene set is tested. When a large number of gene sets are tested, there can be many false positives among the gene sets that receive seemingly highly significant P-values; this is called the multiple hypothesis testing problems. The simplest procedure is to choose a P-value which, when multiplied by the number of hypotheses, i.e. the total number of tested gene sets, gives a sufficiently low corrected P-value, e.g. <0.05. This Bonferroni correction [31] is, however, very conservative and sometimes results in an unacceptably large number of false negatives. An alternative is to control the expected fraction of false positives among the predictions, or the false discover rate (FDR), using the method by Benjamini and Hochberg [32]. The original Benjamini–Hochberg procedure assumes a uniform distribution for the P-values [32]. In some cases when there are relatively many 'non-null' tests, i.e. when low P-values are prevalent, an FDR variant, positive FDR (pFDR) can be applied [32, 33]. The corrected P-value is called Q-value, defined as the 'minimum FDR at which a test is called significant' [32, 34]. The relationship between Q-value and FDR is analogous to that between P-value and type I error [32]. The final significant gene sets are the ones whose Q-values are smaller than an FDR threshold.



Probe set ID	Gene	T score
1424992_at.A	Tprkb	2.16
1425410_at.A	Tprkb	1.35
1416034_at.A	Cd24a	0.15

Figure 1.1 Key components of pathway analysis

Jui-Hung Hung et al.
Gene set enrichment analysis: performance evaluation and usage guidelines
Brief Bioinform (2012) 13 (3): 281-291 first published online September 7, 2011

1.1.3 Classification of existing pathway analysis approaches

Existing pathway analysis techniques can further be classified into three generations: Over representation analysis (ORA), Functional Class Scoring (FCS) and Pathway topology based (PT) [1] (Figure 1.2). List of existing tools collected by Khatri et al. for each generation can be found at table 1.1-1.3.

First Generation: Over-Representation Analysis (ORA) Approaches

The immediate need for functional analysis of microarray gene expression data and the emergence of gene ontology (GO) during that period gave rise to over-representation analysis (ORA), which statistically evaluates the fraction of genes in a particular pathway found among the set of genes showing changes in expression. It is also referred to as “2×2 table method” in the literature [35]. ORA uses one or more variations of the following strategy [36]–[43] (Figure 1.2): first, an input list is created using a certain threshold or criteria. For example, a researcher may choose genes that are differentially over- or under-expressed in a given condition at a false discovery rate (FDR) of 5%. Then, for each pathway, input genes that are part of the pathway are counted. This process is repeated for an appropriate background list of genes (e.g., all genes measured on a microarray). Next, every pathway is tested for over- or under-representation in the list of input genes. The most commonly used tests are based on the hypergeometric, chi-square, or binomial distribution. We refer the readers to recent comparisons of ORA tools for more details [44], [45]. Many of the ORA tools differ very slightly from each other as they use the same statistical tests as well as overlapping pathway databases

Limitations: Over-Representation Analysis (ORA) Approaches

Despite the availability of a large number of tools and their widespread usage, ORA has a number of limitations. First, the different statistics used by ORA (e.g., hypergeometric distribution, binomial distribution, chi-square distribution, etc.) are independent of the measured changes. This means that these tests consider the number of genes alone and ignore any values associated with them such as probe intensities. By discarding this data, ORA treats each gene equally. However, the information about the extent of regulation (e.g., fold-changes, significance of a change, etc.) can be useful in assigning different weights to input genes, as well as to the pathways they are involved in, which in turn can provide more information than current ORA approaches. Second, ORA typically uses only the most significant genes and discards the others. For instance, the input list of genes from a microarray experiment is usually obtained using an arbitrary threshold (e.g., genes with fold-change and/or p-values). With this method, marginally less significant genes (e.g., fold-change = 1.999 or p-value = 0.051) are missed, resulting in information loss. Breitling et al. addressed this problem by proposing an ORA method for avoiding thresholds. It uses an iterative approach that adds one gene at a time to find a set of genes for which a pathway is most significant [46]. Third, by treating each gene equally, ORA assumes that each gene is independent of the other genes. However, biology is a complex web of interactions between gene products that constitute different pathways. One goal of gene expression analysis might be to gain insights into how interactions between gene products are manifested as changes in gene expression. A strategy that assumes the genes are independent is significantly limited in its ability to provide insights in this regard. Furthermore, assuming independence between genes amounts to “competitive null hypothesis” testing (see below), which ignores the correlation structure between genes. Consequently, the estimated significance of a pathway may be biased or incorrect. Fourth, ORA assumes that each pathway is independent of other pathways, which is erroneous. For instance, GO defines a biological process as a series of events accomplished by one or more ordered assemblies of molecular functions (http://www.geneontology.org/GO.doc.shtml). Another example of dependence between pathways is the cell cycle pathway in KEGG (http://www.genome.jp/kegg/pathway/hsa/hsa04110.html), where the presence of a growth factor activates the MAPK signaling pathway. This, in turn, activates the cell cycle pathway. No ORA methods account for this dependence between molecular functions in GO and signaling pathways in KEGG.

Second Generation: Functional Class Scoring (FCS) Approaches

The hypothesis of functional class scoring (FCS) is that although large changes in individual genes can have significant effects on pathways, weaker but coordinated changes in sets of functionally related genes (i.e., pathways) can also have significant effects. With few exceptions [47]–[49], all FCS methods use a variation of a general framework that consists of the following three steps [11] (Figure 1.2): first, a gene-level statistic is computed using the molecular measurements from an experiment. This involves computing differential expression of individual genes or proteins. Statistics currently used at gene-level include correlation of molecular measurements with phenotype [50], ANOVA [51], Q-statistic [47], signal-to-noise ratio [52], t-test [51], [23], and Z-score [53]. Although the choice of a gene-level statistic has a negligible effect on the identification of significantly enriched gene sets [11], when there are few biological replicates, a regularized statistic may be better. Furthermore, untransformed gene-level statistics can fail to identify pathways with up- and down-regulated genes. In this case, transformation of gene-level statistics (e.g., absolute values, squared values, ranks, etc.) is preferable [11], [26]. Second, the gene-level statistics for all genes in a pathway are aggregated into a single pathway-level statistic. This statistic can be multivariate [49], [54]–[57] and account for interdependencies among genes, or it can be univariate [23], [26] and disregard interdependencies among genes. The pathway-level statistics used by current approaches include the Kolmogorov-Smirnov statistic [52], [58], sum, mean, or median of gene-level statistic [26], the Wilcoxon rank sum [59], and the maxmean statistic [60]. Irrespective of its type, the power of a pathway-level statistic can depend on the proportion of differentially expressed genes in a pathway, the size of the pathway, and the amount of correlation between genes in the pathway. Interestingly, although multivariate statistics are expected to have higher statistical power, univariate statistics show more power at stringent cutoffs when applied to real biological data, and equal power as multivariate statistics at less stringent cutoffs [61]. The final step in FCS is assessing the statistical significance of the pathway-level statistic. When computing statistical significance, the null hypothesis tested by current pathway analysis approaches can be broadly divided into two categories: i) competitive null hypothesis and ii) self-contained null hypothesis [11], [23], [35], [60]. A self-contained null hypothesis permutes class labels (i.e., phenotypes) for each sample and compares the set of genes in a given pathway with itself, while ignoring the genes that are not in the pathway. On the other hand, a competitive null hypothesis permutes gene labels for each pathway, and compares the set of genes in the pathway with a set of genes that are not in the pathway. FCS methods address three limitations of ORA. First, they do not require an arbitrary threshold for dividing expression data into significant and non-significant pools. Rather, FCS methods use all available molecular measurements for pathway analysis. Second, while ORA completely ignores molecular measurements when identifying significant pathways, FCS methods use this information in order to detect coordinated changes in the expression of genes in the same pathway. Finally, by considering the coordinated changes in gene expression, FCS methods account for dependence between genes in a pathway, which ORA does not.

Limitations: Functional Class Scoring (FCS) Approaches

Although FCS is an improvement over ORA [23], [50], it also has several limitations. First, similar to ORA, FCS analyzes each pathway independently. This is a limitation because a gene can function in more than one pathway, meaning that pathways can cross and overlap. Consequently, in an experiment, while one pathway may be affected in an experiment, one may observe other pathways being significantly affected due to the set of overlapping genes. Such a phenomenon is very common when using the GO terms to define pathways due to the hierarchical nature of the GO. Second, many FCS methods use changes in gene expression to rank genes in a given pathway, and discard the changes from further analysis. For instance, assume that two genes in a pathway, A and B, are changing by 2-fold and 20-fold, respectively. As long as they both have the same respective ranks in comparison with other genes in the pathway, most FCS methods will treat them equally, although the gene with the higher fold-change should probably get more weight. Importantly, however, considering only the ranks of genes is also advantageous, as it is more robust to outliers. A notable exception to this scenario is approaches that use gene-level statistics (e.g., t-statistic) to compute pathway-level scores. For example, an FCS method that computes a pathway-level statistic as a sum or mean of the gene-level statistic accounts for a relative difference in measurements.

Third Generation: Pathway Topology (PT)-Based Approaches

A large number of publicly available pathway knowledge bases provide information beyond simple lists of genes for each pathway. Unlike GO and the Molecular Signatures Database (MSigDB), these knowledge bases also provide information about gene products that interact with each other in a given pathway, how they interact (e.g., activation, inhibition, etc.), and where they interact (e.g., cytoplasm, nucleus, etc.). These knowledge bases include KEGG [61], MetaCyc [62], Reactome [63], RegulonDB [64], STKE (http://stke.sciencemag.org), BioCarta (http://www.biocarta.com), and PantherDB [65]. ORA and FCS methods consider only the number of genes in a pathway or gene coexpression to identify significant pathways, and ignore the additional information available from these knowledge bases. Hence, even if the pathways are completely redrawn with new links between the genes, as long as they contain the same set of genes, ORA and FCS will produce the same results. Pathway topology (PT)-based methods have been developed to utilize the additional information. PT-based methods are essentially the same as FCS methods in that they perform the same three steps as FCS methods. The key difference between the two is the use of pathway topology to compute gene-level statistics. Rahnenfuhrer et al. proposed ScorePAGE, which computes similarity between each pair of genes in a pathway (e.g., correlation, covariance, etc.) [29]. The similarity measurement between each pair of genes is analogous to gene-level statistics in FCS methods, which is averaged to compute a pathway-level score. However, instead of giving equal weight to all pairwise similarities, ScorePAGE divides the pairwise similarities by the number of reactions needed to connect two genes in a given pathway (Figure 1.2). Although the approach is designed to analyze metabolic pathways, it is theoretically also applicable to signaling pathways. A recent impact factor (IF) analytic approach was proposed to analyze signaling pathways. IF considers the structure and dynamics of an entire pathway by incorporating a number of important biological factors, including changes in gene expression, types of interactions, and the positions of genes in a pathway [66], [67] (Figure 1.2). Briefly, IF analysis models a signaling pathway as a graph, where nodes represent genes and edges represent interactions between them. Further, it defines a gene-level statistic, called perturbation factor (PF) of a gene, as a sum of its measured change in expression and a linear function of the perturbation factors of all genes in a pathway (see Equation 1 in Appendix). Because the PF of each gene is defined by a linear equation, the entire pathway is defined as a linear system. Representing a pathway as a linear system also addresses loops in the pathways [67]. The IF of a pathway (pathway-level statistic) is defined as a sum of PF of all genes in a pathway (see Equation 2 in Appendix). IF analysis was recently improved to address the dominating effect of change in expression on PF and high false positive rate for a small list of input genes [28]. FCS methods that use correlations among genes [50], [68] implicitly assume that the underlying network, as defined by the correlation structure, does not change as the experimental conditions change. However, this assumption may be inaccurate. For example, the correlation structure between ARG2 and other genes in the urea-cycle pathway changes with a change in expression of ARG2 [69], suggesting changes in the topology of the pathway. Shojaie et al. proposed a method, called NetGSA, that accounts for the the change in correlation as well as the change in network structure as experimental conditions change [70]. Their approach, like IF analysis, models gene expression as a linear function of other genes in the network. However, it differs from IF in two aspects. First, it accounts for a gene's baseline expression by representing it as a latent variable in the model. Second, it requires that the pathways be represented as directed acyclic graphs (DAGs). If a pathway contains cycles, NetGSA requires additional latent variables affecting the nodes in the cycle. In contrast, IF analysis does not impose any constraint on the structure of a pathway [67].

Limitations: Third Generation: Pathway Topology (PT)-Based Approaches

Although PT-based methods are difficult to generalize, they have several common limitations. One obvious problem is that true pathway topology is dependent on the type of cell due to cell-specific gene expression profiles and condition being studied. However, this information is rarely available and is fragmented in knowledge bases, even if it is fully understood [71]. As annotations improve, these approaches are expected to become more useful. Other limitations of PT-based methods include the inability to model dynamic states of a system and the inability to consider interactions between pathways due to weak inter-pathway links to account for interdependence between pathways.

1.1.4 Challenges in pathway analysis

Khatri et al. introduces two major challenges in pathway analysis: annotation challenge and methodological challenge [1]. As pathway is a repository of knowledge that is acquired by fundamental evidence from experiments, the boundary is limited to the current understanding of Biology.

The annotation challenges can further be categorized to several challenges such as low resolution knowledge bases, incomplete and inaccurate annotations, missing condition- and cell-specific information and Inability to model and analyze dynamic response. Choosing the right pathway databases, adding manual curation for the specific biological problem that pathway analysis be applied will provide more accurate results.

The methodological challenge can further be classified to three challenges: benchmark data sets for comparing different methods, inability to model and analyze dynamic response and inability to model effects of an external stimuli. Especially, comparing different methods is important for new method development. However, the problem of comparing different methods for pathway analysis is made difficult by the lack of a gold standard. One can mine the literature to obtain evidence on whether a gene set is associated with the phenotype change, but this can only be done on a small scale. An alternative is suggested by Jui-Hung et al. to quantify the number of overlapping predictions [2].

However, it is arguable that pathways that have commonly discovered by several methods can always be a gold standard. A sensitive method might have found the true biological knowledge while others have failed. Thus, we suggest to not to consider overlapping pathways from multiple methods as a gold standard. Rather than that, we suggest to collect as many as data set that can be used for independent validation set for your study. In this paper, we follow this philosophy to demonstrate the performance of our method.

Another important challenge, although it is not mentioned by either of review papers [1,2], is personalized pathway analysis. Most pathway analysis methods reviewed so far is for discovering aberrant pathway between two phenotype groups. Methods can hardly provide the molecular aberrance of a single sample in terms of pathway. Issues about personalized pathway analysis will be introduced in the next section.

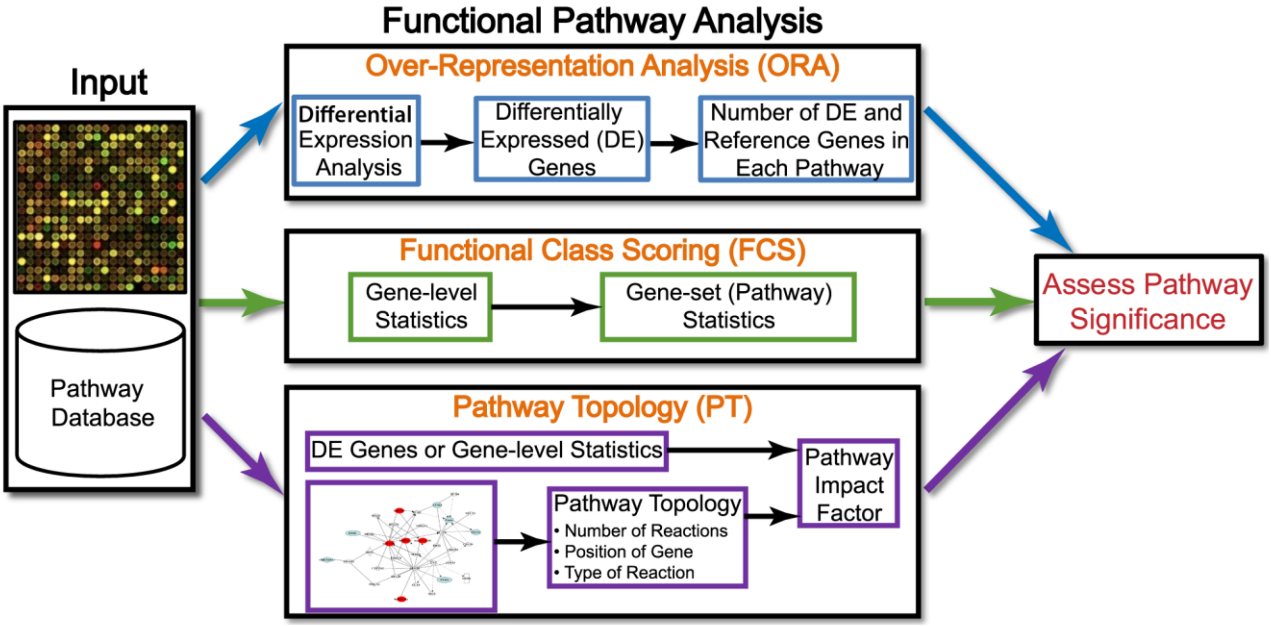
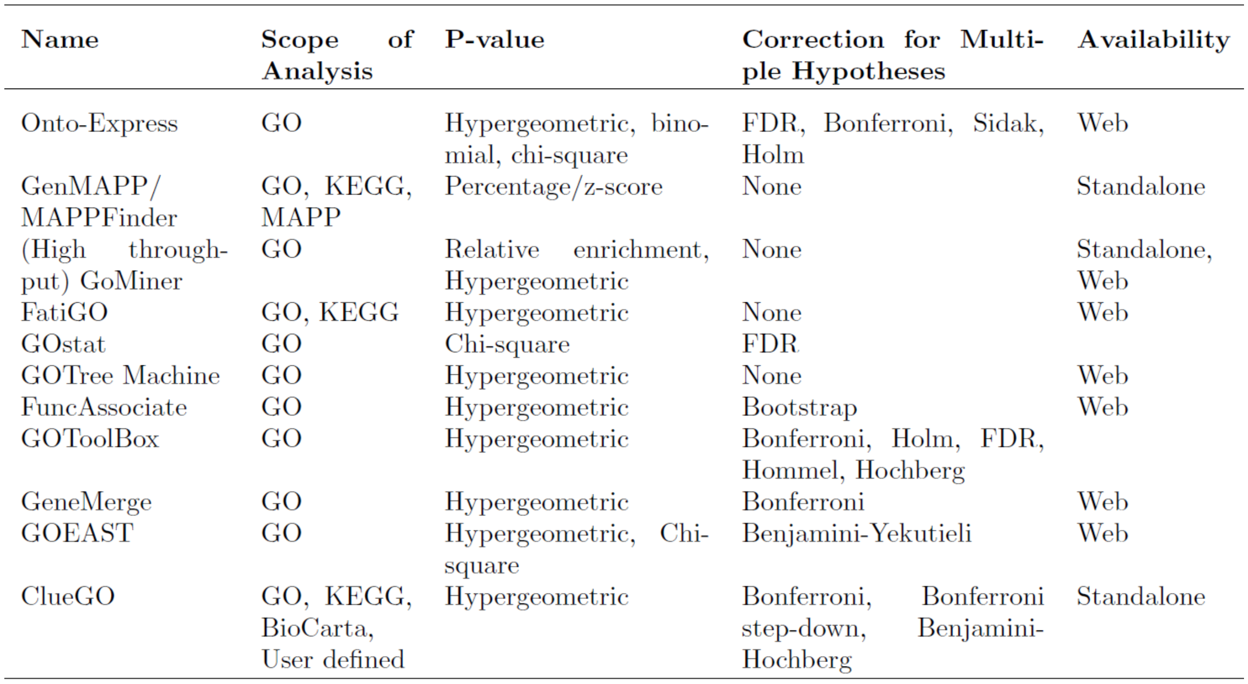

Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Pathway Impact Factor
Assess Pathway Significance

Figure 1.2 Overview of existing pathway analysis methods using gene expression

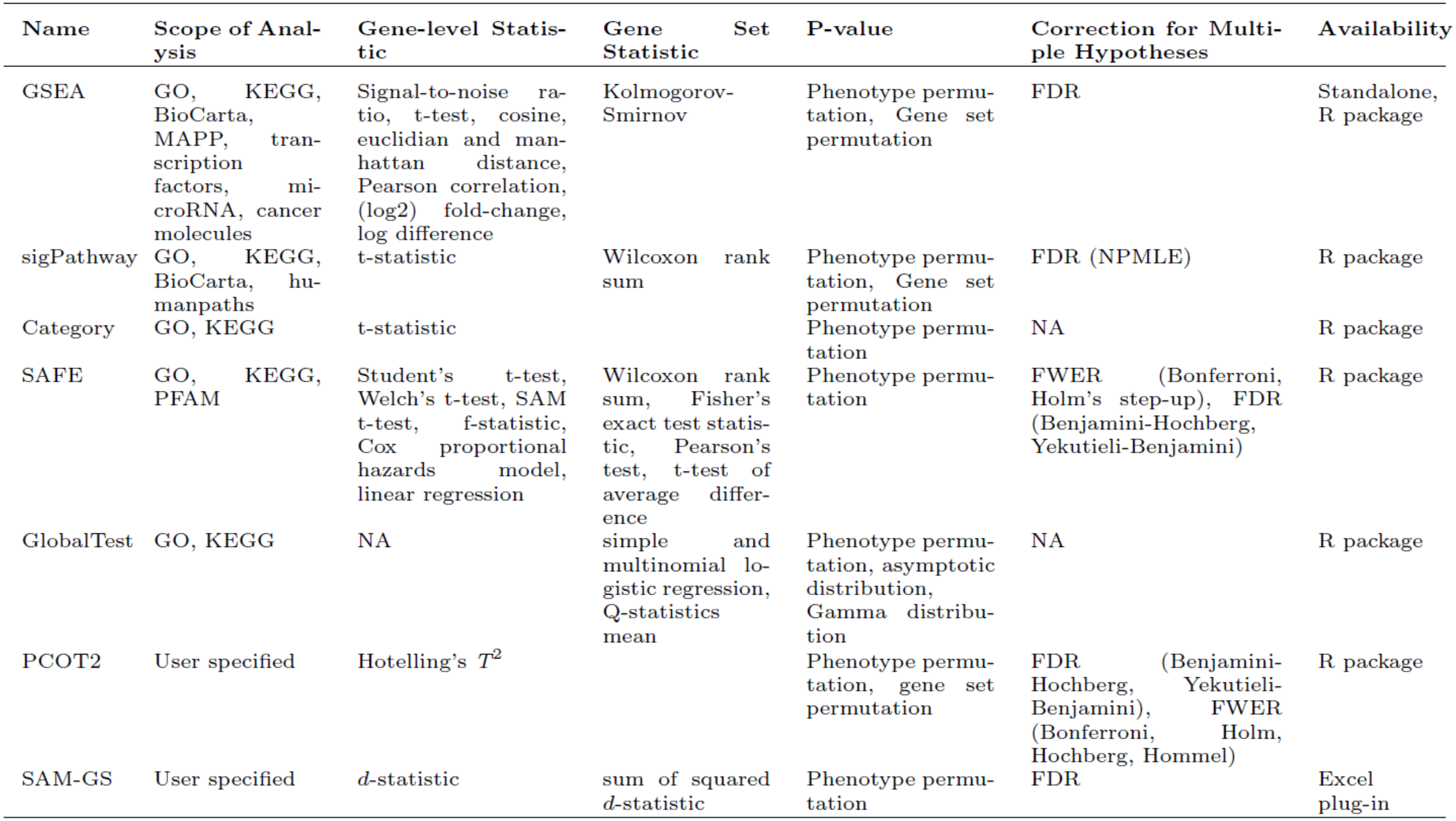
Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol 8(2): e1002375. doi:10.1371/journal.pcbi.1002375

Table 1.1 List of ORA pathway analysis tools



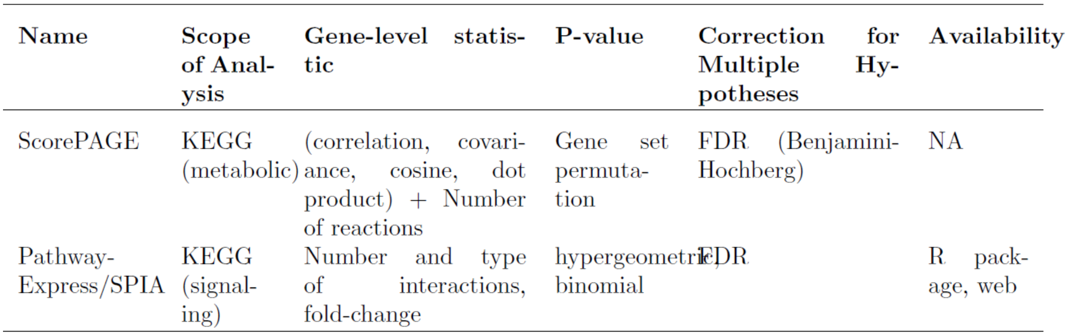
Name	Scope of Analysis	P-value	Correction for Multiple Hypotheses	Availability
Onto-Express	GO	Hypergeometric, binomial, chi-square	FDR, Bonferroni, Sidak, Holm	Web
GenMAPP/ MAPPFinder	GO, KEGG, MAPP	Percentage/z-score	None	Standalone
(High throughput) GoMiner	GO	Relative enrichment, Hypergeometric	None	Standalone, Web
FatiGO	GO, KEGG	Hypergeometric	None	Web
GOstat	GO	Chi-square	FDR	
GOTree Machine	GO	Hypergeometric	None	Web
FuncAssociate	GO	Hypergeometric	Bootstrap	Web
GOToolBox	GO	Hypergeometric	Bonferroni, Holm, FDR, Hommel, Hochberg	
GeneMerge	GO	Hypergeometric	Bonferroni	Web
GOEAST	GO	Hypergeometric, Chi-square	Benjamini-Yekutieli	Web
ClueGO	GO, KEGG, BioCarta, User defined	Hypergeometric	Bonferroni, Bonferroni step-down, Benjamini-Hochberg	Standalone

Table 1.1 List of FCS pathway analysis tools



Name	Scope of Analysis	Gene-level Statistic	Gene Set Statistic	P-value	Correction for Multiple Hypotheses	Availability
GSEA	GO, KEGG, BioCarta, MAPP, transcription factors, microRNA, cancer molecules	Signal-to-noise ratio, t-test, cosine, euclidian and manhattan distance, Pearson correlation, (log2) fold-change, log difference	Kolmogorov-Smirnov	Phenotype permutation, Gene set permutation	FDR	Standalone, R package
sigPathway	GO, KEGG, BioCarta, humanpaths	t-statistic	Wilcoxon rank sum	Phenotype permutation, Gene set permutation	FDR (NPMLE)	R package
Category	GO, KEGG	t-statistic		Phenotype permutation	NA	R package
SAFE	GO, KEGG, PFAM	Student's t-test, Welch's t-test, SAM t-test, f-statistic, Cox proportional hazards model, linear regression	Wilcoxon rank sum, Fisher's exact test statistic, Pearson's test, t-test of average difference	Phenotype permutation	FWER (Bonferroni, Holm's step-up), FDR (Benjamini-Hochberg, Yekutieli-Benjamini)	R package
GlobalTest	GO, KEGG	NA	simple and multinomial logistic regression, Q-statistics mean	Phenotype permutation, asymptotic distribution, Gamma distribution	NA	R package
PCOT2	User specified	Hotelling's T^2		Phenotype permutation, gene set permutation	FDR (Benjamini-Hochberg, Yekutieli-Benjamini), FWER (Bonferroni, Holm, Hochberg, Hommel)	R package
SAM-GS	User specified	*d*-statistic	sum of squared *d*-statistic	Phenotype permutation	FDR	Excel plug-in

Table 1.3 List of PT pathway analysis tools



Name	Scope of Analysis	Gene-level statistic	P-value	Correction for Multiple Hypotheses	Availability
ScorePAGE	KEGG (metabolic)	(correlation, covariance, cosine, dot product) + Number of reactions	Gene set permutation	FDR (Benjamini-Hochberg)	NA
Pathway-Express/SPIA	KEGG (signaling)	Number and type of interactions, fold-change	hypergeometric, binomial	FDR	R package, web

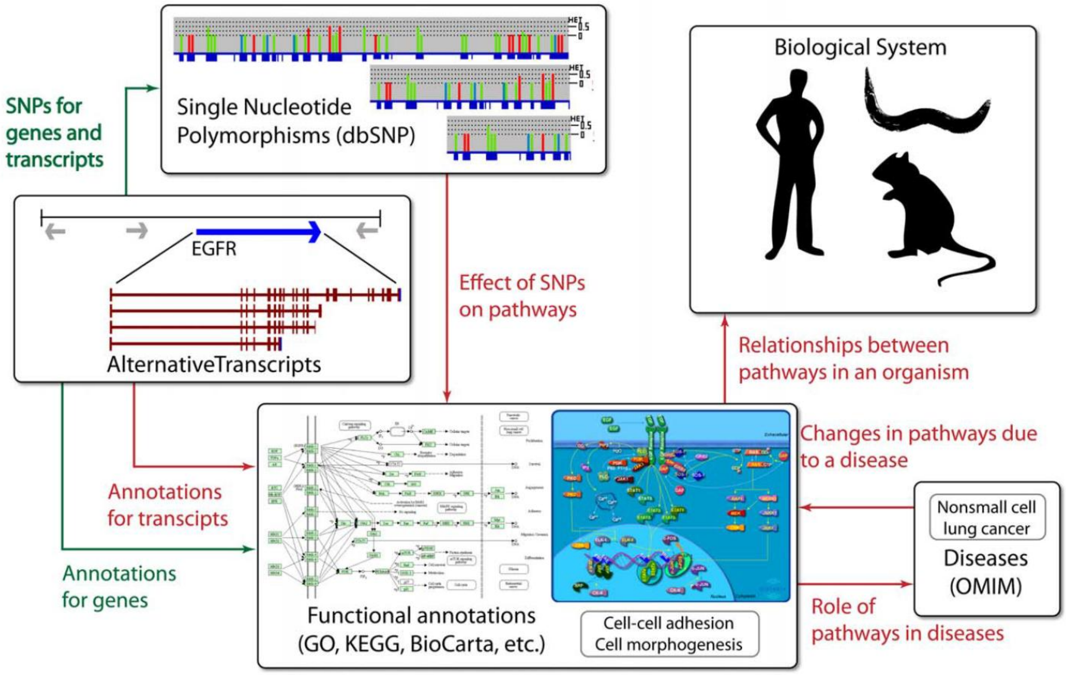

SNPs for genes and transcripts
Single Nucleotide Polymorphisms (dbSNP)
HET
0.5
0
EGFR
AlternativeTranscripts
Effect of SNPs on pathways
Biological System
Relationships between pathways in an organism
Changes in pathways due to a disease
Annotations for transcipts
Annotations for genes
Nonsmall cell lung cancer
Diseases (OMIM)
Role of pathways in diseases
Functional annotations (GO, KEGG, BioCarta, etc.)
Cell-cell adhesion
Cell morphogenesis

Figure 1.3 Overview of challenges in pathway analysis: low resolution, missing, and incomplete information. Green arrows represent abundantly available information, and red arrows represent missing and/or incomplete information. The ultimate goal of pathway analysis is to analyze a biological system as a large, single network. However, the links between smaller individual pathways are not yet well known. Furthermore, the effects of a SNP on a given pathway are also missing from current knowledge bases. While some pathways are known to be related to a few diseases, it is not clear whether the changes in pathways are the cause for those diseases or the downstream effects of the diseases.

Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol 8(2): e1002375. doi:10.1371/journal.pcbi.1002375

1.2 Personalized pathway analysis

Cancer arises from normal cells and can evolve to become malignant, metastatic, and/or resistant to therapy. The analysis of altered pathways in an individual cancer patient may help to understand the disease status and suggest customized anti-cancer therapies.

It is straightforward to compare the molecular profile of an individual's tumor and normal cells to discover molecular aberrances specific to his/her cancer. However, it may not be feasible in the current clinical practice environment to perform a metastatic tumor biopsy at the time of treatment resistance in patients with advanced cancer [1]. A case study of custom tailored med-icine based on an individual's genome and transcriptome highlights this limitation [2]. A patient's tumor had metastasized to the lung after surgery at the primary site. A biopsy from his lung tumor was analyzed by mutation and transcription profiling; however, the patient's normal lung tissue was not biopsied. Since there was no matched normal tissue, mRNA expression in the patient's own blood and information collected from various normal tissues were used to identify differentially expressed genes (DEGs). The results of pathway analysis based on DEGs, integrated copy number variation (CNV), and mutation information led the doctor to change the patient's drug treatment and the disease was stabilized for 3 months.

Although the personalized interpretation of pathways can be demanding, most current pathway analyses have been developed to investigate deregulated pathways between two phenotype groups. Khatri et al. [3] classified these methods into three types: Over-Representation Analysis (ORA), Functional Class Scoring (FCS), and a Pathway Topology based approach (PT).

ORA approaches typically apply an arbitrary threshold value (e.g., fold-change >2 or P-value <0.05) on gene expression to assess if the number of genes beyond threshold are significantly over or under represented in the given pathway. There are two drawbacks to ORA. First, it uses only the most significant genes and discards others, thus resulting in information loss for marginally significant genes [4]. Second, it considers only the number of genes and does not consider the magnitude of expression changes, leading to information loss regarding the importance of genes (e.g. a gene with a fold change of 2.01 and a gene with a fold change of 4 are considered equally). Unlike ORA, FCS methods do not discard genes with an arbitrary threshold but use all available genes, which is an improvement over ORA [5]. PT methods are essentially based on FCS methods with the addition that they consider network topology information. They compensate for the common limitation of ORA and FCS in reporting false positive gene sets due to sets of overlapping genes. In our paper, we focus on ORA and FCS methods, extending and implementing each for personalized pathway analysis.

There are two exceptional studies examining individualized pathway analysis [6],[7]. PARADIGM is a tool that infers a pathway status by using known functional structures. The method models the functional structure of pathway as a set of interconnected variables, where the variables are omic objects such as DNA, mRNA, and Protein where the interaction between variables describes the functional status of a pathway. PARADIGM may perform better with multiple omics as it utilizes known functional relationships between a gene or inter-gene DNA and protein. Hence, it might not perform well with single layer omic data, such as from mRNA microarrays.

Drier et al. (2013) [7] proposed a personal pathway deregulation score (PDS) which represents the distance of a single cancer sample from the median of normal samples on the principal curve. To calculate PDS, they reduced the dimensions by principal component analysis (PCA) and found the best principal curve, utilizing entire cohort samples containing both normal and different stages of cancers. Drier's method performs better than PARADIGM does in the mRNA only data sets of brain and colon cancers. However, calculating PDS requires manual data pre-processing steps, including selecting the number of principal components to be used and filtering out noisy gene data to obtain optimized principal curves. PDS requires information from whole cohort data to interpret an individual's pathway, which can be a limitation when applied to cohorts with small sample sizes, especially, when applied to cohorts that lack normal tissues controls.

Our proposed method is based on the comparison of one cancer sample to many accumulated normal samples (we use "nRef" to refer to the accumulated normal samples) that is different from the previous studies in following sense. The proposed method is suitable to adopt single layer omics data and expendable to interpret a patient in the context of many published or user defined pathway gene sets. PARADIGM has less freedom in terms of data and gene sets as it prefers multi-layered omics data and requires pre-defined functional structure among omics objects. Unlike PDS, which ex-tracts the principal curve from entire cohort data, our method does not assume an individual sample belongs to a cohort. This helps to avoid potential risk of dataset dependent analytic bias and also reduces efforts for data context dependent optimization process.

Our method provides a series of analysis steps, which consists of four parts: data processing, gene level statistics, individualized pathway aberrance scoring (iPAS), and a significance test. To dis-cover the most feasible method for iPAS, we extend existing path-way analysis techniques, namely ORA and FCS, to properly reflect the nature of testing one cancer to many normal samples.

To demonstrate that iPAS do indeed capture biologically and clinically relevant information in a sensible, valid, and useful manner, we apply it to samples of lung and colon adenocarcinoma. We show that our representation generates clinically relevant stratifications and outcome predictors, which would not have been achieved when the same data is analyzed by the conventional method that does not utilize accumulated normal data. We also applied our method to investigate functional impact mutations in the context of pathway.

.

Our empirical study suggests two different strategies, depending on the biological question that iPAS is focused on. In the case of cancer diagnosis, a method that utilizes the inter-gene correlation structure of the accumulated normal samples performs best. In the case of cancer prognosis, a simple averaging of all member genes' standardized gene expression values performs best.

1.3 Purpose and novelty of this study

The main purpose of this thesis is to quantify an individual cancer sample's pathway level aberrance. We followed biological insight that every cancer evolves from normal cell, inspiring comparison of a single cancer sample to many normal samples to quantify the molecular aberrance of single cancer case.

Our approach is unique in that our interpretation is not data set or cohort dependent, but reference dependent. In typical cases, cancer data is provided in the cohort study. Multiple cancer samples in the same cohort are usually normalized together and further be analyzed for differentially expressed gene analysis. In this procedure, a single cancer sample's expression level is going to be affected by other cancer samples in the same data set, thus it is data set dependent interpretation. In our analysis, we collect and provide multiple normal samples' data. A single cancer case is compared to many normals to quantify its pathway level aberrance. A cancer sample's expression level is going to be affected by normal samples' data, but it is not affected to other cancers from the same cohort where the sample of interest originated from. Thus, interpretation of our method can be considered as data set independent.

To prove that our approach is meaningful and have advantage against conventional approach that just use entire cohort for a single cancer case's interpretation, we provide a series of experiments. We have demonstrated that our method truly captures previous published biological knowledge, provide improved validation rate in the survival analysis than typical approach that utilize entire cohort data set. Additionally, our method is useful to capture the clinical important features in an unsupervised clustering of samples. The method is also able to capture commonly dysregulated pathways across all cancer samples which would not have been captured in the approach using cancer cohort data set. The commonly dysregulated pathways discovered by our method is able to classify cancer and normal, possibly to be useful for diagnosing cancer. Quantifying an individual sample’s pathway level aberrance is also useful to investigate mutational impact on the pathways.

An important clinical aspect of our methods is that it enables the interpretation of a cancer case in a single patient, even if matched normal tissue data from the same individual is unavailable. Accumulated information of normal tissues from a data repository will take the place of data unavailable for a specific individual. As the data repository is growing rapidly, it is expected that more normal tissue data will be available for many diseases in the near future.

We hope that our proposed approach can help in the personalized interpretation of tumor data and can be a useful tool in the upcoming era of data-based personalized medicine.

1.4 Outline of thesis

This thesis is organized as follows. Chapter 1 is an introduction of this study including the review of pathway analysis. Chapter 2 is introducing the approach of our study with materials used. Data sets used for this study is provided. Chapter 3 is a study of the effects of personalized pathway analysis. In this chapter, firstly, validation study of previously published survival related pathways are performed. This is to prove our approach is sensitively discovers known biological truth. Secondly, we compare our method to conventional approach in terms of validation rate of survival related pathways. Thirdly, we tested if unsupervised clustering of individualized pathway aberrance score can truly represent clinically important cancer characteristics. At fourth, we investigated that our approach is useful to diagnose cancer, namely pathway based identification of cancer. At fifth, we applied our personalized pathway analysis to discover mutations that have impact on the pathways. Conclusion and summary is presented at Chapter 4.

Chapter 2

Methods and materials

2.1 Gene expression data

I built nRef by the manual curation of data obtained from NCBI GEO [77],[78]. Microarray data of adjacent normal tissues obtained from patients undergoing surgery were selected to serve as the nRef. Data from biopsied samples, primary cultures of normal tissues and post-mortem donors were not included in the nRef. I collected a total of 120 nRef for lung, 60 from GSE19804 [79],[80], 27 from GSE7670 [81] and 33 from GSE10072 [82]. Samples came from individuals with variable smoking histories and different ethnic backgrounds. I collected 101 nRef' for colon, concentrating on normal mucosa tissue samples from six data sets available at GEO.

To evaluate the effectiveness of our method in survival analysis, we used Beer's data of 442 lung adenocarcinomas (LUAD) [83] to discover survival related pathways and validated the associations of 61 LUAD samples of GSE8894 [84]. The pathways identified in of LUAD were tested on 120 cancers and 120 normal samples of GSE19804, GSE7670, and GSE10071. Further validation was conducted with 48 cancers and 35 normal samples collected from GSE19188 [85] and GSE31547. For patient stratification by colon cancer differentiation status, we used 566 microarrays of GSE39582 [86], which provided in a separate manner, 443 for discovery, 123 for validation. GSE17536 [87] was also used for validation. The gene expression data set used is summarized in Table 2.1 and 2.2.

Table 2.1 Data set for pathway based survival analysis

Cancer	Purpose	GEO ID	# Samples	Platform
Lung adenocarcinoma	nRef	GSE19804	60	Affymetrix U133 Plus 2.0
		GSE10072	33	Affymetrix U133A
		GSE7670	27	Affymetrix U133A
	discovery*	Beer et al	442	Affymetrix U133A
	validation*	GSE8894	61	Affymetrix U133 Plus 2.0
	validation*	GSE3141	58	Affymetrix U133 Plus 2.0
Colon	nRef	GSE37364	38	Affymetrix U133 Plus 2.0
		GSE8671	32	Affymetrix U133 Plus 2.0
		GSE9348	12	Affymetrix U133 Plus 2.0
		GSE4107	10	Affymetrix U133 Plus 2.0
		GSE22619	8	Affymetrix U133 Plus 2.0
		GSE22242	1	Affymetrix U133 Plus 2.0
	discovery*	GSE39582	443	Affymetrix U133 Plus 2.0
	validation*	GSE39582	123	Affymetrix U133 Plus 2.0
	validation*	GSE17536	177	Affymetrix U133 Plus 2.0

Table 2.2 Data set for pathway based identification of cancer

Cancer	Purpose	GEO ID	# Samples	Platform
Lung adenocarcinoma	discovery**	GSE19804	60/60	Affymetrix U133 Plus 2.0
		GSE10072	33/33	Affymetrix U133A
		GSE7670	27/27	Affymetrix U133A
	validation**	GSE19188	18/15	Affymetrix U133 Plus 2.0
		GSE31547	28/20	Affymetrix U133 Plus 2.0

2.2 Pathway data

Information from gene sets representing biological pathways were obtained from REACTOME [88] which are also provided in the Molecular Signature Database [4]. I filtered out pathways where the number of genes was greater than 97. This criterion is arbitrary and was decided upon to avoid potential effects that were dependent on large gene set size. Out of 674 pathways that were originally from REACTOME, 583 pathways remained after filtering by the size of gene the set.

2.3 Individualized pathway analysis using the nRef

The aim of our approach is to identify altered pathways in an individual by making use of the nRef. A schematic diagram of our method of individualized pathway analysis is described at Figure 2.1 and the following sections describe each step.

2.3.1 Data preprocessing & Gene level statistics

Expression level was defined by using the robust multichip average [4]. For datasets using different microarrays, only those with probes in common from Affymetrix U133A to Affymetrix U133Plus 2.0 were used for further analysis. For individual tumor cases, we performed quantile normalization [89] after combining the single tumor microarray with all nRef samples. In cases of genes with multiple probes, gene expression level was summarized by averaging probe-level expression.

Individual tumor sample gene expression was stan-dardized using the mean and standard deviation of the reference.

2.3.2 Pathway level statistics & Significance test

I introduce five methods as candidates for iPAS. Each method is our modification of existing pathway analysis techniques, enabling us to test an individual tumor sample's pathway aberrance by using the nRef. A summary is provided in table 2.3.

Average Z

Standardizing the gene expression by mean and standard deviation from data sets is often used in microarray analysis. A vector $Z = (z_1, z_2, \ldots, z_n)$ denotes the expression status of a pathway where zi symbolize the standardized expression value of ith gene, where the number of genes belonging to the pathway is n. In typical settings, standardization is performed using the mean and standard deviation (s.d.) of a given data set, mostly the cancer only cohort data, thus $|Z| /n$ indicates how much the given sample's overall pathway gene expression deviates from the center of the cancer samples. I made the simple modification, $Z' = (z_1', z_2', \ldots, z_n)$, where z_i' is derived from mean and s.d. of the nRef. In this case, $|Z'| /n$ gives the samples deviation from the nRef. I believe this modification is biologically valid because every cancer starts its malignancy from normal cell. Thus, the clinical characteristics of a single cancer can be captured by measuring the difference of it against common characteristic of normal cells, which is represented by the nRef in our study.

Fisher exact test

I generated a 2 x 2 contingency table for a given pathway (S) and differentially expressed genes (D) for the test. For individualized interpretation, we define D by the ranking of z-value, which is standardized gene expression for the mean and s.d. of the nRef. For each individual sample, 5% (highest 2.5% and lowest 2.5%) of the total genes are defined as D. I applied a two-tailed test to detect alteration of pathways due to enrichment or depletion of differential genes. The result of this statistic can be interpreted as how many differentially expressed genes are enriched in the given pathway, where the expression difference refers to how much a patient's gene expression deviates from the nRef.

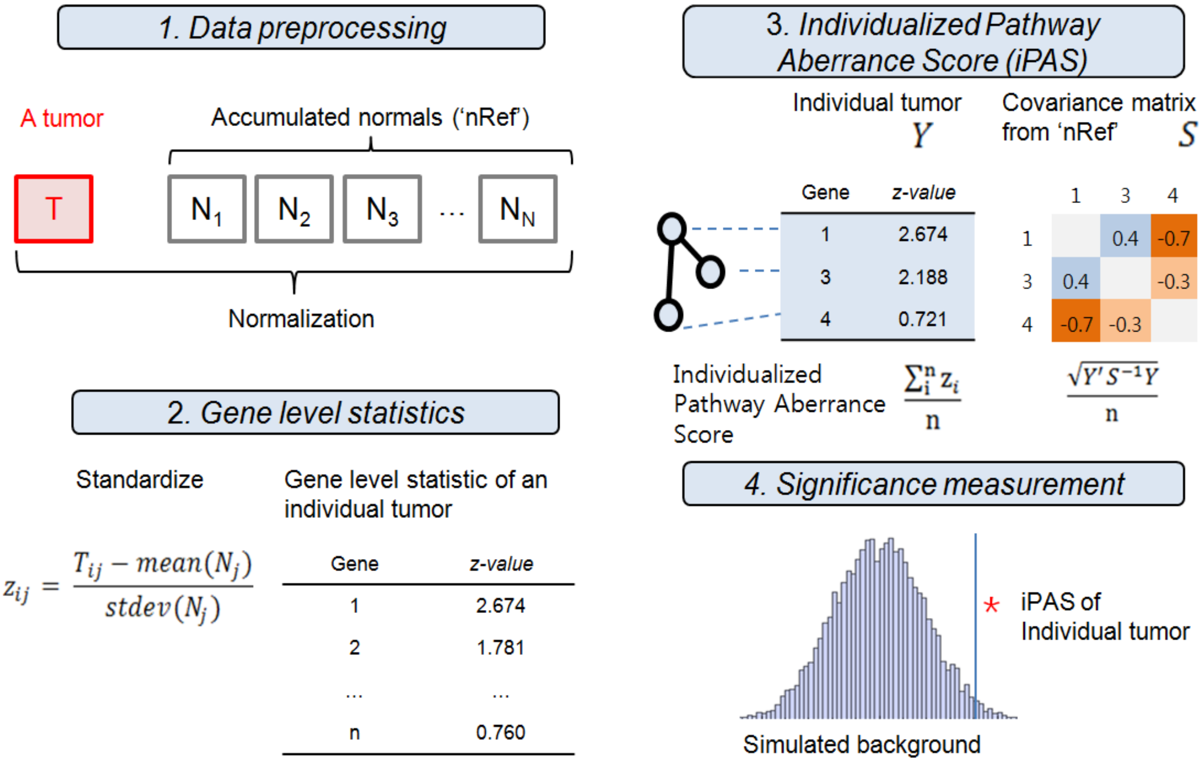
Gene set enrichment analysis (GSEA)

I adopted the original version of GSEA proposed by Subramanian et al. [52]. Typically, inputs for GSEA are generated by testing whole cohort samples using phenotype label; t-statistic, correlation coefficients, and fold changes are usually used. In the personalized analysis setting, we use the z-value as an input for the GSEA algorithm, which is standardized gene expression by mean and s.d. of the nRef. The GSEA output Enrichment Score (ES) reflects the degree to which a gene set in the pathway is overrepresented at the extremes (low or high) of the entire ranked list of z-values from a single patient.

Non parametric quadratic test

Gene expression in a pathway of a tumor sample is represented by vector $Z = (z_1, z_2, \ldots, z_n)$, where z_i standardized expression level of ith gene by mean and s.d. of the nRef, where n is the number of genes belonged to the pathway. A pathway characteristic of an individual patient's pathway can be represented by the averaged Euclidean distance ($Z^T Z/n$). This gives the distance of a single patient from the center of the nRef due to the square of standardized expression difference, and thus does not reflect increased or decreased expression, only the extent of the expression difference. Genes in the pathway are usually functionally correlated, therefore use of the correlation structure of the normal samples may increase sensitivity enough to capture the aberrance of a single cancer case. I also consider the averaged Mahalanobis (Z^TSZ/n) distance, that utilizes the covariance matrix calculated from the nRef. This value describes the statistical distance from the center of normal samples taking into account correlation among normal samples. The covariance matrix S is calculated for each pathway from the nRef.

Significance can be obtained against the null distribution generated from normal samples. All the collected normal samples for the nRef are one by one compared to the nRef to yield statistics of the null distribution. A statistic from a single cancer case is compared to this null distribution to yield P-value.



1. Data preprocessing
A tumor
Accumulated normals ('nRef')
T
N_1
N_2
N_3
…
N_N
Normalization
2. Gene level statistics
Standardize
Gene level statistic of an individual tumor
$z_{ij} = \frac{T_{ij} - mean(N_j)}{stdev(N_j)}$
3. Individualized Pathway Aberrance Score (iPAS)
Individual tumor Y
Covariance matrix from 'nRef' S
Individualized Pathway Aberrance Score
$\frac{\sum_i^n z_i}{n}$
$\frac{\sqrt{Y'S^{-1}Y}}{n}$
4. Significance measurement
iPAS of Individual tumor
Simulated background

Gene	*z-value*
1	2.674
2	1.781
…	…
n	0.760

Gene	*z-value*
1	2.674
3	2.188
4	0.721

	1	3	4
1		0.4	-0.7
3	0.4		-0.3
4	-0.7	-0.3	

Figure 2.1 Schematic description of individualized pathway analysis using accumulated normal data (nRef). An individual tumor data is normalized with the nRef. Gene expression is standardized by mean and standard deviation of the nRef. iPAS is calculated from standardized gene expression values in the pathway. Null distribution calculated from the nRef provides significance.

Method	Gene statistics	Pathway statistics
Average Z	$z_i = \frac{g_{\mathrm{T}i} - mean(g_{\mathrm{nRef}})}{stdev(g_{\mathrm{nRef}})}$	$\frac{\sum_{\mathrm{i}}^{\mathrm{n}} \mathrm{z}_i}{\mathrm{n}}$
Fisher	$z_i = \frac{g_{\mathrm{T}i} - mean(g_{\mathrm{nRef}})}{stdev(g_{\mathrm{nRef}})}$ DEG: top 2.5% bottom 2.5%	$\sum_{\mathrm{i=k}}^{\mathrm{min}\ (\lvert\mathrm{DEG}\rvert,\mathrm{n})} \frac{\binom{\lvert\mathrm{DEG}\rvert}{\mathrm{k}}\binom{\mathrm{N}-\lvert\mathrm{DEG}\rvert}{\mathrm{n-k}}}{\binom{\mathrm{N}}{\mathrm{n}}}$
GSEA	$z_i = \frac{g_{\mathrm{T}i} - mean(g_{\mathrm{nRef}})}{stdev(g_{\mathrm{nRef}})}$	$\max(P_{\mathrm{hit}} - P_{\mathrm{miss}})$ $P_{\mathrm{hit}}(P,i) = \sum_{\substack{g_i \in P \\ j \leq i}} \frac{\lvert z_i \rvert^x}{N_R}$ $P_{\mathrm{miss}}(P,i) = \sum_{\substack{g_i \notin P \\ j \leq i}} \frac{1}{(N - N_H)}$ $N_R = \sum_{g_i \in P} \lvert z_i \rvert^x$
Euclidean	$z_i = \frac{g_{\mathrm{T}i} - mean(g_{\mathrm{nRef}})}{stdev(g_{\mathrm{nRef}})}$	$\frac{\sqrt{\sum_{\mathrm{i}}^{\mathrm{n}} {z_i}^2}}{\mathrm{n}}$
Mahalanobis	$z_i = \frac{g_{\mathrm{T}i} - mean(g_{\mathrm{nRef}})}{stdev(g_{\mathrm{nRef}})}$	$\frac{\sqrt{Z'S^{-1}Z}}{\mathrm{n}}$

Table 2.3. Pathway statistics for individualized pathway aberrance score (iPAS)

2.4 Pathway based identification of rare mutation effect in cancer

To evaluate functional impact of mutations on pathway, somatic mutation (WUSM mutation calling) and normalized gene expression (UNC Agilent G4502A_07, level 3) data of breast cancer is downloaded from TCGA web site [91]. The level 3 TCGA mRNA data provides gene level summary of mRNA expression, which is standardized by mean and standard deviation of entire data set. Samples having both mutation and gene expression data (n=513) are used for analysis. Missing gene expressions are replaced by zeros.

Pathway gene sets are downloaded from molecular signature database. Total 543 gene sets from Biocarta, NCI cancer pathway [92] and KEGG pathway [93],[94] are used for our analysis. I also manually defined additional gene sets to assess rare mutations impact on pathways that include more than one drug target; we defined 16 receptor tyrosine kinase (RTK) genes which have more than one approved targeted drugs. Associated drug information is retrieved from Ingenuity SystemsTM [95]. For each of RTK genes, we expended the gene set by adding its first neighbors by adopting protein-protein interaction data in HipathDB [96]. Through this annotation process, we obtained gene sets representing 16 RTK pathways. Summary of RTK pathways used in our study is provided in table 2.4

Target	Drugs	First neighbors
EGFR	cetuximab, AEE 788, panitumumab, BMS-599626, ARRY-334543, XL647, canertinib, gefitinib, HKI-272, PD 153035, lapatinib, vandetanib, erlotinib	125
PDGRFB	dasatinib, sunitinib, pazopanib, axitinib, KRN-951, tandutinib, imatinib, sorafenib, becaplermin	61
ERBB2	trastuzumab, BMS-599626, ARRY-334543, XL647, CP-724,714, HKI-272, lapatinib, erlotinib	59
MET	crizotinib	55
ERBB4	BMS-599626	44
KIT	dasatinib, sunitinib, pazopanib, KRN-951, OSI-930, telatinib, tandutinib, imatinib, sorafenib	38
FLT4	sunitinib, pazopanib, CEP 7055, KRN-951, telatinib, sorafenib, vandetanib	36
PDGFRA	sunitinib, pazopanib, axitinib, telatinib, imatinib, becaplermin	35
TEK	vandetanib	35
RET	sunitinib, vandetanib	30
FGFR1	pazopanib	29
EPHA2	dasatinib	22
FGFR3	pazopanib	18
FLT3	CHIR-258, tandutinib, sorafenib, lestaurtinib, CGP 41251	14
FGFR2	palifermin	13

Table 2.4. Curated drug target centric pathways

In our analysis, multiple rare mutations on the same gene are defined as a single mutation event. Our algorithm iteratively adds genes to mutation event while the mutation event makes significant difference between mutation having and non-having sample group. Final output is a set of genes. The output can be interpreted as a set of rare mutations from a(several) gene(s) that have significant pathway level impact. Pseudo code description of our procedure is as follows.

```
In the given pathway, define 'gene list'
gene list : gene names in the given pathway

for each gene in the 'gene list'
while ( no more genes in the 'gene list' or
                no more significant mutational event found ){

    update : gene(s) name(s) to be considered as
             mutational event
    x: a vector containing pathway statistics of mutation event having
             patients
    y: a vector containing pathway statistics of mutation event non having
             group

    do t-test on x and y
    if( t-test p-value < fdr_threshold){
        add one more gene for mutation event
    }
   update fdr_threshold
   remove the gene from the 'gene list'
}
```

Chapter 3

Results

3.1 Capturing published survival related pathways

To assess whether our method can sensitively detect pathway aberrances that are associated with a patient's clinical outcome, a known survival pathway that showed strong association with patient survival from Beer's data was tested. Bryant et al. [97] reported that the "cell cycle stimulatory" pathway of 51 genes is significantly associated with patient survival (Cox proportional-hazards model, P = 0.000113). In that study, pathway gene expression was represented as an average of z-values, where the z-value indicates the standardized expression level, by the mean and s.d., of all cancer samples. The high-risk group was defined as those in which pathway expressions were greater than zero and the pathway showed poor prognostic outcome. The association was significant with or without adjusted clinical covariates, and thus the pathway alone is a strong indicator of cancer prognosis. This finding was also validated in the Japanese LUAD cohort (n = 87, survival data is not provided to public) in Bryant's study. As studies have shown a clear association between the cell cycle pathway and cancer, in terms of driving cancer proliferation, we considered this pathway as a pathway that should be detected. All of the methods proposed as candidates for iPAS showed significant associations of the "cell cycle stimulatory" pathway from Beer's data (Table 3.1). The same pathway analyzed using GSE8894 (n = 61) data yielded significant associations in all proposed methods with the marginal exception of mahalanobis, (P = 0.0549).

Prognostic gene expression signatures for stage II and III colon cancers have been reported in seven papers in, yielding 207 genes in total [90],[98-103]. The genes are enriched in 32 REACTOME pathways (FDR < 0.05, pathway size < 96, Table 3.2). I assumed the 32 pathways were valid as ground truth to be identified and analyzed in the colon cancer dataset GSE39585 (stage II and III were only considered). Average Z provided best performer (sensitivity = 0.88) with 28 pathways deemed as significant. GSEA, Fisher, Euclidean, Mahalanobis gave the following values, 0.78, 0.66, 0.06, 0.03, respectively.

These results satisfied us that our approach captures the fundamental knowledge of cancer, thus it is reasonably considered as individualized pathway aberrance score.

Table 3.1. Survival analysis of “cell cycle” pathway reported by Bryant et al.

Data set	Pathway Statistics	Coef	*P* value
Beer (*N* = 432), Bryant et al., Overall survival	Average Z*	0.37	0.00011
Beer (*N* = 442) Overall survival	Average Z**	0.62	0.00003
	Fisher	0.50	0.00068
	GSEA	0.65	0.00001
	Euclidean	0.65	0.00001
	Mahalanobis	0.67	0.00001
GSE8894 (*N* = 61) Recurrent Free Survival	Average Z**	0.90	0.01163
	Fisher	0.91	0.01076
	GSEA	0.78	0.02899
	Euclidean	0.87	0.01544
	Mahalanobis	0.68	0.05485

Table 3.2. Enriched pathways of 207 colorectal cancer survival genes

Gene Set Name	# Genes in Gene Set (K)	# Genes inOverlap (k)	k/K	p-value	FDR q-value
REGULATION_OF_MITOTIC_CELL_CYCLE	85	5	0.0588	4.32E-05	4.16E-03
SYNTHESIS_OF_DNA	92	5	0.0543	6.32E-05	4.73E-03
REGULATION_OF_ORNITHINE_DECARBOXYLASE_ODC	49	4	0.0816	7.23E-05	4.87E-03
CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	57	4	0.0702	1.31E-04	7.32E-03
ER_PHAGOSOME_PATHWAY	61	4	0.0656	1.71E-04	8.22E-03
APC_C_CDH1_MEDIATED_DEGRADATION_OF_CDC20_AND_OTHER_APC_C_CDH1_TARGETED_PROTEINS_IN_LATE_MITOSIS_EARLY_G1	72	4	0.0556	3.24E-04	1.28E-02
APC_C_CDC20_MEDIATED_DEGRADATION_OF_MITOTIC_PROTEINS	73	4	0.0548	3.41E-04	1.28E-02
ANTIGEN_PROCESSING_CROSS_PRESENTATION	76	4	0.0526	3.98E-04	1.34E-02
M_G1_TRANSITION	81	4	0.0494	5.07E-04	1.63E-02
REGULATION_OF_MRNA_STABILITY_BY_PROTEINS_THAT_BIND_AU_RICH_ELEMENTS	84	4	0.0476	5.82E-04	1.78E-02
MITOTIC_PROMETAPHASE	87	4	0.046	6.64E-04	1.87E-02
CDK_MEDIATED_PHOSPHORYLATION_AND_REMOVAL_OF_CDC6	48	3	0.0625	1.34E-03	3.23E-02
CROSS_PRESENTATION_OF_SOLUBLE_EXOGENOUS_ANTIGENS_ENDOSOMES	48	3	0.0625	1.34E-03	3.23E-02
AUTODEGRADATION_OF_THE_E3_UBIQUITIN_LIGASE_COP1	51	3	0.0588	1.60E-03	3.26E-02
P53_INDEPENDENT_G1_S_DNA_DAMAGE_CHECKPOINT	51	3	0.0588	1.60E-03	3.26E-02
SCF_BETA_TRCP_MEDIATED_DEGRADATION_OF_EMI1	51	3	0.0588	1.60E-03	3.26E-02
HEPARAN_SULFATE_HEPARIN_HS_GAG_METABOLISM	52	3	0.0577	1.69E-03	3.26E-02
VIF_MEDIATED_DEGRADATION_OF_APOBEC3G	52	3	0.0577	1.69E-03	3.26E-02
REPAIR_SYNTHESIS_FOR_GAP_FILLING_BY_DNA_POL_IN_TC_NER	14	2	0.1429	1.77E-03	3.26E-02
DESTABILIZATION_OF_MRNA_BY_AUF1_HNRNP_D0	53	3	0.0566	1.79E-03	3.26E-02
CDT1_ASSOCIATION_WITH_THE_CDC6_ORC_ORIGIN_COMPLEX	56	3	0.0536	2.09E-03	3.53E-02
SCFSKP2_MEDIATED_DEGRADATION_OF_P27_P21	56	3	0.0536	2.09E-03	3.53E-02
P53_DEPENDENT_G1_DNA_DAMAGE_RESPONSE	57	3	0.0526	2.20E-03	3.62E-02
REGULATION_OF_APOPTOSIS	58	3	0.0517	2.32E-03	3.72E-02
ACTIVATION_OF_NF_KAPPAB_IN_B_CELLS	64	3	0.0469	3.07E-03	4.32E-02
AUTODEGRADATION_OF_CDH1_BY_CDH1_APC_C	64	3	0.0469	3.07E-03	4.32E-02
INTERFERON_ALPHA_BETA_SIGNALING	64	3	0.0469	3.07E-03	4.32E-02
ASSEMBLY_OF_THE_PRE_REPLICATIVE_COMPLEX	65	3	0.0462	3.20E-03	4.32E-02
CYCLIN_E_ASSOCIATED_EVENTS_DURING_G1_S_TRANSITION	65	3	0.0462	3.20E-03	4.32E-02
SIGNALING_BY_WNT	65	3	0.0462	3.20E-03	4.32E-02
ORC1_REMOVAL_FROM_CHROMATIN	67	3	0.0448	3.49E-03	4.61E-02
NEPHRIN_INTERACTIONS	20	2	0.1	3.64E-03	4.71E-02

3.2 Proposed approach vs conventional approach

To investigate which of the candidates for iPAS most robustly reflect phenotype association, we evaluated the proposed methods by determining whether survival associated pathways are validated in datasets never used for discovery using LUAD and colon cancer (LUAD: Beer's set n=442 for discovery, GSE8894(n=61) GSE3141(n=58) for validation, colon cancer: GSE39582d (n=443) for discovery, GSE39582v(n=123) and GSE17536(n=109) for validation, logrank P < 0.05, comparing tumors in the top 50th percentile of aberrance scores to those in the bottom 50th percentile). Validation rates varied depending on the dataset and these were possibly affected by the small sample size compared to that of the discovery set. In these cases, we were not able to determine a superior method that outperformed the others (Table 3.3, Figure 3.1, 3.2). Average Z gave the highest validation rate in three out of four dataset with validation rates of GSE8894 (43.6%, 92/211), GSE3141 (13.3%, 28/211) and GSE17536 (10.7%, 24/224). When validation rates from four datasets are averaged, Average Z gave the highest validation rate, (21.9%, Figure 3.3, blue bars). Pathways validated as significantly associated with patient survival for each cancer are listed in Table 3.4. I also investigated the validation rate of iPAS candidates under the conditions where the same data is not standardized by the nRef, but instead standardized by the mean and s.d. of the cohort dataset, which consists of only cancers (Figure 3.3, red bars). It is noteworthy that use of the nRef increased the validation rate for every iPAS candidate investigated. This implies that the strategy of using accumulated normal samples as a reference is beneficial in terms of pathway based survival analysis.

I further investigate the reason why the Ave Z generally performs better than the other methods in terms of validation rate. Thirty six survival related pathways were concordantly validated in two independent lung adenocarcinoma data sets. Visual inspections of the 36 pathways provide an insight about difference of its performance. Among 36 pathways, selected 5 pathways will be discussed. Each of 5 pathways is representing a pathway that is validated in two data sets by the method of Ave Z, GSEA, ED, MD and both of Ave Z and GSEA, respectively (Figure 3.5).

A pathway 'G1 S specific transcription' is well explaining survival difference in three data sets by the methods of Ave Z and GSEA (Figure 3.6). The common feature of Ave Z and GSEA vs others (ED, MD, Fisher) is that Ave Z and GSEA use the information of '+' or '-' of pathway statistics. This possibly explains in what circumstance if the methods perform better. If a true biology of a given pathways goes in one direction such as up-regulation of all member gene, or down-regulation of all member gene, it is more appropriate to represent pathway characteristics with direction. In case true biology has a clear direction in the pathway function, ED and MD which performs square of standardized gene expression value, Fisher which provide test statistic of enrichment, do not represent the direction, thus might perform poorly.

This is making sense in the specific case of 'G1 S specific transcription'. In this pathway, almost all member genes are supposed to be up-regulated due to the tumurogenic signal. Figure 3.7 describe this phenomenon in detail. There are three distinguishing sample groups, down-regulated (S1), up-regulated (S3) and intermediate (S2). Up-regulated sample subgroup (S3) has poor outcome due to the highly proliferating nature of this pathway. In both discovery data set and validation data set, methods using directionality Ave Z and GSEA are well representing the patient subgroup. However, in validation data set S3, other methods that do not utilize directionality information of gene expression perform poorly in prognostic prediction (Figure 3.6). This might be caused by different interpretations of genes which expressions are negatively regulated the way it usually work in cancer proliferation signal.

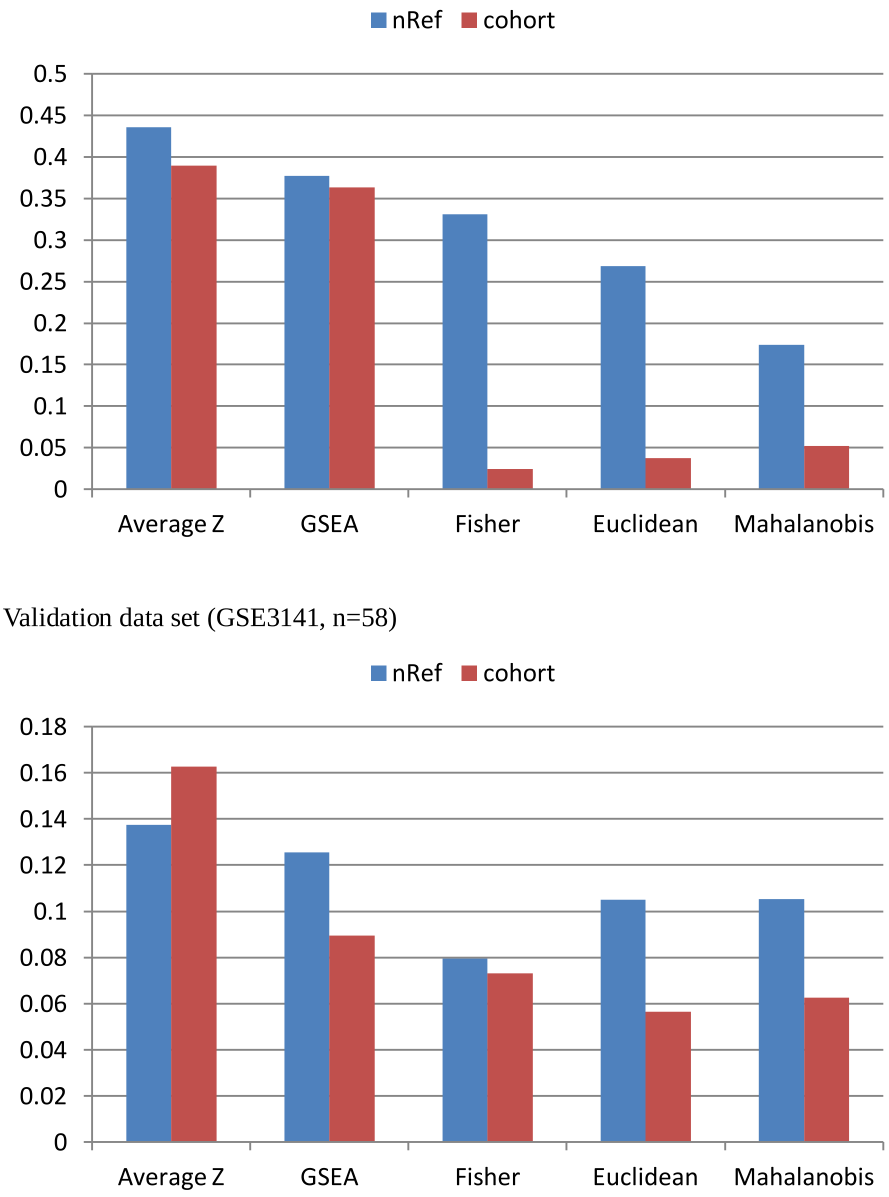
Yellow circle in Figure 3.7 describe this case. In case Ave Z and GSEA, genes in the yellow circle can be interpreted as negative factors. The genes reduce the overall activity of the pathway in the interpretation of Ave Z and GSEA, which is concordant to biological facts. However, in the interpretation of ED, MD and Fisher, the genes can be interpreted as positive factors, increasing overall extent of gene expression difference of the pathway, which is not concordant to biological facts. Figure 3.8 depicts the case GSEA performs better than the Ave Z in terms of validation. Both Ave Z and GSEA utilize gene expression directionality information, but the difference is considering the ranks pathway members genes in the whole genome context in GSEA analysis. Yellow circles at Figure 3.8 show discrepancy between Ave Z and GSEA.

Some pathways validated in ED specific manner. Figure 3.9 describes this case. The ‘Cyclin A B1 asssociated events during G2 M transition’ pathway is consists of 15 genes. In tumorigenic situation, the pathway is generally up-regulated in transcription level. Two genes CCNH and CCNA1 show down regulated pattern while the other 13 genes in the pathway generally up-regulated. Actually, it has known that low expression of CCNH is activating cyclin dependent kinase activity, thus it is biologically proper that low expression of CCNH is associated with poor outcome in patient survival. In this pathway, contribution of CCNH is explained by its down regulation. Thus, methods using directionality of gene expression like Ave Z, GSEA miss-interpretate the contribution of CCNH by simply adding the negative value into overall sum (Ave Z) or little contribution of enrichment score while majority genes are up-regulated (GSEA). Unlike Ave Z and GSEA, methods ED and MD only consider the extent of gene expression changes, thus under expression of CCNH is also positively contribute to overall extent of gene expression change when these methods are considered. However, in this pathway, only ED is validated in two different data sets, while the MD is not. This might be the reason why the methodological difference between the two methods.

One thing obvious is that not one method performs dominant against other method in all cases. Some pathway is consisted of up-regulated genes or down-regulated genes. Some pathway is consisted of mixture of up and down regulated genes with biological meaning. It is important to know the characteristic of pathway member genes and choose the best performing pathway statistics.

Discovery of survival related pathway (Beer's data, n=442)

Validation data set (GSE 8894, n=61)

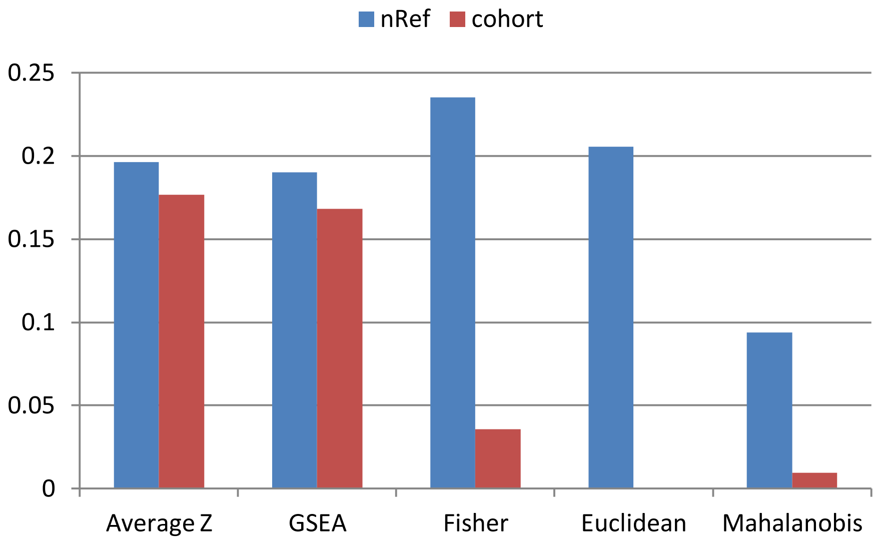


Validation data set (GSE3141, n=58)

Figure 3.1 Validation rate of discovered survival related pathways in lung adenocarcinoma. Proposed approach using nRef (blue) vs. Conventional approach. Conventional approach standardizes individual sample by mean and s.d of entire cohort.

Discovery of survival related pathway (GSE39582d, n=443)

Validation data set (GSE39582v, n=123)



Validation data set (GSE39582v, n=177)

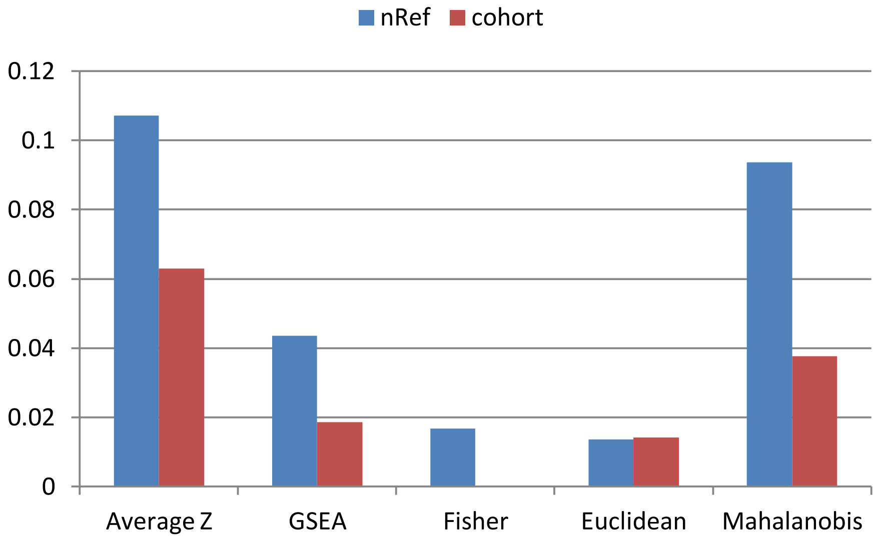


Figure 3.2 Validation rate of discovered survival related pathways in colon cancer. Proposed approach using nRef (blue) vs. Conventional approach. Conventional approach standardizes individual sample by mean and s.d of entire cohort.

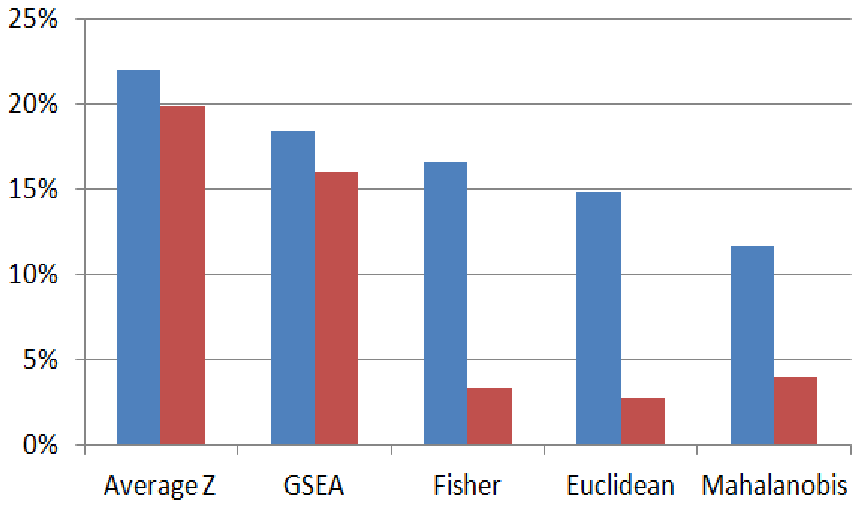


Figure 3.3 Averaged validation rate of discovered survival related pathway at four data sets. Proposed approach using nRef (blue) vs. Conventional approach that standardizes individual sample by mean and s.d. of entire cohort data set (red)

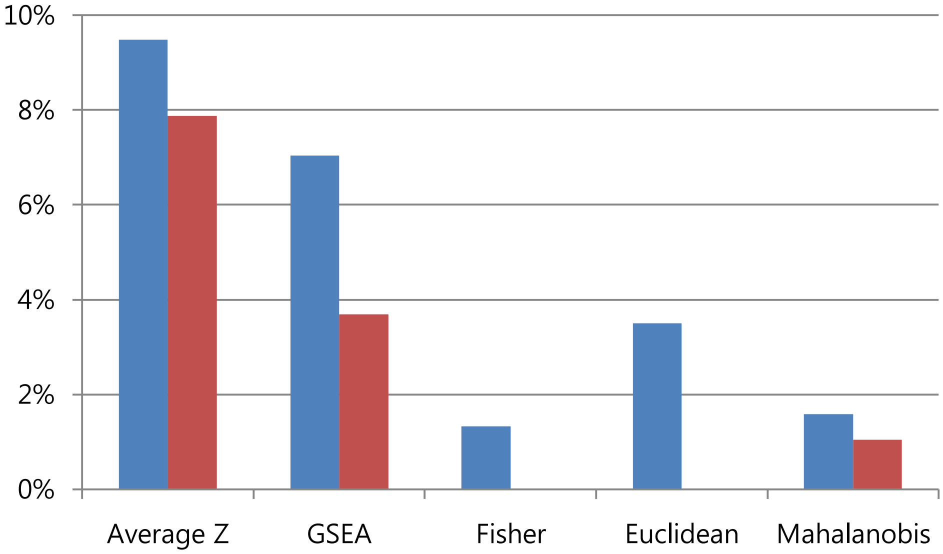


Figure 3.4 Validation rate of discovered survival related pathways (# of pathways commonly validated at two data sets / # of survival related pathways discovered at discovery data set)

		nRef-based approach				Cancer cohort data based approach*			
		Discovery	Validation			Discovery	Validation		
	Method	Beer	GSE8894 (A)	GSE3141 (B)	A ∩ B	Beer	GSE8894 (C)	GSE3141 (D)	C ∩ D
LUAD	Average Z	211	92	29	20	203	79	33	16
	GSEA	199	75	25	14	190	69	17	7
	Fisher	151	50	12	2	41	1	3	0
	Euclidean	257	69	27	9	53	2	3	0
	Mahalanobis	190	33	20	3	96	5	6	1
		dGSE39582	vGSE39582 (A)	GSE17536 (B)	A ∩ B	dGSE39582	vGSE39582 (C)	GSE17536 (D)	C ∩ D
Colon Cancer	Average Z	224	44	24	7	238	42	15	3
	GSEA	184	35	8	2	214	36	4	0
	Fisher	119	28	2	1	28	1	0	0
	Euclidean	73	15	1	0	71	0	1	0
	Mahalanobis	32	3	3	2	106	1	4	1

Table 3.3. Comparison of validation rates of survival related pathway using nRef vs. using entire cohort data

* Cancer cohort data-based approach standardized gene expression values of an individual with the mean and s.d. of the entire cancer cohort data set where the sample originated from. This is counterpart experiment based on assumption what if an individual tumor sample is compared to many cancer data instead of many normal data. “Many cancer” data in this context is defined by cancer cohort dataset where an individual sample of interest is originated from. For example, if a tumor sample is one of member of GSE8894, the entire microarray data of GSE8894 is served for normalization and standardizing the expression value of the sample.

Result shows that using nRef-based approach generally increase the validation rate

	Beer			GSE8894			GSE3141		
	coef	pval	threshold	coef	pval	threshold	coef	pval	threshold
ASSEMBLY_OF_THE_PRE_REPLICATIVE_COMPLEX	0.520099	0.000415	0.386737	1.21574	0.000892	0.861136	0.710268	0.046145	0.55635
CYCLIN_E_ASSOCIATED_EVENTS_DURING_G1_S_TRANSITION_	0.42974	0.003513	0.247331	0.792707	0.027953	0.547786	0.716713	0.044535	0.213168
DNA_STRAND_ELONGATION	0.71499	1.00E-06	1.274833	1.071312	0.003607	1.391335	0.716089	0.04458	1.193017
FACILITATIVE_NA_INDEPENDENT_GLUCOSE_TRANSPORTERS	0.354506	0.01565	0.168691	0.763299	0.033131	-0.24277	1.184847	0.001134	-0.27284
G1_S_SPECIFIC_TRANSCRIPTION	0.830124	0	0.890295	0.944739	0.00934	1.179604	0.702067	0.049373	0.167741
G2_M_CHECKPOINTS	0.705606	2.00E-06	0.816311	0.904154	0.011586	1.161367	0.709231	0.04604	1.045875
GLUCOSE_TRANSPORT	0.770402	0	0.678839	0.700776	0.049534	0.284518	0.913949	0.010858	-0.28531
INTERACTIONS_OF_VPR_WITH_HOST_CELLULAR_PROTEINS	0.545955	0.000228	0.742698	0.841117	0.019373	0.558675	0.830168	0.019384	0.185536
M_G1_TRANSITION	0.552641	0.000178	0.513903	1.032637	0.00431	0.877324	0.733152	0.039056	0.637421
METABOLISM_OF_NUCLEOTIDES	0.415754	0.004697	0.674579	0.987276	0.006085	0.720526	0.701953	0.048824	0.607988
PHOSPHORYLATION_OF_THE_APC_C	0.688238	3.00E-06	0.487527	0.897167	0.013935	0.775273	0.743429	0.035953	0.891884
POL_SWITCHING	0.765461	0	0.759954	1.298079	0.000486	1.169592	0.742413	0.038171	0.909319
PROCESSING_OF_CAPPED_INTRONLESS_PRE_MRNA	0.44213	0.002712	0.238756	1.199786	0.001042	0.637821	0.715653	0.043672	0.673775
PYRIMIDINE_METABOLISM	0.496914	0.000737	0.670047	0.865313	0.015118	0.505466	0.76726	0.030959	0.15402
RECRUITMENT_OF_NUMA_TO_MITOTIC_CENTROSOMES	0.591678	6.10E-05	0.478313	0.850128	0.01803	1.178261	0.740686	0.038929	0.960324
REGULATION_OF_MITOTIC_CELL_CYCLE	0.657655	9.00E-06	0.47527	0.888198	0.013234	0.760152	0.74365	0.036574	0.567474
REPAIR_SYNTHESIS_FOR_GAP_FILLING_BY_DNA_POL_IN_TC_NER	0.761335	0	1.102294	1.298079	0.000486	1.339255	0.906748	0.011015	1.208045
SYNTHESIS_AND_INTERCONVERSION_OF_NUCLEOTIDE_DI_AND_TRIPHOSPHA	0.495164	0.000785	0.81846	0.731939	0.040875	1.201097	0.767719	0.032327	1.280663
SYNTHESIS_OF_DNA	0.575068	9.80E-05	0.589823	1.032637	0.00431	0.963818	0.733152	0.039056	0.718498
UNWINDING_OF_DNA	0.664002	7.00E-06	1.968073	0.884732	0.013615	1.719352	0.777411	0.028587	1.531343

Table 3.4. Pathways associated with LUAD survival validated in two independent datasets
logrank *P*-value, comparing tumors with the top half pathway aberrance score (Average Z) to the bottom.

Figure 3.5 Selected pathway for visual inspection of gene level expression and pathway level statistics.

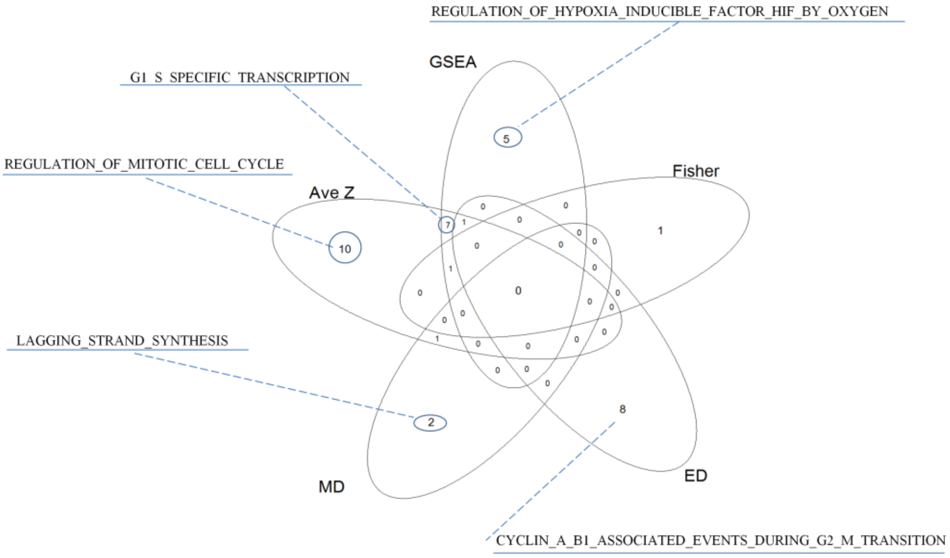
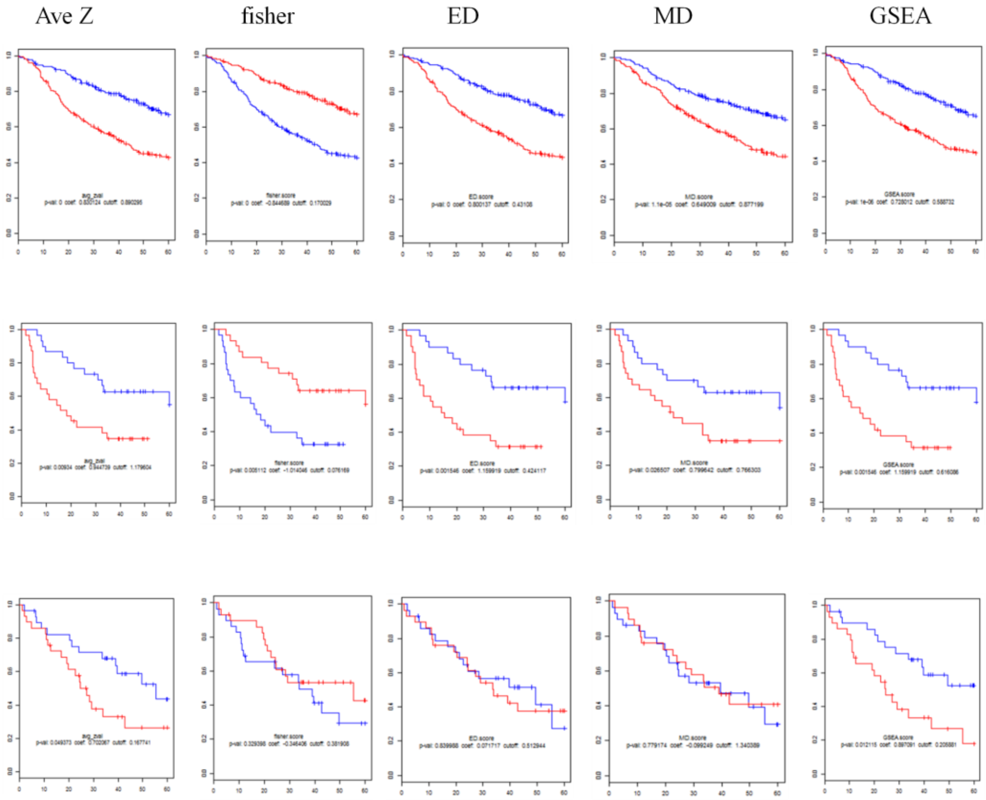

REGULATION_OF_HYPOXIA_INDUCIBLE_FACTOR_HIF_BY_OXYGEN
GSEA
G1_S_SPECIFIC_TRANSCRIPTION
5
REGULATION_OF_MITOTIC_CELL_CYCLE
Fisher
Ave Z
0
0
7
1
0
1
0
0
10
0
0
0
1
0
0
0
0
0
0
0
0
1
0
0
0
0
0
0
0
LAGGING_STRAND_SYNTHESIS
8
2
ED
MD
CYCLIN_A_B1_ASSOCIATED_EVENTS_DURING_G2_M_TRANSITION

Figure 3.6 Pathway based survival analysis (Reactome G1 S specific transcription) at discovery set (first row, Beer data set) and two independent validation set (second and third row, GSE8894 and GSE3141 respectively). AveZ and GSEA methods provide significant association in all data sets.



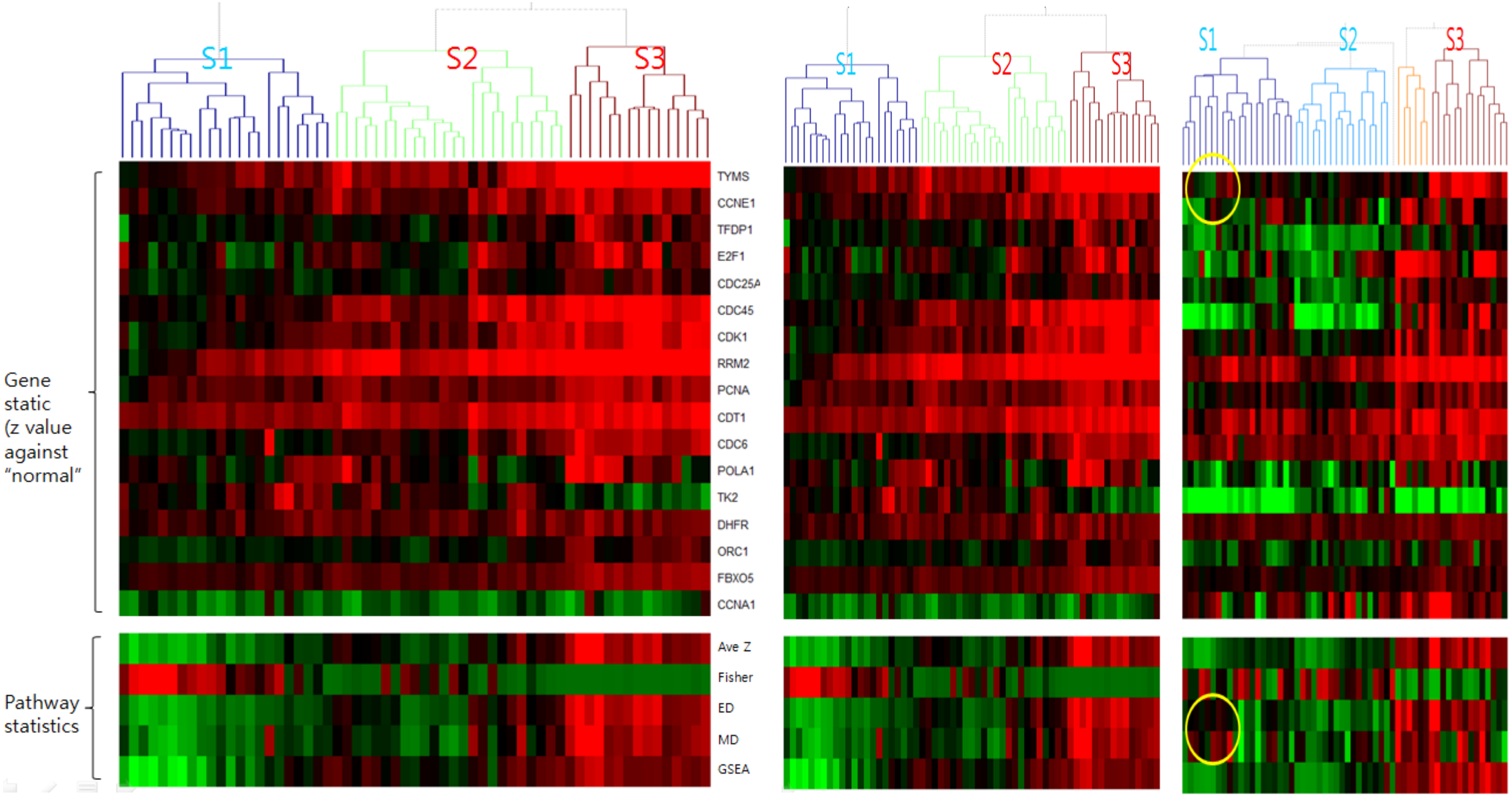


Figure 3.7 Gene and pathway level mRNA level expression profile of "G S1 specific transcription" pathway. The yellow circles indicate down regulated genes that cause pathway level difference in ED and MD methods.

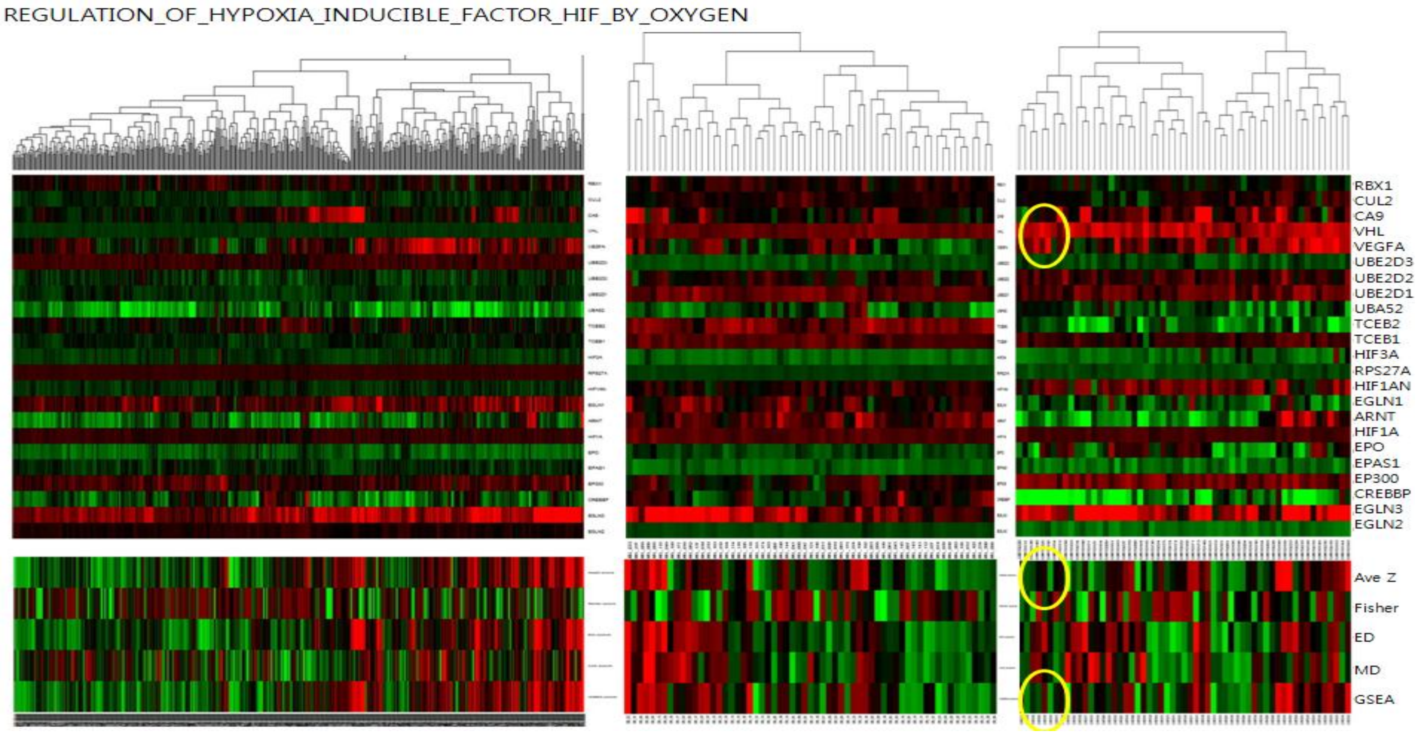


Figure 3.8 Gene and pathway level mRNA level expression profile of "Regulation of hypoxia inducible factor HIF by oxygen" pathway. The yellow circles indicate down regulated genes that cause pathway level difference in Ave Z and GSEA methods.

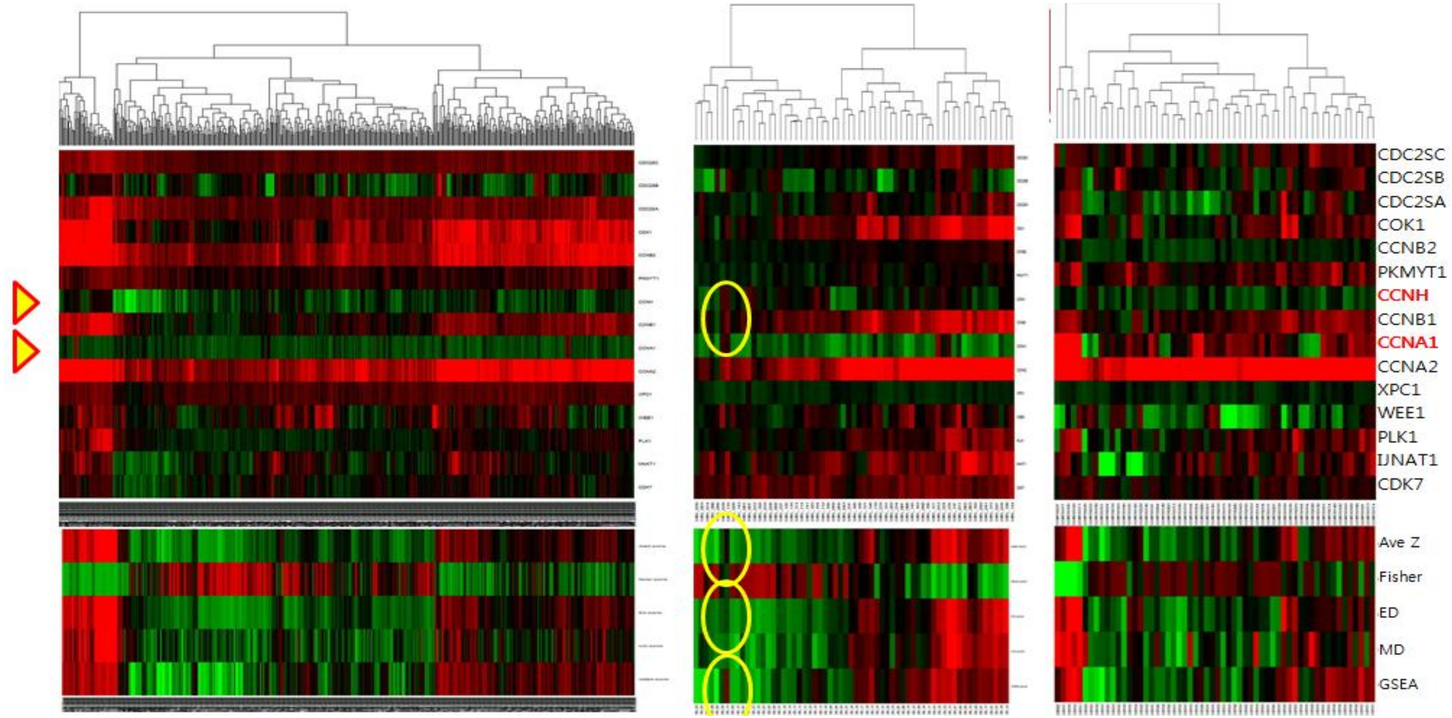


Figure 3.9 Gene and pathway level mRNA level expression profile of "Cyclin A B1 asssociated events during G2 M transition" pathway. The yellow circles indicate down regulated genes that cause pathway level difference between ED and MD.

3.3 Discovery of clinical importance

Cluster analysis of using Average Z as the iPAS method on Beer's data identified 12 pathway clusters (denoted by 1~12 in Figure 3.10) and 3 sample clusters (S2~S4, S1 is from the nRef, Figure 3.2). Sample clusters S2 and S4 represent well the differentiation status of LUAD (Fisher exact test, P<4.65*10-15). Well-differentiated adenocarcinoma resembles the normal glandular structure; therefore, it is a reasonable result that cluster S2 is close to the nRef. The survival outcome of S2 and S4 are significantly different (P<0.0028), and this assures us that unbiased clustering based iPAS has enough sensitivity to capture clinically important associations. This finding is concordant with prior knowledge that well-differentiated LUAD patients are likely to have better prognosis [104]. Pathway cluster P9 is distinguished as commonly up-regulated in tumor samples. The pathways are tRNA aminoacylation, amino acid or purine synthesis, DNA elongation, and the extension of telomeres.

Unbiased pathway based clustering of colon cancer data also captures clinically important associations by revealing sample clusters that are survival related (S2 and S3, P = 0.0037, Figure 3.11). It is important to note that iPAS is not only sensitive enough to identify clinically meaningful substructure of patients, but also reveals common characteristic of cancers at the same time. For example, pathways commonly up- or down-regulated in all cancer samples, for example, P9 or P2, would have not been discovered if the analysis had been performed by a conventional approach that does not make use of "nRef" (Figure 3.12).

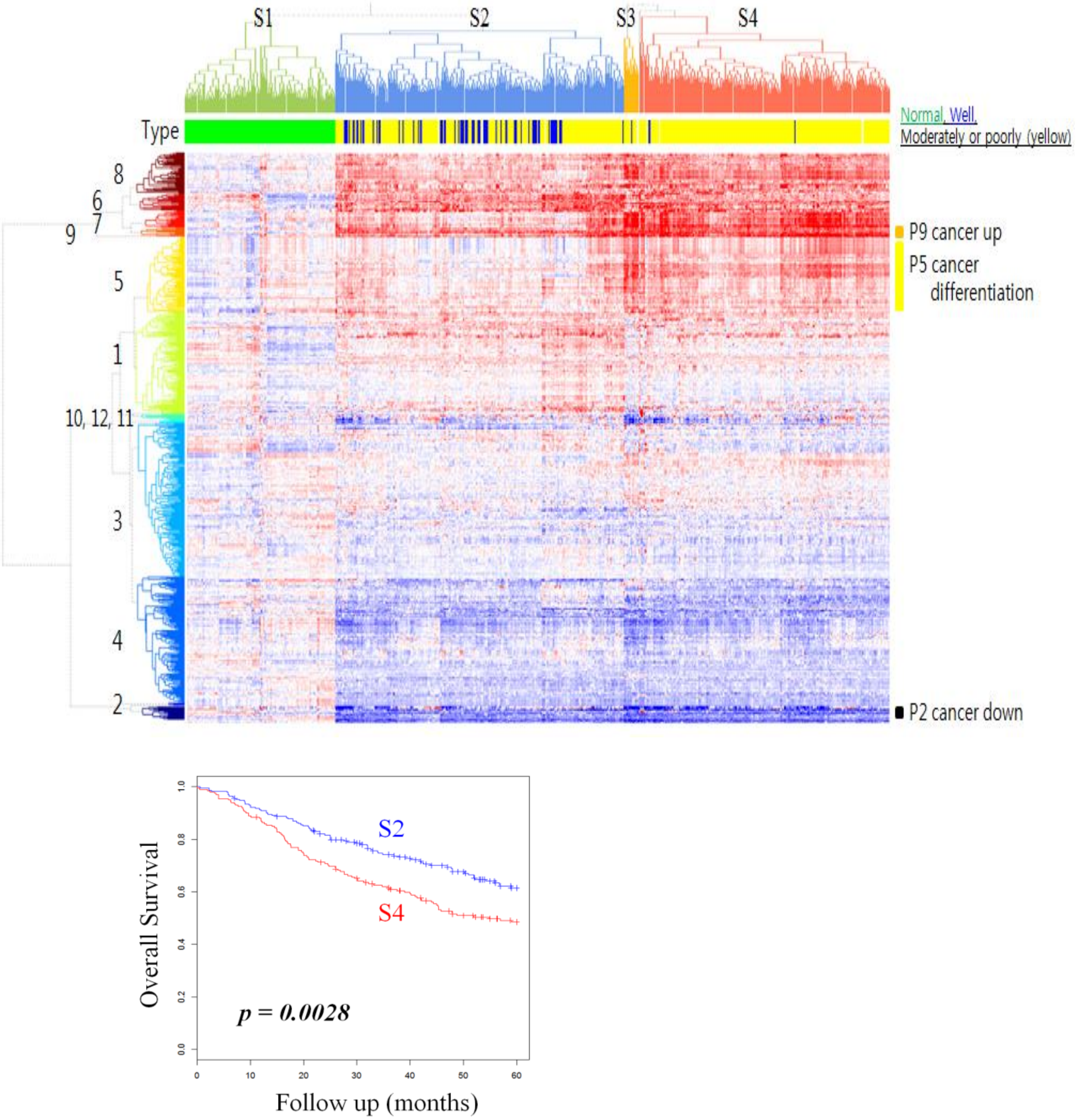


Figure 3.10 Clustered iPAS of lung adenocarcinoma dataset. Pathways (n = 583) and samples (n = 442) are clustered according to iPAS. Normal samples are clustered at left (S1). Tumors (S2~S4) deviate from normal in both up and down regulated directions (darker red and blue, respectively). Sample clusters are well representing histopathological differentiation status (S2 : for well-differntiated lung adenocarcinoma, P < 4.65 x 10-15) and overall survival.

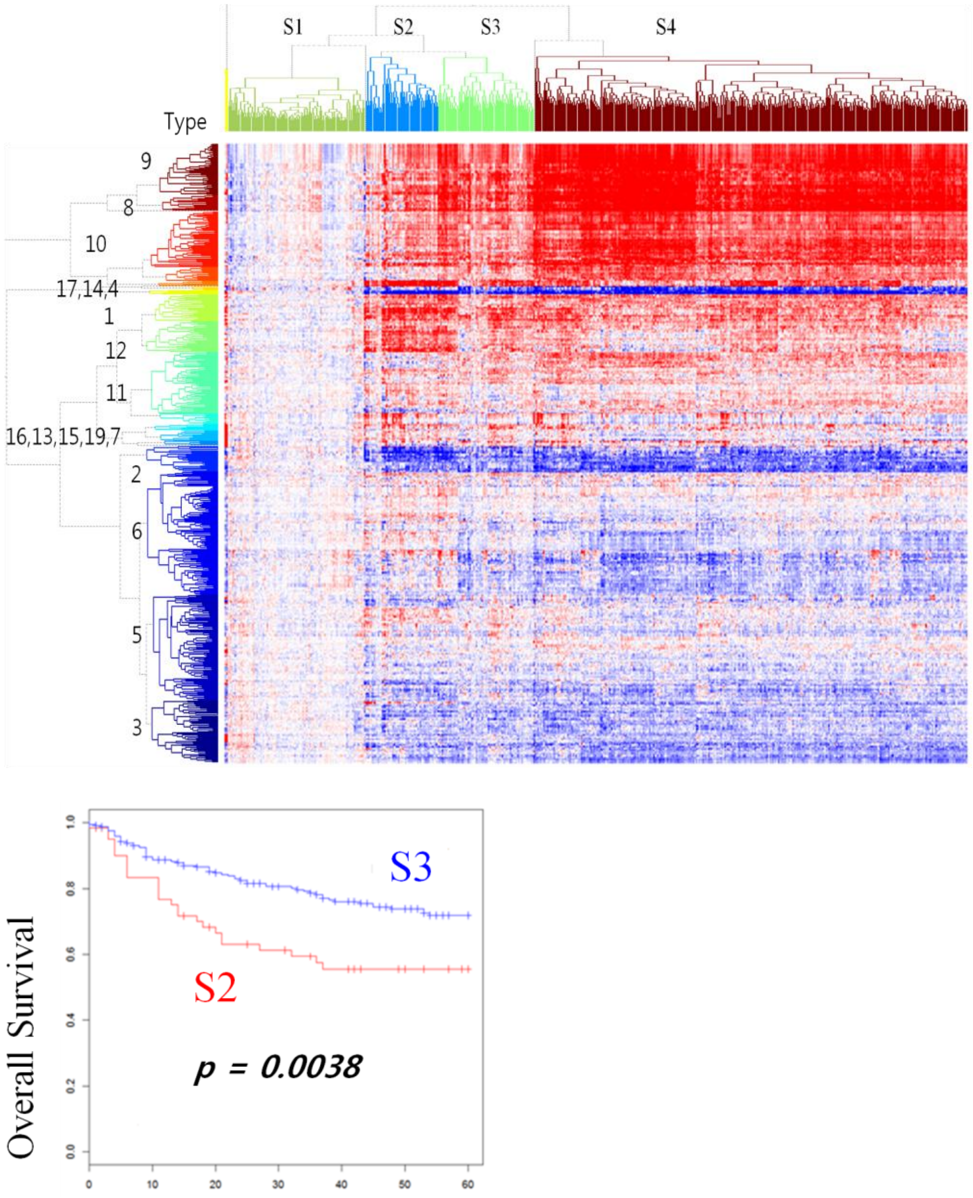


Figure 3.11 Unbiased clustering using iPAS of colon cancer dataset. Pathways (n = 583) and samples (n = 443) are clustered according to iPAS (Average Z). Normal samples are clustered at left (S1). Tumors (S2~S4) deviate from normal, being both up- and down-regulated, (darker red and blue, respectively). Sample clusters represent well the overall survival of patients (P = 0.0038). There were no survival differences between the sample cluster pairs of (S2,S4) and (S3,S4).

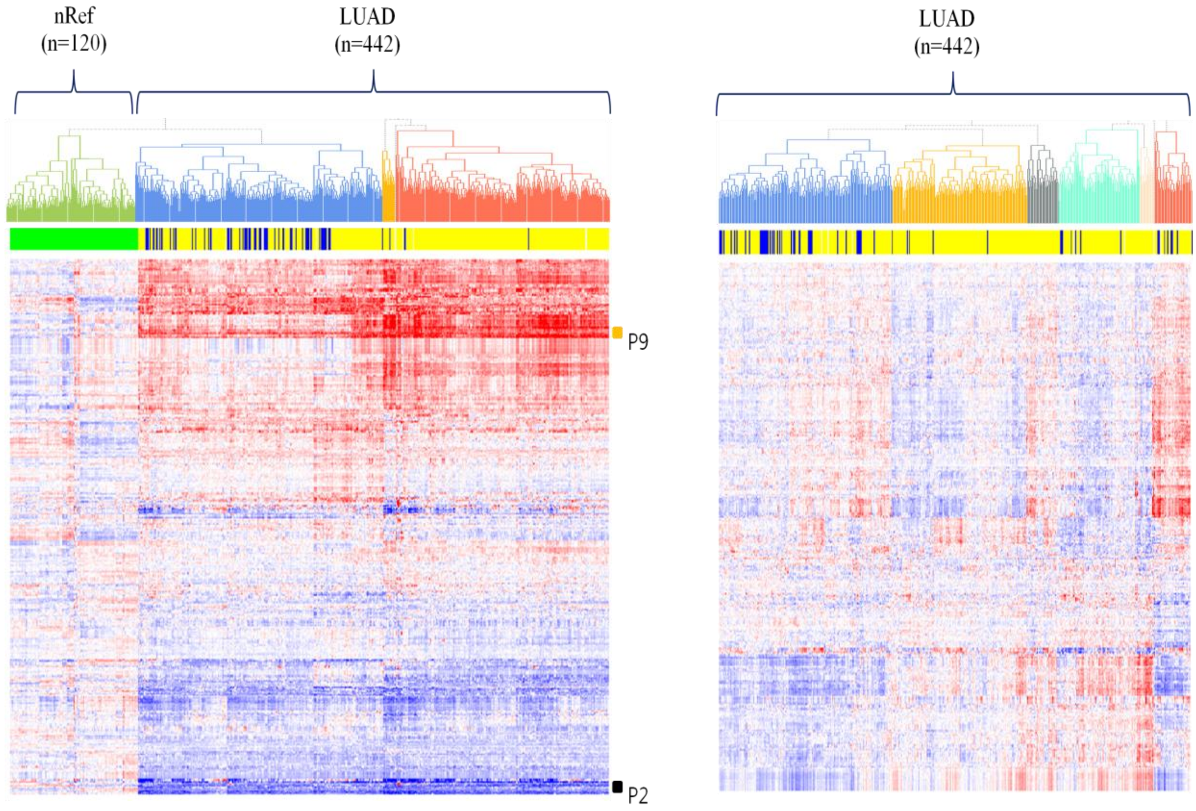


Figure 3.12 nRef-based approach (left) and the conventional approach (right) provide different interpretations. The same data was processed with two different approaches. The approach using nRef can identify globally up- or down-regulated pathways in cancer samples can also identify variant pathways across cancer samples. The conventional approach considers all cancer samples as a cohort, and then normalizes all of the samples together. Pathway score is represented in the conventional approach by averaging standardized gene expression values, where the mean and standard deviation of all cancer samples are used for standardization. Because an individual samples' pathway statistic is affected by the context of other cancer samples, pathways commonly up- or down-regulated in all cancer samples can be obscured.

3.4 Pathway based identification of cancer

Cancer develops unique mechanisms for malignancy. Therefore, it is reasonable to believe that identifying the unique molecular aberrances of cancer will aid in cancer diagnosis. Our empirical study of iPAS-based clustering of LUAD revealed several pathways commonly up- or down-regulated in all of the cancer samples. Further analysis was performed to determine if iPAS could be successfully used in the accurate identification of cancer. I tested this in a simple unsupervised way by judging whether an unknown sample is significantly different against the nRef, as a tumor, if not as normal. I performed a 5-fold cross validation one hundred times with the LUAD data set, which consisted of 120 cancers and 120 normal samples. Microarray data from the normal samples was randomly divided into five groups, and four of the five served as the reference groups. The remaining group was used as the test for the true normal set. The same amount of data was randomly picked from the cancer microarray data set and served as the true cancer set. I considered 583 pathways in REACTOME, giving a total of 293,500 (583 pathways × 5 fold × 100 repeats) AUCs and accuracy values. I averaged AUCs and accuracies from the five candidate methods for iPAS and used this as a representative AUC and accuracy of a given pathway.

By ranking the pathways by AUC, top pathways that marked averagely high performance by all iPAS candidates, are listed. The “amino acid synthesis and interconversion & transamination” pathway showed the highest classification performance. Unsurpri-singly, this pathway was one of the commonly up-regulated path-ways in the analysis of the Beer’s data (Figure 3.5, pathway cluster P9).

Among the tested iPAS candidates for this pathway, Mahala-nobis yielded the highest AUC (0.980), while Average Z gave 0.936 and Fisher's exact test gave the lowest value, (0.914) (Table 3.5). The standardized gene expression pattern for this pathway differed between tumor and normal. Many of the genes deviated from mean of the nRef, by more than two orders of sigma, contributing to its best performance out of all iPAS candidate methods, including ORA method like Fisher's exact test (Figure 3.13).

The "amino acid synthesis and interconversion & transamination" pathway consists of 17 genes involved in three major reactions, as it is described at REACTOME. The pathways are responsible for; (1) synthesis of three amino acids (aspartate, asparagine, glutamate), (2) the synthesis of glucose under fasting conditions by utilizing carbon atoms from these four amino acids, (3) conversion of amino acids to their corresponding alpha-keto acids, coupled to their conversion to glutamate, which is the first step in the catabolism of most amino acids. This function makes sense in terms of the "glutamine addiction" of cancer cells. The nutrients glucose and glutamine are specifically required by cancer cells as metabolites for growth and for production of ATP [105]. Myc and p53 have been revealed to be associated with this "addiction" by up-regulating glutamine synthesis in cancer cells. Thus, our finding is in accordance with prior knowledge regarding the up-regulation of glutamine synthetase.

I further validated our findings with an independent set that were not used in the discovery set. I collected two more LUAD gene expression data sets with normal data at GEO (GSE19188, GSE31547). Aggregated data sets of 48 microarrays from tumor tissues, and 35 microarrays from normal tissues were used for independent validation. The pathway was also altered in a cancer specific way in a validation set yielding an AUC of 0.982 by Mahalanobis-based iPAS (Figure 3.14, validation 1). I also assessed the same validation set in a different manner by using the nRef from the discovery set. Normal sample microarrays from the discovery sets (GSE10082, GSE7670, GSE10072) served as the nRef to classify samples in the independent validation set. The resulting AUC was 0.982 by the Mahalanobis method (Figure 3.14, validation 2). In our experiments using LUAD samples, the Mahalanobis distance, which used a pre-calculated covariance matrix from the 'nRef,' gave the best performance. Based on these results, we conclude that iPAS using Mahalanobis is best method to use in the pathway- based identification of cancer.

The biological role of this identified pathway is to supply nutrients and energy to cancer cells. This may be the reason why this pathway is universally aberrant in all the LUAD samples we assessed. I also investigated other cancers, if this pathway is useful to identify cancer. I collected microarray data of 156 breast cancers and 114 breast normals, 149 colon cancers and 145 colon normals, 151 ovarian cancers and 120 ovarian normals. Pathway based identification of cancer by using "Aminoacid synthesis and interconversion" pathway results AUC of 0.921 at breast cancer, 0.900 at colon cancer, 0.953 at ovarian cancer (Figure 3.15). Our analysis of this pathway in other cancer types demonstrated less of a role for this pathway. This suggests that the biological role of this pathway in cancer is more LUAD specific than other cancers.

I believe that the common disruption of this pathway is a novel discovery as this pathway, consisting of 17 genes, has not been reported as an indicator of LUAD in any of the studies we acquired data sets from (GSE10082, GSE7670, GSE10072), nor in a literature search with key words.

Pathway name (REACTOME)	Aver Z	GSEA	Fisher	ED*	MD*	Mean
AMINO_ACID_SYNTHESIS_AND_INTERCONVERSION_TRANSAMINATION	0.936	0.950	0.914	0.958	0.980	0.947
UNWINDING_OF_DNA	0.937	0.942	0.833	0.920	0.937	0.914
O_LINKED_GLYCOSYLATION_OF_MUCINS	0.925	0.939	0.833	0.955	0.910	0.912
SYNTHESIS_AND_INTERCONVERSION_OF_NUCLEOTIDE_DI_AND_TRIPHOSPHA	0.941	0.953	0.738	0.932	0.946	0.902
APC_CDC20_MEDIATED_DEGRADATION_OF_NEK2A	0.885	0.906	0.799	0.945	0.948	0.897
PURINE_RIBONUCLEOSIDE_MONOPHOSPHATE_BIOSYNTHESIS	0.905	0.915	0.820	0.921	0.912	0.895
PURINE_METABOLISM	0.915	0.918	0.729	0.945	0.936	0.889
DNA_STRAND_ELONGATION	0.889	0.920	0.783	0.930	0.916	0.888
DEGRADATION_OF_THE_EXTRACELLULAR_MATRIX	0.839	0.906	0.804	0.964	0.918	0.886
G1_S_SPECIFIC_TRANSCRIPTION	0.837	0.876	0.813	0.945	0.957	0.886
G0_AND_EARLY_G1	0.873	0.888	0.767	0.948	0.948	0.885
KINESINS	0.862	0.878	0.809	0.928	0.938	0.883
GLUCONEOGENESIS	0.890	0.910	0.783	0.903	0.902	0.877
ACTIVATION_OF_THE_PRE_REPLICATIVE_COMPLEX	0.892	0.908	0.737	0.909	0.918	0.873
E2F_MEDIATED_REGULATION_OF_DNA_REPLICATION	0.903	0.903	0.725	0.925	0.893	0.870
DEPOSITION_OF_NEW_CENPA_CONTAINING_NUCLEOSOMES_AT_THE_CENTROM	0.911	0.930	0.828	0.966	0.706	0.868
EXTENSION_OF_TELOMERES	0.840	0.889	0.762	0.913	0.926	0.866
CELL_CYCLE	0.894	0.922	0.766	0.955	0.780	0.863
INHIBITION_OF_THE_PROTEOLYTIC_ACTIVITY_OF_APC_C_REQUIRED_FOR_	0.852	0.881	0.745	0.888	0.951	0.863

Table 3.5. AUC of pathway-based classification of tumor sample via different pathway summary methods (ED*: Euclidean distance, MD*: Mahalanobis distance)

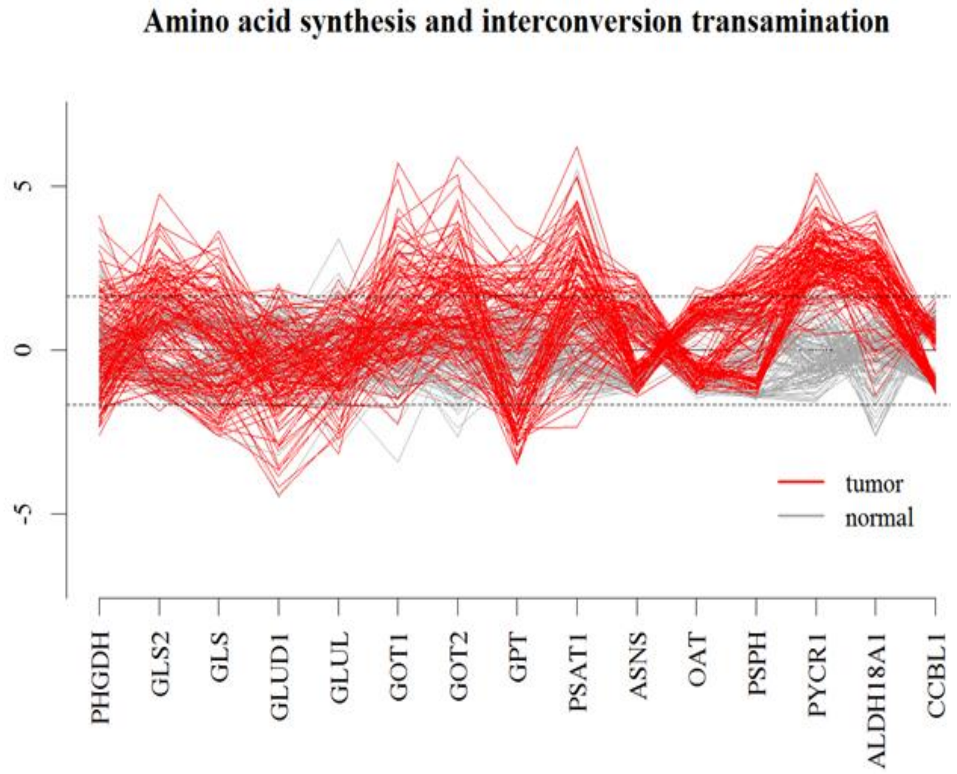


Figure 3.13 Expression pattern of genes in the pathway. Each line represents sample. (Grey: Normal, Red: Tumor). Dashed line represents expression value deviated 1.96σ from the mean expression value of normal tissues.

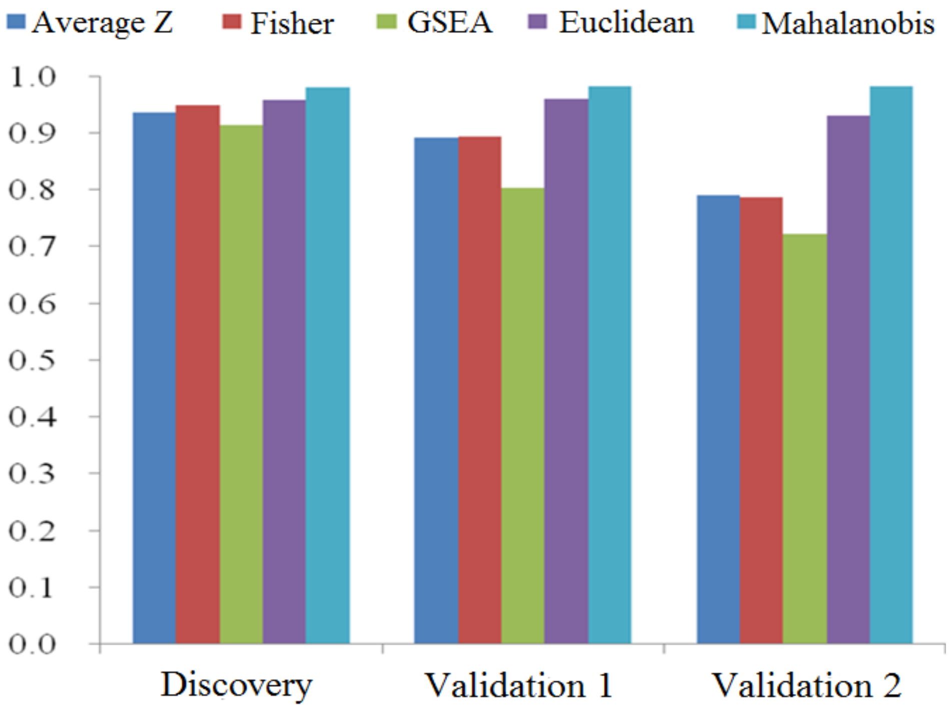


Figure 3.14 Performance of classification of cancer by ‘Amino acid synthesis and interconversion transamination’. AUC of 0.980 has marked in discovery set, independent validation set results AUC of 0.982 (validation 1: normal samples in validation set served as reference) and 0.982 (validation 2: normal samples in discovery set served as reference).

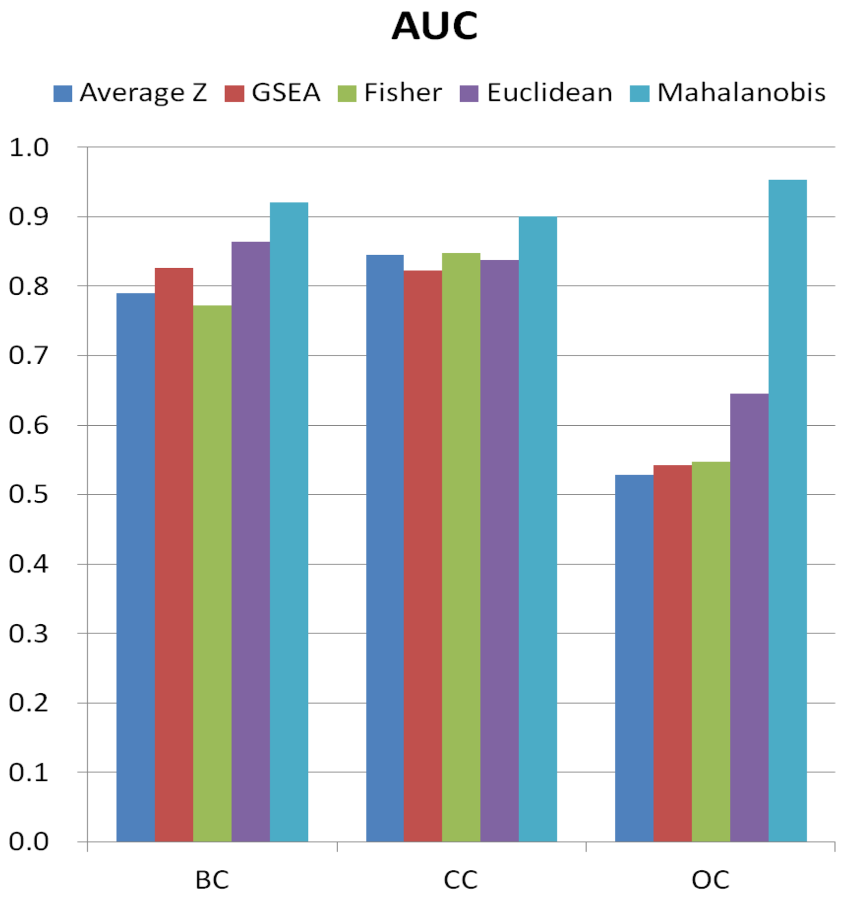


Figure 3.15 Identification of cancer by the ‘amino acid synthesis and interconversion transamination’ pathway. The pathway identified in lung adenocarcinoma was investigated to determine its usefulness in the identification of other cancer types (BC: Breast Cancer, CC: Colon Cancer, OC: Ovarian cancer).

Table 3.6. Identification of cancer by "amino acid synthesis and interconversion transamination" pathway in other cancer types. The pathway shows decreased performance in the identification of cancer in samples from other cancer types, suggesting the pathway aberrance is more specific to lung adenocarcinoma.

	CANCER TYPE	Average Z	GSEA	Fisher	Euclidean	Mahalanobis
Accuracy	BC	0.689	0.689	0.467	0.737	0.863
	CC	0.776	0.738	0.667	0.544	0.738
	OC	0.424	0.443	0.446	0.458	0.838
AUC	BC	0.790	0.827	0.772	0.864	0.921
	CC	0.844	0.823	0.848	0.837	0.900
	OC	0.528	0.542	0.547	0.645	0.953

BC: Breast cancer (Test normal: 114, Test tumor: 156)
CC: Colon cancer (Test normal: 145, Test tumor: 149)
OC: Ovarian cancer (Test normal: 120, Test tumor: 151)

3.5 Pathway based identification of rare mutation effect in cancer

Mutation information is getting more useful for stratified medicine. For example, KRAS mutation is a test recommended to targeted drugs of colon cancer therapy [106], EGFR activation mutations and resistant mutations for targeted therapy for non small cell lung cancers [107]. Only a few mutations are known to be clinically actionable, most of less frequent mutations are remained to be understood.

In conjunction with iPAS, we performed integrated analysis of mutation and pathway level gene expression to discover a combination of rare mutations that causes pathway level gene expression changes. Assuming combination of functionally related rare mutations can influence the pathway of mRNA expression, we consider multiple rare mutations to be counted as a single mutation event. At the first step of our analysis, we assess if pathway level mRNA expression is significantly different between the group of with mutation and without mutation. If it is different, we add another mutation sites to be counted as a single mutational event, then assess if the new event can still differentiate pathway level mRNA expression. I iterate this procedure until no more rare mutation can be added into mutational event under the certain significance threshold for pathway level gene expression difference. The output can biologically be interpreted as a set of mutations that influence the pathway level gene expression, thus the mutations can be considered functional in cancer, further be prioritized as cancer drivers.

To assess whether our method can sensitively capture the impact of mutational event to the pathway level gene expression, we analyzed breast cancer data set from TCGA having paired somatic mutation and expression data (n=513). Mutational event of single gene is described at Figure 3.16. Mutations that changes gene expression level of the mutated genes (FDR q-value < 0.1) are shown in blue. Mutated genes those gene level expression are changed, but pathway level mRNA expression are changed (FDR q-value < 0.1) are shown in red. In the latter case, there were three mutated genes causing pathway level difference in 24 pathways. The three genes are TP53 (187 are mutated out of 513 samples, 36.4%), PIK3CA (173/513, 33.7%), RB1 (11/513, 2.1%), previously implicated in breast cancer.

It is noteworthy that the pathway of the three genes are addressed as crucial by the original research [91]. In the TCGA [91], three pathways (PI3K, TP53 and RB pathways) are considered as representative pathways for breast cancer, and addressed its functional status with protein level phosphorylation. In our analysis, we discovered mutations on TP53, PIK3CA and RB1 have significant impact on pathways, without any prior knowledge, but by analyzing the mutation and mRNA expression data. This indirectly proves that our approach is sensitive enough to capture the important biological features; thus it can be properly considered to measure pathway level impact of a somatic mutation.

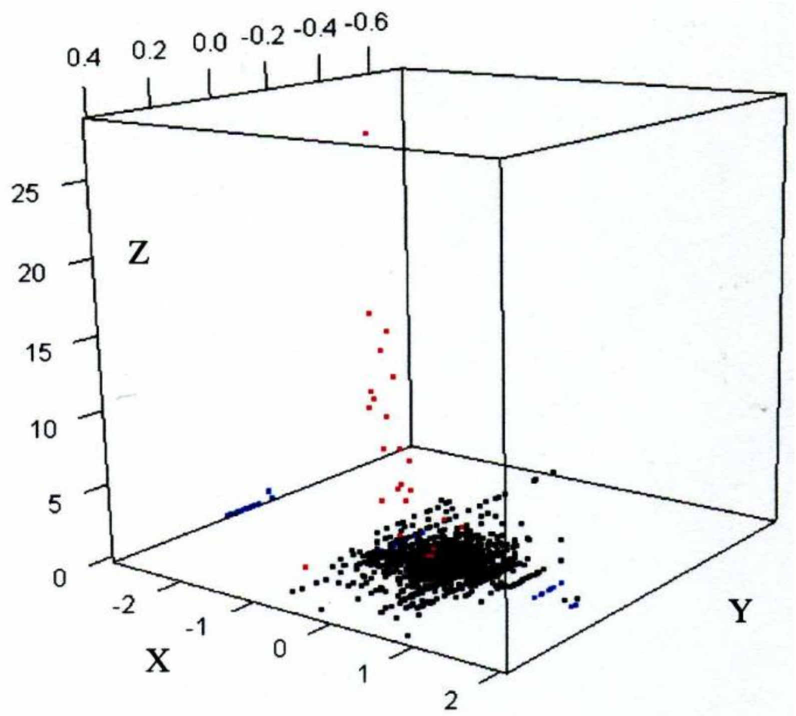


Figure 3.16 Single gene's mutational influence on mRNA expression at gene level (X axis) and pathway level (Y axis). X : averaged gene expression difference of mutation having group minus non having group). Y : averaged pathway level difference of mutation having group minus non-having group). Z : -log10p score, where p is from t-test of pathway statistics between mutation having group vs non-having group. Red : mutational event where its influence on pathway level is significant (FDR q-value<0.1). Blue dots : mutational event its influence on gene level is significant (FDR q-value<0.1) but not significant at pathway level

Twenty four pathways showed differential mRNA expression between groups with and without mutations of TP53, PIK3CA and RB1 mutations (Figure 3.17). TP53 and PIK3CA are not mutually exclusive at the observation of TCGA breast cancer data, which is concordant to the previous report [40]. Heatmap visualization of unsupervised clustering of pathway level characteristics shows distinguishing subgroup pattern between "TP53 mutated and PIK3CA non-mutated samples enriched" subgroup (C and D) and "TP53 non-mutated and PIK3CA mutated samples enriched" subgroup (A and B). This characteristic might be explained based on to previous findings that TP53 gene product regulates PIK3CA in a transcriptional level.

Astanehe et al., [109] demonstrated that direct binding of TP53 reduces the expression of PIK3CA, thus decreases the expression of PIK3CA expression. In our analysis, sample cluster of C and D, representing un-mutated samples on PIK3CA keeps this mRNA deregulation functionality of PIK3CA, showing pathway level down-regulation in PIK3CA related pathway cluster P1. Unlike sample cluster C and D, PIK3CA mutation enriched sample cluster of A and B might have TP53 mediated regulation of PIK3CA gene product, showing pathway level up-regulation in the pathways P1.

Mutations on RB1 are enriched at sample group C and D, showing a tendency that it is coupled to TP53 mutation status. Subgroup having mutations on both of RB1 and TP53 has unfavourable outcome when it is compared to the others. The observation is concordant to the known biological knowledge that breast

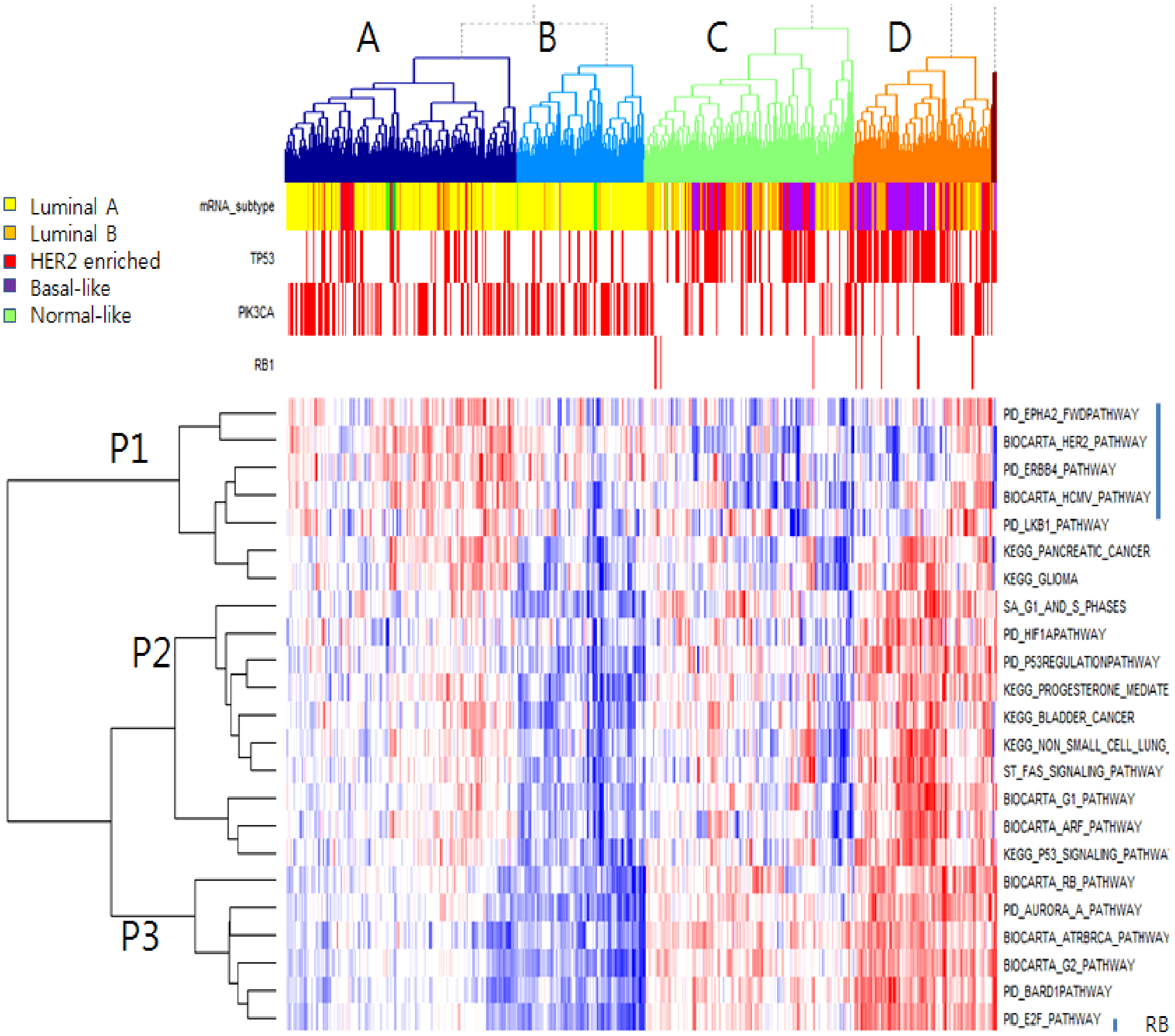


Figure 3.17 Normalized heatmap description of top pathway influencing mutations (PIK3CA, TP53 and RB1, q-value < 0.1). Column side bar describes mutation events on RB1, TP53, PIK3CA with red bars, respectively.

In the analysis of impact of single gene mutation into the pathway, rare mutations having less than 11 mutated samples were not reported as significant at q-value < 0.1. To further assess combinatorial rare mutation genes in the pathway level mRNA expression, we iteratively combine mutated genes so that multiple mutational events can be considered on multiple genes into one mutational event increasing the sample size of group with mutations. Mutations reported as significant in the single gene mutation analysis is discarded in the combinatorial iteration, to avoid its strong impact and hopefully to find additive impact of rare mutations only.

Through our additive combination of rare mutations into one event, we have found 19 mutation events affecting on pathway level at the cut of FDR q-value < 0.25. Figure 3.18 depicts the relationship of pathway level mRNA expression difference between groups with and without mutational events (X axis), number of samples having mutation events (Y axis), and the significance of impact of the mutational event (Z axis).

Among 15,819 genes with somatic mutations reported in 513 breast cancer samples, 63 genes were shown to have pathway level impact when it is combinatorially considered as a mutational event. Gene ontology analysis using g:Profiler [110,111] showed significant functional enrichment of these genes into cancer related biological processes like ‘cell proliferation (p-value = 1.17e-07)’, ‘cellular component organization or biogenesis (p-value = 6.45e-15)’, ‘response to stimulus (p-value = 1.7e-13)’, ‘cell cycle (p-value = 2.16e-19)’ and ‘developmental process (p-value = ‘1.3e-14)’. This is a supportive information that our method not just coincidentally picked rare mutation genes with pathway impact, but sensitively reveal the additive impact of rare mutations in the context of pathway level mRNA expression change.

Based on these results, we suggest the 63 genes along with 3 genes that showed single gene’s mutational influence on pathways as potential tumor driver mutations, having more functional importance than the other 15,750 somatic mutations. Table 3.7 provides mutational events that have shown pathway level impact in the combinatorial summary of mutation event among the 63 genes.

I further investigated pathway level impact of somatic mutations against gene networks having actionable drugs. Among 19 receptor tyrosine kinase related pathways, three pathways have shown pathway level difference with two mutation event (Table 3.8).

LYN is a member src kinase superfamily and is known to be involved in the regulation of cell activation. Recent publication [112] addressed SH2 domain missense mutation D197Y at breast cancer is functional. Over expression of D197Y is more potent than wild type LYN at inducing signaling cascade, rendering the treatment of ER downreguatlor fulvestrant or PI3K inhibitor BKM120 less effective. This indicates LYN may play a role for ER+ breast cancer acquiring hormone-independent growth. Two LYN mutations in our discovery for RTK pathway related mutation event were located SH2 domain (E159K, K188N), they may be considered to have similar contribution to D197Y.

NCK1 is downstream of signal cascade of LYN. Its major function is activating actin cytoskeleton reorganization. There is no evidence how LYN and NCK1 regulate the transcription level of PDGFR pathways. Our observation indicates that the group with either mutation of the two genes has lower level of gene expression in the PDGFR pathway than the group without mutations.

Among mutation event of 16 samples either of three genes (PIK3R1, PIK3CD and GRB2), PIK3R1 is the most frequent (# of event samples : 14). Most of 14 mutations clustered in the PIK3CA interaction domain. BKM120 and GDC-0941 are the suggested drugs for patients having mutation at PIK3R1 sites [91]. PIK3R1, PIK3CD and GRB2 are all interacting together, having mutation on these genes can cause similar impact on downstream pathways.

In summary of receptor tyrosine kinase pathways, we additionally discovered two mutational events that have significant pathway level mRNA change. Literature survey on discovered mutations also revealed that the mutations are potential drug target. This is another supporting evidence that our method can sensitively detect functional rare mutations, in other words, measuring pathway level impact of summarized rare mutation events is useful to prioritize functional ones.

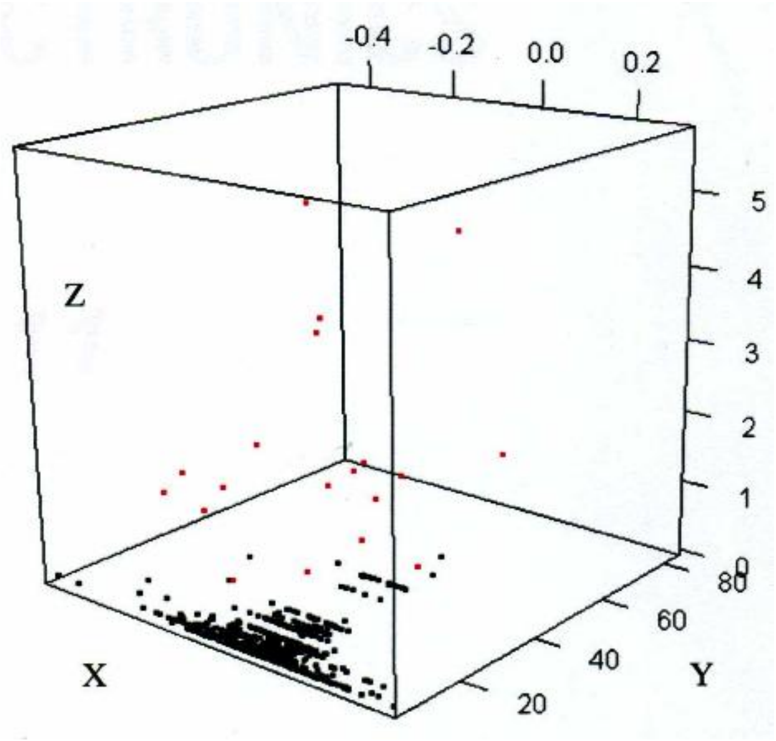


Figure 3.18 Multiple gene's mutational influence on mRNA expressino at pathway level (X axis). X : averaged pathway level difference of mutation having gorup minus non-having group. Y : number of samples having summarized multi-gene mutation event. Z : -log10p score, where p is from t-test of pathway statistics between mutation event having group vs non-having group. Red : mutational event where its influence on pathway level is significant (FDR q-value<0.25).

Table 3.7. List of multiple gene's mutation events having impact on pathways

Pathway	Mutation event (# of distinct genes 63)	# event sample	t-stat	q-value*
KEGG_GLIOMA	PIK3CB,HRAS	5	12.587	0.000
KEGG_MELANOMA	PIK3CB,BRAF	7	10.302	0.000
PID_CDC42	CDH1,MAP3K1	70	-5.244	0.000
PID_TRAIL	RIPK1,MAPK3	6	9.624	0.000
BIOCARTA_PPARA	NCOR1,EHHADH	21	-4.562	0.007
PID_A6B1_A6B4_INTEGRIN	COL17A1,GRB2	6	-6.239	0.009
BIOCARTA_MTOR	TSC1,TSC2	7	-5.081	0.010
SIG_PIP3_SIGNALING_IN_B_LYMPHOCYTES	ITPR3,RPS6KA3	7	6.174	0.014
PID_CERAMIDE	MAP2K4,AKT1,RIPK1	35	4.803	0.017
SA_PTEN	AKT3,BPNT1	5	-6.573	0.018
ST_FAS_SIGNALING	MAP3K1,EZR	42	-3.727	0.029
PID_EPHBFWDPATHWAY	EPHB1,EFNB1	9	-4.848	0.032
KEGG_RENAL_CELL_CARCINOMA	EPAS1,GRB2,PDGFB	7	7.628	0.057
PID_ECADHERIN_KERATINOCYTE	CDH1,FMN1,PIP5K1A,EGFR,AKT2,CDH1,RAC1,CDH1,RAC1	49	-5.564	0.074
KEGG_BLADDER_CANCER	CDH1,THBS1,MDM2,RAF1	41	-4.966	0.083
KEGG_PROGESTERONE_MEDIATED_OOCYTE_MATURE~	ADCY9,RPS6KA3,MAD2L1,MAPK3,PRKX,ARAF,MAPK13,PIK3CD,RPS6KA1,CDC25A,IGF1,HSP90AA1,MAPK10,GNAI2,IGF1R,MAPK1,CDK1	29	13.209	0.144
PID_CDC42	MAP3K1,CDH1,APC,CDC42BPA,VAV2,IQGAP1	84	-5.985	0.183
KEGG_ERBB_SIGNALLING_PATHWAY	PRKCG,PAK1,GSK3B,PLCG2,TGFA,PRKCA,PAK4	11	-11.81	0.242

Table 3.8. List of multiple gene's mutation events having impact on drug target centric pathways

Pathway	Mutation event	# event samples	t-stat	q-value
PDGFRB_neighbors	LYN, NCK1	5	-8.3139	0.005
FGFR2_neighbors	PIK3R1, PIK3CD, GRB2	18	3.506	0.141

Chapter 4. Summary and Conclusion

In this thesis, we have proposed personalized extensions to ORA- and FCS-based pathway analysis by introducing the concept of comparing an individual tumor to many normal samples. Exploratory analyses of our methods with previously published survival pathway signatures reproduced the correct survival outcomes. I have also demonstrated that using nRef improves the validation rate. Unbiased clustering with individualized pathway aberrance scores revealed sample clustering, which is indicative of the cancer differentiation status of LUAD and of different survival outcomes. Clustering also identifies pathway characteristics from patients displaying common up- or down-regulations and sub-group specific deregulations.

Pathways that are commonly deregulated across all cancer patients may be useful in identifying cancer from unknown samples. I explored the pathway-based identification of cancer with “amino acid synthesis and interconversion transamination” pathway, which is commonly up-regulated in LUAD patients. Validation using independent data sets demonstrated that this pathway is useful in classifying LUAD and normal lung samples.

Based on our results, we conclude that individualized pathway scores using nRef can provide a sensitive measure of a patient’s clinical features and can be useful for identifying cancer.

In our empirical study, Average Z performed best in highlighting pathway aberrance and in further revealing clinical importance. It had the best statistical power when identifying a previously known survival related pathway and the best averaged validation rate of survival related pathways for LUAD and colon cancer. In the pathway-based identification of cancer, the Mahalanobis method performed best.

Additionally, we developed an approach that assesses mutational impact on pathway level mRNA expression. I suggest combinatorial summary of mutational events. I have demonstrated that the approach sensitively discovers important mutations that have been known to have pathway level impact. Important mutations that cause deregulation of representative breast cancer pathways reported by previous study have been captured..

Combinatorial mutation summary found 63 genes that showed pathway level difference between groups with and without mutation events. Based on pathway level impact analysis result, we suggested functional importance of somatic mutations on the 63 genes is bigger than the others. I also investigated impact of rare mutations on drug target pathways; we found two mutation events that consisted of two and three genes, respectively. Two of total five mutations were mentioned as potential drug target in the literature, indirectly supporting that our approach is useful to prioritize druggable mutations.

Due to the innovation of next generation sequencing technology, more cancer patients' genome and transcriptome are expected to be available.

Accumulation of data will enable more accurate estimation of functional impact of rare mutations.

There are three challenging issues further to be investigated as an extension of this work. Firstly, determining platform dependency should be considered. I have applied our approach into data sets from the same platforms. To address the pathway level aberrance of a single cancer sample accumulated normal data from the same platform to the cancer case must be collected. Addressing if different platform data (e.g next generation sequencing data of cancer case against microarray data of normals) is interpretable is next challenge. Secondly, regarding the methodology, extending our approach by considering pathway topology (PT) is considerable. Impact factor analysis introduced at chapter 1 and appendix considers the impact of functional neighbors [66,67]. Thirdly, extending our philosophy to multiple omics data is considerable. Although we have demonstrated the concept of our approach in mRNA level data only, it is not limited to a single layered omics data. In our methods, gene level statistics standardized to represent how much a single cancer sample data is deviated from accumulated normal reference data. This philosophy can be implemented in multiple omics data by various statistical methods such as combining p-values or summarized statistics [113].

An important clinical aspect of our methods is that it enables the interpretation of a cancer case in a single patient, even if matched normal tissue data from the same individual is unavailable. Accumulated information of normal tissues from a data repository will take the place of data unavailable for a specific individual. As the data repository is growing rapidly, it is expected that more “nRef” data will be available for many diseases in the near future. I hope that our proposed approach can help in the personalized interpretation of tumor data and can be a useful tool in the upcoming era of data-based personalized medicine.

I hope that our proposed approach can be used to discover mutations having functional impact, thus further to prioritize mutations for the consideration of customized therapy.

Appendix

The code for this study is available at

http://bibs.snu.ac.kr/ipas

A quick start instruction is as follows.

```
#Required R library
install.packages("preprocessCore")
install.packages("MASS")
```

(make sure to unzip the folder bgDistribution)
Execute R and set the working directory to the unzipped folder of downloaded codes.
Run the following codes.

```
/*
tumorFile = "Beer_10samples.txt"
refFile = "LUAD_nREF.txt"
pathwayListFile= "reactome.txt"
whoami = "testRun2"
runMode = "nRef"

sourcePath = "iPAS_library.r"
source(sourcePath )
bgDistributionPath = "./bgDistribution"

iPAS( tumorFile, refFile, pathwayListFile, whoami, runMode,
bgDistributionPath )
*/
\
```

Impact Factor Analysis

Impact Factor (IF) analysis [66, 67] combines both ORA and FCS approach, while accounting for the topology of the pathway. IF analysis computes Perturbation Factor (PF) for each gene in each pathway, which is a gene-level statistic, as follows

$$PF(g_i) = \Delta F(g_i) + \sum_{j=1}^{n} \beta_{ji} \cdot \frac{PF(g_j)}{N_{ds}(g_j)} \qquad \text{Eq. 1.}$$

In Eq. 1, the first term, ⊿F(g_i), represents the signed normalized measured expression change (i.e.,fold change) of the gene g_i. The second term in Eq. 1 accounts for the topology of the pathway, where gene g_j is upstream of gene g_i. In the second term, *ji* represents the type and strength of interaction between g_i and g_j . If g_j activates g_i, *ji* = 1, and if g_j inhibits g_i, *ji* = -1. Note that the PF of the upstream gene gj is normalized by the number of downstream genes it interacts with, $N_{ds}(g_i)$. The second term is repeated for every gene g_j that is upstream of gene g_i.

After computing PF for each gene, pathway-level statistic, Impact Factor (IF), is computed using Eq. 2:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)} \qquad \text{Eq. 2.}$$

In Eq. 2, the first term captures the significance of the given pathway P_i as provided by ORA, where pi corresponds to the probability of obtaining a value of the statistic used at least as extreme as the one observed when the null hypothesis is true. Because IF should be large for severely impacted pathways (i.e., small p-values), the first term uses 1=pi rather than pi. The log function is necessary to map the exponential scale of the p-values to a linear scale in order to keep the model linear. The second term sums up the values of the PFs for all genes g on the given pathway P_i, and is normalized by the number of differentially expressed genes on the given pathway P_i.

Note that Eq. 1 essentially describes the perturbation factor PF for a gene g*i* as a linear function of the perturbation factors of all genes in a given pathway. Therefore, the set of all equations defining the PFs for all genes in a given pathway Pi form a system of simultaneous equations. Expanding and re-arranging Equation 1 for all genes g_1; g_2; : : : ; g_n in a pathway Pi can be re-written as follows:

$$\begin{pmatrix} PF(g_1) \\ PF(g_2) \\ \cdots \\ PF(g_n) \end{pmatrix} = \begin{pmatrix} 1-\frac{\beta_{11}}{N_{ds(g_1)}} & -\frac{\beta_{21}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n1}}{N_{ds(g_n)}} \\ -\frac{\beta_{12}}{N_{ds(g_1)}} & 1-\frac{\beta_{22}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n2}}{N_{ds(g_n)}} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{\beta_{1n}}{N_{ds(g_1)}} & -\frac{\beta_{2n}}{N_{ds(g_2)}} & \cdots & 1-\frac{\beta_{nn}}{N_{ds(g_n)}} \end{pmatrix}^{-1} \begin{pmatrix} \alpha(g_1)\cdot\Delta E(g_1) \\ \alpha(g_2)\cdot\Delta E(g_2) \\ \cdots \\ \alpha(g_n)\cdot\Delta E(g_n) \end{pmatrix}$$

After computing the PFs of all genes in a given pathway as the solution of this linear system, Eq. 2 is used to calculate the impact factor of each pathway. The impact factor of each pathway is then used as a score to assess the impact of a given gene expression data set on all pathways (the higher the impact factor the more significant the pathway).

References

1. Khatri, P., Sirota, M. and Butte, A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges, PLoS Comput Biol, 8, e1002375.
2. Jui-Hung Hung et al. Gene set enrichment analysis: performance evaluation and usage guidelines. Brief Bioinform (2012) 13 (3): 281-291 first published online September 7, 2011
3. Lu, C. (2004) Improving the scaling normalization for high-density oligonucleotide GeneChip expression microarrays, BMC Bioinformatics, 5, 103.
4. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics, 4, 249-264.
5. Steinhoff, C. and Vingron, M. (2006) Normalization and quantification of differential expression in gene expression microarrays, Brief Bioinform, 7, 166-177.
6. Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003) Summaries of Affymetrix GeneChip probe level data, Nucleic Acids Res, 31, e15.
7. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nat Methods, 5, 621-628.
8. Celton, M., A. Malpertuy, et al. (2010). Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. BMC Genomics 11: 15.
9. Brock, G.N., Shaffer, J.R., Blakesley, R.E., Lotz, M.J. and Tseng, G.C. (2008) Which missing value imputation method to use in expression profiles: a

comparative study and two selection schemes, BMC Bioinformatics, 9, 12.

10. Bo, T.H., Dysvik, B. and Jonassen, I. (2004) LSimpute: accurate estimation of missing values in microarray data with least squares methods, Nucleic Acids Res, 32, e34.

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

12. Saxena, V., Orgill, D. and Kohane, I. (2006) Absolute enrichment: gene set enrichment analysis for homeostatic systems, Nucleic Acids Res, 34, e151.

13. Leone FC, Nelson LS, Nottingham RB. The folded normal distribution. Technometrics 1961;3(4):543–50. 25. Irizarry RA, Wang C, Zhou Y, et al. Gene set enrichment

14. Irizarry, R.A., Wang, C., Zhou, Y. and Speed, T.P. (2009) Gene set enrichment analysis made simple, Stat Methods Med Res, 18, 565-575.

15. Leong, H.S., Yates, T., Wilson, C. and Miller, C.J. (2005) ADAPT: a database of affymetrix probesets and transcripts, Bioinformatics, 21, 2552-2553.

16. Harbig, J., Sprinkle, R. and Enkemann, S.A. (2005) A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array, Nucleic Acids Res, 33, e31.

17. Hong, F. and Breitling, R. (2008) A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments, Bioinformatics, 24, 374-382.

18. Fisher R. Statistical Methods for ResearchWorkers. 4th edn. London: Oliver and Boyd, 1932. 30. Rosenthal R, Hiller JB, Bornstein RF, et al. Meta-analytic

19. Rosenthal R, Hiller JB, Bornstein RF, et al. Meta-analytic methods, the Rorschach, and the MMPI. Psychol Assess 2001;13(4):449–51.

20. Fundel, K., Kuffner, R., Aigner, T. and Zimmer, R. (2008) Normalization and gene p-value estimation: issues in microarray data processing, Bioinform Biol Insights, 2, 291-305.

21. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. and Dewey, C.N. (2010) RNA-

Seq gene expression estimation with read mapping uncertainty, Bioinformatics, 26, 493-500.

22. Faulkner, G.J., Forrest, A.R., Chalk, A.M., Schroder, K., Hayashizaki, Y., Carninci, P., Hume, D.A. and Grimmond, S.M. (2008) A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE, Genomics, 91, 281-288.

23. Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S. and Park, P.J. (2005) Discovering statistically significant pathways in expression profiling studies, Proc Natl Acad Sci U S A, 102, 13544-13549.

24. Nam, D. and Kim, S.Y. (2008) Gene-set approach for expression pattern analysis, Brief Bioinform, 9, 189-197.

25. Dinu, I., Potter, J.D., Mueller, T., Liu, Q., Adewale, A.J., Jhangri, G.S., Einecke, G., Famulski, K.S., Halloran, P. and Yasui, Y. (2009) Gene-set analysis and reduction, Brief Bioinform, 10, 24-34.

26. Jiang, Z. and Gentleman, R. (2007) Extensions to gene set enrichment, Bioinformatics, 23, 306-313.

27. Hung, J.H., Whitfield, T.W., Yang, T.H., Hu, Z., Weng, Z. and DeLisi, C. (2010) Identification of functional modules that correlate with phenotypic difference: the influence of network topology, Genome Biol, 11, R23.

28. Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.S., Kim, C.J., Kusanovic, J.P. and Romero, R. (2009) A novel signaling pathway impact analysis, Bioinformatics, 25, 75-82.

29. Rahnenfuhrer, J., Domingues, F.S., Maydt, J. and Lengauer, T. (2004) Calculating the statistical significance of changes in pathway activity from gene expression data, Stat Appl Genet Mol Biol, 3, Article16.

30. Keller, A., Backes, C. and Lenhof, H.P. (2007) Computation of significance scores of unweighted Gene Set Enrichment Analyses, BMC Bioinformatics, 8, 290.

31. Shaffer J. Multiple hypothesis testing: a review. Annu Rev Psychol 1995;46:561–84. 41. Benjamini Y, Hochberg Y. Controlling the false

discovery

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

33. Tang, Y., Ghosal, S. and Roy, A. (2007) Nonparametric bayesian estimation of positive false discovery rates, Biometrics, 63, 1126-1134.

34. Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies, Proc Natl Acad Sci U S A, 100, 9440-9445.

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

36. Khatri, P., Draghici, S., Ostermeier, G.C. and Krawetz, S.A. (2002) Profiling gene expression using onto-express, Genomics, 79, 266-270.

37. Draghici, S., P. Khatri, et al. (2003). Global functional profiling of gene expression. Genomics 81(2): 98-104.

38. Berriz, G.F., King, O.D., Bryant, B., Sander, C. and Roth, F.P. (2003) Characterizing gene sets with FuncAssociate, Bioinformatics, 19, 2502-2504.

39. Beissbarth, T. and Speed, T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes, Bioinformatics, 20, 1464-1465.

40. Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M. and Sherlock, G. (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, Bioinformatics, 20, 3710-3715.

41. Castillo-Davis, C. I. and D. L. Hartl (2003). GeneMerge--post-genomic analysis, data mining, and hypothesis testing. Bioinformatics 19(7): 891-2.

42. Martin D, Brun C, Remy E, Mouren P, Thieffry D, et al. (2004) GOToolBox: functional analysis of gene datasets based on gene ontology. Genome Biol 5: R101. doi: 10.1186/gb-2004-5-12-r101.

43. Doniger, S. W., N. Salomonis, et al. (2003). MAPPFinder: using Gene

Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol 4(1): R7.

44. Khatri P, Draghici S (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 21: 3587–3595. doi: 10.1093/bioinformatics/bti565.

45. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1–13. doi: 10.1093/nar/gkn923.

46. Breitling, R., Amtmann, A. and Herzyk, P. (2004) Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments, BMC Bioinformatics, 5, 34.

47. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucleic Acids Res, 37, 1-13.

48. Mansmann U, Meister R (2005) Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. Methods Inf Med 44: 449–53. doi: 10.1267/METH05030449.

49. Kong, S.W., Pu, W.T. and Park, P.J. (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge, Bioinformatics, 22, 2373-2380.

50. Pavlidis P, Qin J, Arango V, Mann J, Sibille E (2004) Using the Gene Ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. Neurochem Res 29: 1213–1222. doi: 10.1023/B:NERE.0000023608.29741.45.

51. Al-Shahrour, F., Diaz-Uriarte, R. and Dopazo, J. (2005) Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information, Bioinformatics, 21, 2988-2993.

52. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005) Gene set enrichment analysis: a knowledge-based

approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci U S A, 102, 15545-15550.

53. Kim SY, Volsky DJ (2005) PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics 6: 144. doi: 10.1186/1471-2105-6-144.

54. Lu Y, Liu PY, Xiao P, Deng HW (2005) Hotelling's T2 multivariate profiling for detecting differential expression in microarrays. Bioinformatics 21: 3105–3113. doi: 10.1093/bioinformatics/bti496.

55. Xiong H (2006) Non-linear tests for identifying differentially expressed genes or genetic networks. Bioinformatics 22: 919–923. doi: 10.1093/bioinformatics/btl034.

56. Hummel M, Meister R, Mansmann U (2008) GlobalANCOVA: exploration and assessment of gene group effects. Bioinformatics 24: 78–85. doi: 10.1093/bioinformatics/btm531.

57. Klebanov L, Glazko G, Salzman P, Yakovlev A, Xiao Y (2007) A multivariate extension of the gene set enrichment analysis. J Bioinform Comput Biol 5: 1139–1153. doi: 10.1142/S0219720007003041.

58. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34: 267–273. doi: 10.1038/ng1180.

59. Barry, W.T., Nobel, A.B. and Wright, F.A. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach, Bioinformatics, 21, 1943-1949.

60. Witten, D. M. and R. Tibshirani (2008). Testing Significance of Features by Lassoed Principal Components. Ann Appl Stat 2(3): 986-1012.

61. Glazko, G.V. and Emmert-Streib, F. (2009) Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets, Bioinformatics, 25, 2348-2354.

61. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30. doi: 10.1093/nar/28.1.27.

62. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, et al. (2002) The MetaCyc database. Nucleic Acids Res 30: 59–61. doi: 10.1093/nar/30.1.59.

63. Joshi-Tope G, Vasrik I, Gopinath GR, Matthews L, Schmidt E, et al. (2003) The genome knowledgebase: a resource for biologists and bioinformaticists. Cold Spring Harb Symp Quant Biol 68: 237–243. doi: 10.1101/sqb.2003.68.237.

64. Huerta AM, Salgado H, Thieffry D, Collado-Vides J (1998) RegulonDB: a database on transcriptional regulation in Escherichia coli. Nucleic Acids Res 26: 55–59. doi: 10.1093/nar/26.1.55.

65. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13: 2129–2141. doi: 10.1101/gr.772403.

66. Draghici, S., P. Khatri, et al. (2007). A systems biology approach for pathway level analysis. Genome Res 17(10): 1537-45.

67. Khatri P, Drăghici S, Tarca AL, Hassan SS, Romero R (2007) A system biology approach for the steady-state analysis of gene signaling networks. Proc 12th Iberoamerican Congress on Pattern Recognition, CIARP 2007; Valparaiso, Chile.

68. Pavlidis P, Lewis DP, Noble WS (2002) Exploring gene expression data with class scores. Pac Symp Biocomput 7: 474–485.

69. Li KC (2002) Genome-wide coexpression dynamics: theory and application. Proc Natl Acad Sci U S A 99: 16875–16880. doi: 10.1073/pnas.252466999.

70. Shojaie A, Michailidis G (2009) Analysis of gene sets based on the underlying regulatory network. J Comput Biol 16: 407–426. doi: 10.1089/cmb.2008.0081.

71. Bauer-Mehren, A., Furlong, L.I. and Sanz, F. (2009) Pathway databases and tools for their exploitation: benefits, current limitations and challenges, Mol Syst Biol, 5, 290.

72. Dancey, J.E., Bedard, P.L., Onetto, N. and Hudson, T.J. (2012) The genetic basis for cancer treatment decisions, Cell, 148, 409-420.

73. Jones, S.J., Laskin, J., Li, Y.Y., Griffith, O.L., An, J., Bilenky, M., Butterfield, Y.S., Cezard, T., Chuah, E., Corbett, R., Fejes, A.P., Griffith, M., Yee, J., Martin, M., Mayo, M., Melnyk, N., Morin, R.D., Pugh, T.J., Severson, T., Shah, S.P., Sutcliffe, M., Tam, A., Terry, J., Thiessen, N., Thomson, T., Varhol, R., Zeng, T., Zhao, Y., Moore, R.A., Huntsman, D.G., Birol, I., Hirst, M., Holt, R.A. and Marra, M.A. (2010) Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors, Genome Biol, 11, R82.

74. Tian,L. et al. (2005) Discovering statistically significant pathways in expression profiling studies, Proc Natl Acad Sci U S A, 102, 13544-13549.

75. Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D. and Stuart, J.M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM, Bioinformatics, 26, i237-245.

76. Drier, Y., Sheffer, M. and Domany, E. (2013) Pathway-based personalized analysis of cancer, Proc Natl Acad Sci U S A, 110, 6388-6393.

77. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muertter, R.N. and Edgar, R. (2009) NCBI GEO: archive for high-throughput functional genomic data, Nucleic Acids Res, 37, D885-890.

78. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S. and Soboleva, A. (2012) NCBI GEO: archive for functional genomics data sets--update, Nucleic Acids Res, 41, D991-995.

79. Lu,T.P. et al. (2010) Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women, Cancer Epidemiol Biomarkers Prev, 19, 2590-2597.

80. Lu, T.P., Lai, L.C., Tsai, M.H., Chen, P.C., Hsu, C.P., Lee, J.M., Hsiao, C.K.

and Chuang, E.Y. (2011) Integrated analyses of copy number variations and gene expression in lung adenocarcinoma, PLoS One, 6, e24829.

81. Su, L.J., Chang, C.W., Wu, Y.C., Chen, K.C., Lin, C.J., Liang, S.C., Lin, C.H., Whang-Peng, J., Hsu, S.L., Chen, C.H. and Huang, C.Y. (2007) Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme, BMC Genomics, 8, 140.

82. Landi, M.T., Dracheva, T., Rotunno, M., Figueroa, J.D., Liu, H., Dasgupta, A., Mann, F.E., Fukuoka, J., Hames, M., Bergen, A.W., Murphy, S.E., Yang, P., Pesatori, A.C., Consonni, D., Bertazzi, P.A., Wacholder, S., Shih, J.H., Caporaso, N.E. and Jen, J. (2008) Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival, PLoS One, 3, e1651.

83. Beer, D.G., Kardia, S.L., Huang, C.C., Giordano, T.J., Levin, A.M., Misek, D.E., Lin, L., Chen, G., Gharib, T.G., Thomas, D.G., Lizyness, M.L., Kuick, R., Hayasaka, S., Taylor, J.M., Iannettoni, M.D., Orringer, M.B. and Hanash, S. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nat Med, 8, 816-824.

84. Lee, E.S., Son, D.S., Kim, S.H., Lee, J., Jo, J., Han, J., Kim, H., Lee, H.J., Choi, H.Y., Jung, Y., Park, M., Lim, Y.S., Kim, K., Shim, Y., Kim, B.C., Lee, K., Huh, N., Ko, C., Park, K., Lee, J.W., Choi, Y.S. and Kim, J. (2008) Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression, Clin Cancer Res, 14, 7397-7404.

85. Hou, J., Aerts, J., den Hamer, B., van Ijcken, W., den Bakker, M., Riegman, P., van der Leest, C., van der Spek, P., Foekens, J.A., Hoogsteden, H.C., Grosveld, F. and Philipsen, S. (2010) Gene expression-based classification of non-small cell lung carcinomas and survival prediction, PLoS One, 5, e10312.

86. Marisa, L., de Reynies, A., Duval, A., Selves, J., Gaub, M.P., Vescovo, L., Etienne-Grimaldi, M.C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S.,

Chazal, M., Flejou, J.F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., Parc, Y., Olschwang, S., Milano, G., Laurent-Puig, P. and Boige, V. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value, PLoS Med, 10, e1001453.

87. Smith, J.J., Deane, N.G., Wu, F., Merchant, N.B., Zhang, B., Jiang, A., Lu, P., Johnson, J.C., Schmidt, C., Bailey, C.E., Eschrich, S., Kis, C., Levy, S., Washington, M.K., Heslin, M.J., Coffey, R.J., Yeatman, T.J., Shyr, Y. and Beauchamp, R.D. (2010) Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer, Gastroenterology, 138, 958-968.

88. Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P. and Stein, L. (2011) Reactome: a database of reactions, pathways and biological processes, Nucleic Acids Res, 39, D691-697.

89. Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics, 19, 185-193.

90. Lin, Y.H., Friederichs, J., Black, M.A., Mages, J., Rosenberg, R., Guilford, P.J., Phillips, V., Thompson-Fawcett, M., Kasabov, N., Toro, T., Merrie, A.E., van Rij, A., Yoon, H.S., McCall, J.L., Siewert, J.R., Holzmann, B. and Reeve, A.E. (2007) Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer, Clin Cancer Res, 13, 498-507.

91. TCGA (2012) Comprehensive molecular portraits of human breast tumours, Nature, 490, 61-70.

92. Schaefer, C. F., K. Anthony, et al. (2009). PID: the Pathway Interaction Database. Nucleic Acids Res 37(Database issue): D674-9.

93. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets, Nucleic Acids Res, 40, D109-114.

94. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res, 28, 27-30.

95. Ingenuity® Systems, www.ingenuity.com

96. Yu, N., Seo, J., Rho, K., Jang, Y., Park, J., Kim, W.K. and Lee, S. (2011) hiPathDB: a human-integrated pathway database with facile visualization, Nucleic Acids Res, 40, D797-802.

97. Bryant, C.M., Albertus, D.L., Kim, S., Chen, G., Brambilla, C., Guedj, M., Arima, C., Travis, W.D., Yatabe, Y., Takahashi, T., Brambilla, E. and Beer, D.G. (2010) Clinically relevant characterization of lung adenocarcinoma subtypes based on cellular pathways: an international validation study, PLoS One, 5, e11712.

98. Bandres, E., Malumbres, R., Cubedo, E., Honorato, B., Zarate, R., Labarga, A., Gabisu, U., Sola, J.J. and Garcia-Foncillas, J. (2007) A gene signature of 8 genes could identify the risk of recurrence and progression in Dukes' B colon cancer patients, Oncol Rep, 17, 1089-1094.

99. Barrier, A., Boelle, P.Y., Roser, F., Gregg, J., Tse, C., Brault, D., Lacaine, F., Houry, S., Huguier, M., Franc, B., Flahault, A., Lemoine, A. and Dudoit, S. (2006) Stage II colon cancer prognosis prediction by tumor gene expression profiling, J Clin Oncol, 24, 4685-4691.

100. Barrier, A., Roser, F., Boelle, P.Y., Franc, B., Tse, C., Brault, D., Lacaine, F., Houry, S., Callard, P., Penna, C., Debuire, B., Flahault, A., Dudoit, S. and Lemoine, A. (2007) Prognosis of stage II colon cancer by non-neoplastic mucosa gene expression profiling, Oncogene, 26, 2642-2648.

101. Eschrich, S., Yang, I., Bloom, G., Kwong, K.Y., Boulware, D., Cantor, A., Coppola, D., Kruhoffer, M., Aaltonen, L., Orntoft, T.F., Quackenbush, J. and Yeatman, T.J. (2005) Molecular staging for survival prediction of colorectal cancer patients, J Clin Oncol, 23, 3526-3535.

102. Kopetz, S. and Abbruzzese, J.L. (2009) Barriers to Integrating Gene Profiling for Stage II Colon Cancer, Clin Cancer Res, 15, 7451-7452.

103. Wang, Y., Jatkoe, T., Zhang, Y., Mutch, M.G., Talantov, D., Jiang, J., McLeod, H.L. and Atkins, D. (2004) Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer, J Clin Oncol, 22, 1564-1571.

104. Barletta, J.A., Yeap, B.Y. and Chirieac, L.R. (2010) Prognostic significance of grading in lung adenocarcinoma, Cancer, 116, 659-669.

105. Munoz-Pinedo, C., El Mjiyad, N. and Ricci, J.E. (2012) Cancer metabolism: current perspectives and future directions, Cell Death Dis, 3, e248.

106. Lievre, A., Bachet, J.B., Le Corre, D., Boige, V., Landi, B., Emile, J.F., Cote, J.F., Tomasic, G., Penna, C., Ducreux, M., Rougier, P., Penault-Llorca, F. and Laurent-Puig, P. (2006) KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer, Cancer Res, 66, 3992-3995.

107. Gazdar, A.F. (2009) Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors, Oncogene, 28 Suppl 1, S24-31.

108. Mangone, F. R., I. G. Bobrovnitchaia, et al. (2012). PIK3CA exon 20 mutations are associated with poor prognosis in breast cancer patients. Clinics (Sao Paulo) 67(11): 1285-90.

109. Astanehe, A., Arenillas, D., Wasserman, W.W., Leung, P.C., Dunn, S.E., Davies, B.R., Mills, G.B. and Auersperg, N. (2008) Mechanisms underlying p53 regulation of PIK3CA transcription in ovarian surface epithelium and in ovarian cancer, J Cell Sci, 121, 664-674.

110. Reimand, J., Arak, T. and Vilo, J. (2011) g:Profiler--a web server for functional interpretation of gene lists (2011 update), Nucleic Acids Res, 39, W307-315.

111. Reimand, J., Kull, M., Peterson, H., Hansen, J. and Vilo, J. (2007) g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale

experiments, Nucleic Acids Res, 35, W193-200.

112. E. Fox, J. Balko, C. Arteaga. (2013) “Deep kinome sequencing identifies a novel D189Y mutation in the Src family kinase LYN as a possible mediator of antiestrogen resistance in ER+ breast cancer,” Cancer Research, vol. 72, (no. 24), Dec 15 2012

113. Mangone, F.R., Bobrovnitchaia, I.G., Salaorni, S., Manuli, E. and Nagai, M.A. (2012) PIK3CA exon 20 mutations are associated with poor prognosis in breast cancer patients, Clinics (Sao Paulo), 67, 1285-1290.

국문초록 (요약)

유전자 네트웍의 이상을 탐지하는 것은 질병의 기작을 이해하고 나아가 개인의 유전자 결함에 맞춤 치료를 선정하는 일에 중요하다. 현재 존재하는 유전자 조절/생체 대사 경로 분석 알고리즘은 대부분 정상과 대조군 집단에서의 차이를 판별하는 데에 초점이 맞추어져 있다. 이러한 방법은 한 개인에 초점을 맞추어 분석을 하는 용도로는 적합하지 못하다.

한 개인의 유전자 네트웍의 이상을 분석함에 있어 가장 이상적인 방법은 같은 환자의 정상 조직과 질병 조직의 정보를 비교하는 것이다. 하지만, 임상적인 이유에서 환자의 정상 조직의 정보는 항상 가용한 것은 아니다. 정상 조직을 채취 하는 것은 임상적인 위험이 따르며, 특별하고 명확한 이유가 없는 한 권장되지 않는다. 따라서 질병 시료의 개인 맞춤 분석에 있어서, 같은 사람의 정상 조직 정보는 가용하지 않은 경우가 많다. 본 논문에서는 개인 분석이라는 측면과 해당 환자의 정상 조직 정보가 가용하지 않을 때 유전자 네트웍을 분석하는 것에 초점을 맞추었다. 본 논문의 방법의 철학은 한 사람의 암 환자 유전자 정보를 많은 수의 집적된 정상 조직의 유전자 정보와 비교하여 이상 유무를 판단하는 것에 있다.

본 논문은 Over-Representation Analysis (ORA), Functional Class Score (FCS) 등의 기존에 알려진 그룹 대 그룹에서의 유전자 네트웍 분석법의 개인향 분석법을 제공한다. 이 방법을 사용하여 본 논문에서는 개인의 유전자 네트웍 이상 점수 (individualized pathway aberrance score : iPAS)를 제시 한다.

본 논문의 방법을 두가지 종류의 암종 (폐 선암종, 대장암) 유전자 발현 데이터에 적용하여 유용성을 보였다. 페 정상 조직과 대장 점막 정상 조직의 유전자 발현 데이터를 참조 표준으로 삼고, 각 암 환자 한 사람씩의 유전자 네트웍의 이상을 분석 하였다. 본 논문의 방법은 기존의 연구에서 밝혀진 환자 생존률과 관련된 유전자 네트웍 이상을 정확히 탐지 하였다. 본 논문의 방법은 기존에 방법이라고 할 수 있는, 환자 한명의 정보를 해당 환자가 속한 코호트의 정보를 참조 표준으로 사용하여 해석하는 것 보다, 더 높은 재현성을 보였다. 재현성 측정은 서로 다른 데이터군을 사용하여, 유전자 네트웍 발굴군에서 발굴한 생존 관련 유전자 네트웍이, 발굴에 사용되지 않았던 데이터군에서도 생존에 유의한 영향을 미치는지 측정하였다.

또한 해당 방법은 유전자 네트웍의 특징을 기반으로 환자와 정상을 구분할 수 있다. 특별히 ‘amino acid synthesis and interconversion’ pathway 의 경우 폐 선암을 독립적인 검증을 위한 데이터군에서도 AUC 0.982 로 잘 구분할 수 있다. 또한 본 논문에서 제시한 방법은

돌연변이가 유전자 발현 네트웍에 미치는 영향을 정량화 할 수 있는 방법으로 사용될 수 있다. 본 방법을 사용하였을 때 유방암의 유전자 발현 네트웍에 통계적으로 유의한 영향을 미치는 PI3KCA, TP53, RB1 의 세 유전자를 찾을 수 있었고, 이는 알려진 유방암의 지식과 일치한다.

본 논문의 임상적인 의의는 환자 한 사람에서 정상 조직 정보가 없을 때, 한 사람의 암을 유전자 네트웍 측면에서 해석 할 수 있도록 한 것이다. 이러한 방법은 데이터에 기반한 것으로서, 축적되고 있는 정상 조직 데이터를 사용하여, 더욱 정확한 데이터 기반 의사 결정을 하는 데에 기여할 수 있다. 본 논문의 방법은 유전자 발현 뿐 아니라 돌연변이 분석과도 연계되어, 환자의 암을 유발하는 유전자 네트웍을 발굴하고, 맞춤 치료제를 선정하는 일에 기여할 수 있다.

주요어 : 유전자 발현, 개인 유전체 네트워크

학 번 : ２００８－３０１３４