

國立臺灣大學電機資訊學院生醫電子與資訊學研究所

碩士論文

Graduate Institute of Biomedical Electronics and Bioinformatics

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

利用已知基因傳遞機制及蛋白質交互作用圖譜

來開發新的基因互動途徑

Integrate Pathway Information and Protein Interaction Network
to Explore Possible Interactions Between Genes



翁小涵

Siao-Han Wong

指導教授：莊曜宇 博士

Advisor : Dr. Eric Y. Chuang

中華民國九十八年七月

July, 2009

謝誌

本篇論文與研究的完成，仰仗了許多人的鼓勵與幫助。首先要感謝的是碩士班指導教授莊曜宇老師給予我機會加入這個可愛的跨領域研究團隊，在這兩年間提供優良的環境與豐沛的研究資源、藉各式晶片分析訓練帶領我進入生物資訊領域與開展眼界，並常常對我們的行事方向提出建議與樹立典範。再來要感謝賴亮全老師與蔡孟勳老師在平日的數據分析與論文內容修正的辛勤指導。

另外要特別感謝自宏學長從論文题目的啟蒙、內容討論與一路上的帶領，並不斷地把我從牛角尖裡揪出來，確是辛苦萬分的差事！感謝學長姐佩君、子彬永遠不會在我們提出淺顯問題時皺起眉頭，總是專業地回答與指導，並且在論文寫作上提出建議與協助修改；感謝瑞徽學長耐心仔細地教導、亦不時為我們帶來風趣的言語及正向的力量。感謝學長姐方瀚與音秀、常不辭辛勞地為大家訂購午餐飲料或點心，且給予我們窩心的鼓勵與歡樂散布；感謝郁禾學姐，從進實驗室以來的 service 帶領與一路上許許多多熱心的經驗分享，讓我能順利地融入這個環境與這個領域，並且不時地帶給我們溫馨與滿點的歡笑。感謝助理琇瑜、承桓的鼓勵和殷勤地餵食，讓我在實驗室的每一天都充滿驚喜與飽足感；感謝可愛又有趣的學弟妹若陽、榕芝、欣穎這一年來為我們分攤了許多實驗室的事務、使我們得以全力準備論文與研究。感謝兩位同學建樂與建鴻，在這個實驗室同甘共苦了兩年，從一開始籌畫實驗室的活動、修課、學習分析、接工作以及最後一起熬過將論文成形的兩個多月趕捷運的日子。感謝生醫電資所的師長、眾所辦人員在行政事務上的幫助，以及本屆近三十位同學兩年中的扶持與照顧。感謝我身邊的每一位朋友，能認識你們並擁有你們的陪伴是十分幸運亦很幸福的事。

最後，感謝我的家人們一直是我精神上的支柱，給予我不斷往前的動力。沒有你們，就沒有今日與明日的我。

小涵 2009.7.29 基因體中心

摘要

近年來，基因晶片 (microarray) 的數據分析已經從傳統上僅利用統計分析的方法，轉向加入更多已知基因功能性註解來協助分析。生物反應路徑分析法 (pathway analysis) 針對資料庫中每一個已定義好的反應路徑 (pathway)，測量於實驗設計中其是否存在訊息核糖核酸 (mRNA) 層級上的顯著變異。而另外一種網絡圖譜分析法 (network analysis) 則旨在搜尋一個由基因間所有可能存在的交互作用所組成的圖譜 (global interaction network)，看其中是否有顯著改變的子圖譜 (subnetwork)。這兩種分析方法各擅勝場，且可望補足對方的缺點：前者僅分析已知的反應路徑，故將結果局限在熟知的生物知識中；後者雖蘊藏了許多可能的基因互動途徑，但直接從 global interaction network 中搜尋容易找到無法以現有生物知識呈現其一致性生物意義的 subnetwork。

本文提出一個能基於 pathway analysis 但更進一步結合 network analysis 優點的分析方式來改善現有的分析方式，值得一提的是此一結合概念在目前的分析領域中並不常見。此方法首先利用 Tian *et al.* 發展的 pathway analysis 方式測定有顯著變異的 pathway。接著以這些 pathway 的成員作為出發點，採用 Nacu *et al.* 的 network analysis 方法進行一個目的導向式的 subnetwork 搜尋。

本篇論文將此方法應用在台大醫院肺癌病人的基因晶片數據上。一開始嘗試在有變異的 pathway 中尋找其最具代表性的成分 (subnetwork)。這組數據產生的 subnetwork 在另一組台北榮總肺癌病人的數據中亦得到了呈現高度一致的結果。此外我們針對找到的 subnetwork 進行其成員基因的功能性分析，發現從原本完整 pathway 縮減到 subnetwork 的過程中，整體的功能由原本具有的多樣性，專一化到特定的功能上。這暗示了該 pathway 的某部分功能在實驗中是明顯地被改變的，而這樣的改變得以用這篇論文的方法被察覺。更進一步，我們展現了本方法承繼了

network analysis 而來的優點。立基於 pathway 的已知成員去搜尋其可能有互動的鄰近基因，我們得到的 subnetwork 是以此 pathway 為出發點，但包含許多未被認識的交互作用，這樣的結果可以協助研究者對於未知的部分設計實驗做更深入的探究。從另外一方面來說，這樣的結果也有異於傳統 network analysis 的方式，它使得研究可以立基於研究者感興趣的 pathway，基於已知的生物知識去拓展相關未知的可能性。

總結所有分析的結果，它們從不同方向指出了此分析方法不同於以往的許多優點：它可以從顯著改變的 pathway 中萃取出一個最重要且大小適合研究的 subnetwork、也可以針對研究者感興趣的 pathway 或特定的調控機制進行主題式的深入探討、此外除了立基於原有生物知識外，它亦具有開發基因間新的互動機制的的能力。

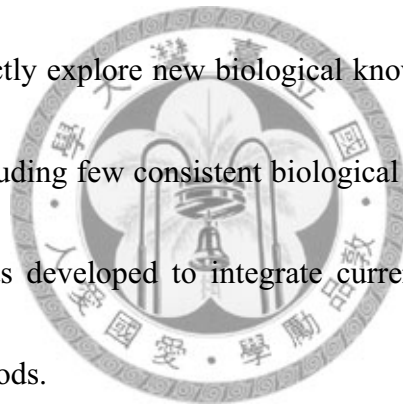


關鍵詞

基因晶片、數據分析

ABSTRACT

Currently the analysis of microarray data had turned into integrating with prior biological knowledge: pathway analysis interprets transcriptomic data on pathway level and identified predefined groups of genes with dysregulation; network analysis takes gene-gene interactions information into consideration and searches for modules associated to the phenotypes under study. The two analyses have its own advantages respectively and they complement the weaknesses of each other: pathway analysis provides little clues to directly explore new biological knowledge and network analysis usually yields modules including few consistent biological information. In this study an analytical methodology was developed to integrate current pathway analysis method with network analysis methods.



Initially, dysregulated pathways are identified by modified pathway analysis method in Tian *et al.*. Subsequently, a focus-oriented investigation on dysregulated pathways are performed by network analysis following the work of Nacu *et al.*, and this step is using modules within or related to members of the pathways to be further investigated. Several improvements were made, such as the scoring functions and the module identification algorithms.

To illustrate the benefits of this methodology, a lung cancer study with 30 paired

cancer and normal tissues was explored. The results derived within dysregulated pathways were also identified consistently in another public dataset GSE7670. Furthermore, GO term enrichment analysis was applied to show that the modules have a specialized functionality than the original pathways. In brief, original large modules were reduced from the entire pathway to a smaller size of relevant interconnected members, which are much easier to be manipulated but still remain their biological information. Moreover, the ability of this methodology to explore novel interactions related to pathway members were also demonstrated by extending the module search algorithm beyond the pre-defined pathways. This would not be achieved by traditional pathway analysis methods, which usually don't include biomolecular interaction information. Yet, modules identified in this methodology were based on dysregulated pathways with specific biological meaning since their members were mainly associated.

In conclusion, these data all indicated the advantages to integrate both pathway and network information during microarray analysis: to uncover manageable size of molecular interaction networks important for pathway dysregulation, to focus on interested pathways, functions or even specific regulatory events, and to possess the potential of performing exploratory researches on mechanisms that are not yet well understood. Undoubtedly, this concept could be extensively applied to other array

experiments of similar design regardless of the disease under study.

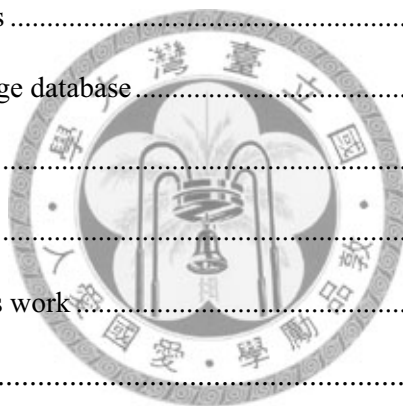
Key words

microarray, pathway analysis, network analysis, module



CONTENTS

口試委員會審定書	I
謝誌	II
摘要	III
ABSTRACT	V
Chapter 1 Introduction	1
1.1 Lung cancer	1
1.2 Microarray	2
1.3 Data analysis	2
1.3.1 Single gene analysis	3
1.3.2 Biological knowledge database	4
1.3.3 Pathway analysis	8
1.3.4 Network analysis	11
1.3.5 Methodology in this work	13
Chapter 2 Materials	17
2.1 Lung cancer datasets	17
2.2 Databases	17
2.3 Program environment and public/ commercial tools	19
Chapter 3 Methods	22
3.1 Database construction	22
3.2 Single gene analysis	24
3.3 Pathway analysis	25
3.4 Network analysis	27
3.5 Results demonstration	30
Chapter 4 Results	31
4.1 Database	31



4.2 Single gene analysis	32
4.3 Pathway analysis	34
4.4 Network analysis - within pathways	36
4.5 Result demonstration and comparison	39
4.5.1 Mapping the main component and leading edge subset on KEGG pathways.....	39
4.5.2 GO term enrichment analysis.....	41
4.6 Network analysis – protruding pathways	48
Chapter 5 Discussion	50
6.1 Input matrix creation.....	50
6.2 Pathway analysis	51
6.3 Network analysis.....	55
6.4 Future perspectives.....	60
Chapter 6 Conclusions	63
REFERENCES	66
APPENDIX	70



LIST OF FIGURES

Figure 3-1. Flow chart of this methodology	22
Figure 3-2. Entity Relationship Diagram (ERD) for database constructed here	22
Figure 4-1. Overlaps between array, pathway and interaction data	31
Figure 4-3. t-score distribution of probe sets before and after representative filtering .	33
Figure 4-4. Histograms of unadjusted p-values for 560 pathways in database	35
Figure 4-5. Main components extracted from two pathways	38
Figure 4-6. Map the main component and leading edge subset to original pathway	40
Figure 4-7. GO terms (cellular component 5) enriched in focal adhesion pathway.....	44
Figure 4-8. GO terms (molecular function 1-5) enriched in focal adhesion pathway...	45

Figure 4-9. GO terms (biological process 5) enriched in cell cycle pathway.....	46
Figure 4-10. GO terms (cellular component 5) enriched in cell cycle pathway.....	47
Figure 4-11. Extend subnetwork search to the global interaction network	49
Figure A-1. Top cancer killers in Taiwan in 2008	70
Figure A-2. The most significant subnetwork identified by GXNA	71
Figure A-3. Main components obtained by f_2 scoring method.....	76
Figure A-4. GO term hierarchy for cluster C in Fig. 4-6.....	79
Figure A-5. GO term hierarchy for terms involved in Fig. 4-7	80
Figure A-6. GO term hierarchy for cluster A and term B,C in Fig. 4-8.....	81
Figure A-7. GO term hierarchy for cluster C,D,E in Fig. 4-9	82
Figure A-8. Histogram of pathway sizes in our database.....	83

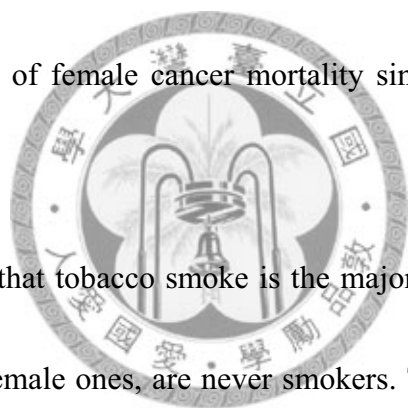
LIST OF TABLES

Table 1-1. Definition of gene sets and exemplary databases.....	5
Table 4-1. Statistics of current database.....	31
Table 4-2. Network analysis procedure and the consistency with GSE7670	36
Table A-1. Statistics of t-scores in two datasets	70
Table A-2. Parameters set during pathway analysis	72
Table A-3. Significant pathways identified by different methods.....	72
Table A-4. Lists of significant pathways identified by f_0 and f_1 scoring function.....	73
Table A-5. Annotations of genes present in Figure 4-5.....	74
Table A-6. Annotations of genes present in Figure A-3.....	77
Table A-7. Overlap of main components and the leading edge subset.....	78
Table A-8. Annotations of genes present in Figure 4-11.....	83

Chapter 1 Introduction

1.1 Lung cancer

According to figures released by Department of Health in June 2009, cancer has topped the major 10 causes of death in Taiwan for 27 consecutive years. The statistics showed that, in 2008, malignant tumors were responsible for 27.3 percent of all deaths, among which the proportions of top cancer killers are displayed in Figure A-1. Moreover, when gender is taken into consideration, lung cancer, in particular, has occupied the leading cause of female cancer mortality since it first overtook cervical cancer in 1986 [1].

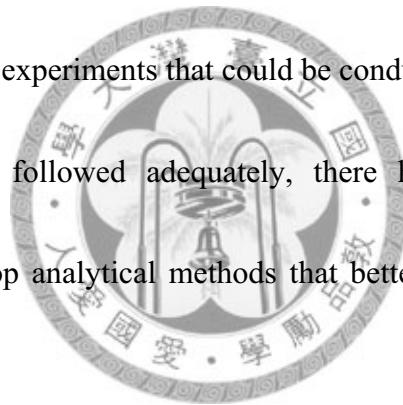


Noteworthy, despite that tobacco smoke is the major risk factor for lung cancer, many patients, especially female ones, are never smokers. This situation is not unusual in countries other than Taiwan, as it had already been discussed in the review paper [2]. It was suggested in the article that pathways of carcinogenesis for lung cancer in never smokers and tobacco-associated lung cancer are not exactly the same due to the clinical and biological differences observed in patients of the two types. However, specific mechanisms are still under investigation, which in further requires good experimental techniques and analytical methodologies that help researchers focus on clues to carcinogenesis process with efficiency.

1.2 Microarray

Microarray is a powerful tool to screen tens of thousands of genes at one time, giving a semi-quantitative sketch of genome-wide mRNA expression levels in cells, which greatly facilitates and accelerates biological studies. Since its invention in the 1990s [3], gradually improved technologies had lead to more affordable commercial arrays with stable quality, making microarray widely applied in various biomedical researches and beyond question cancer-related studies are no exceptions.

However, unlike array experiments that could be conducted with acceptable quality as long as protocols are followed adequately, there has always been room for bioinformaticians to develop analytical methods that better extract biological insights from array data.



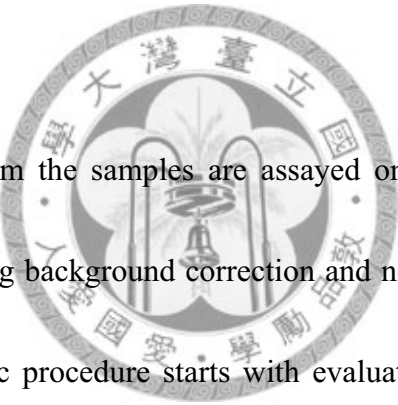
Following this section some methods will be reviewed and in the end of this chapter, the idea about methodology developed here will be introduced.

1.3 Data analysis

Starting from introducing regularly employed single gene analysis that concentrates on independent statistical analysis of individual genes, what followed subsequently are some advanced data analysis methods that take additional biological information into account. Each of these methods detects biological processes being

dysregulated from different entry points, however, major concerns of biologists can yet be satisfied simultaneously: to uncover manageable size of molecular interaction networks important for dysregulation, to focus on interested pathways, functions or even specific regulatory events, and to possess the potential of performing exploratory researches on mechanisms that are not yet well understood. In the end of this chapter, an idea about combining advantages of current methods will be introduced in attempt to fulfill these requirements.

1.3.1 Single gene analysis

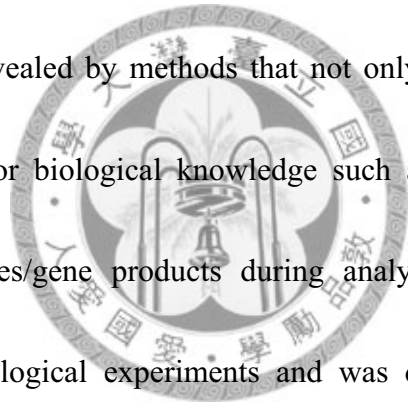


Whatever array platform the samples are assayed on, after image scanning and preprocessing steps including background correction and normalization within/between arrays, conventional analytic procedure starts with evaluation of individual genes: all genes are ordered by the extent of their associations with phenotypes, with a significant degree of association suggesting the gene's being differentially expressed at mRNA level and worthwhile to undergo further biological validations.

However, varied according to topics under study, this approach sometimes ends up with a long list of significant genes even after multiple hypothesis adjustments [4, 5] that are usually required when testing large amount of hypotheses simultaneously. This makes follow-up validations laborious or even infeasible. Such obstacle could not be

surmounted by mere selection from the significant gene list, where the decision of genes deserving further investigations depends on researchers' expertise due to multiple roles a gene might play.

To deal with this problem, analysis shall not be confined to expression data itself anymore. As has been well-known to all, cellular functions are not implemented by individual genes independently, rather, they are accomplished by a group of genes acting together to perform cellular tasks. Thus it is anticipated that a more consistent biological scene can be revealed by methods that not only incorporate transcriptomic data but also consider prior biological knowledge such as functions in common or relationships between genes/gene products during analysis. Such information was derived from previous biological experiments and was deposited in various public databases as described below.



1.3.2 Biological knowledge database

It gains more insight into the interpretation of transcriptomic data if analysis could be integrated with functional annotations or other omic data from different levels. To see how this can be realized, in this section and the next we will introduce what kind of information can be utilized and how they can be integrated.

A. Gene set databases

Generally speaking, a gene set is a group of functionally related genes. Specific instances for gene set definitions and corresponding databases are tabulated in Table 1-1.

Since members in a gene set tend to function in coordination, several methods are developed to analyze genes in groups and identify gene sets instead of genes that are significantly regulated.

Table 1-1. Definition of gene sets and exemplary databases

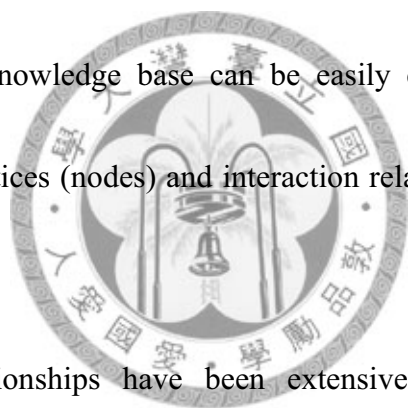
Common feature	Databases
metabolic / signaling pathway member	KEGG, Biocarta, GenMapp, MSigDB (c2)
chromosomal location / cytogenetic band	MSigDB (c1)
target of microRNA / transcription factor	MSigDB (c3)
biological process participants	GO - Biological Process
subcellular location/macromolecular complex	GO - Cellular Component
perform molecular function	GO - Molecular Function

※ Kyoto Encyclopedia of Genes and Genomes (KEGG) [6], BioCarta [7], GenMapp [8], MSigDB [9], Gene Ontology (GO) [10]

In this study, we focus only on metabolic and signaling pathways that are essentially an abstraction of the information flow through physical interaction network in response to a drug, nutrients or external stimuli.

B. Interaction databases

An interaction knowledge base contains genetic [11] and physical interactions between genes/gene products of various types. They can be related to physical binding (protein complex), protein modification (methylation, (de)phosphorylation), promoter binding (transcriptional regulation) or chemical reaction (activation/inhibition). Typically, they are stored as binary interactions and in the form of *gene(product)1-relation-gene(product)2* triplets, from which a global biomolecular network visualizing this knowledge base can be easily constructed by representing genes/gene products as vertices (nodes) and interaction relationships as edges (directed or undirected).



Although these relationships have been extensively studied in small-scaled experiments using synthetic lethality or other biochemical and biophysical techniques, they are recorded in scientific literatures and cannot be directly utilized by computational scientists unless undergone information extraction into machine-readable format. Currently, this can be done by manual curations or by text mining techniques such as applying natural language processing algorithm [12].

The construction of such databases was triggered actually by the expansion of high-throughput techniques in the last ten years, which includes: microarray for

synexpression of genes; yeast two-hybrid (Y2H) technology, tandem affinity purification coupled with mass spectrometry (TAP-MS), high-throughput mass-spectrometric protein complex identification (HMS-PCI) and co-immunoprecipitation (Co-IP) for protein-protein interactions; chromatin immunoprecipitation coupled with DNA microarray (ChIP-chip) or with paired-end ditag (ChIP-PET), DNA adenine methylase identification (DamID) and yeast one-hybrid assays for protein-DNA interactions. The boosted amount of formatted data from aforementioned techniques had received wide attention of bioinformaticians and enabled the development of this field. At present, these data account for more than 70% of current database content. Nonetheless, scientific literatures remain to be the most important and reliable source since a high percentage of high-throughput data were estimated to be spurious [13, 14]. On the other hand, it is also believed that these so called false positives are in fact true physical interactions, yet might not be biologically meaningful [15].

To sum up, a pool of all known biomolecular interactions between genes/gene products are accommodated in public interaction databases such as BIND [16], HPRD [17], MINT [18] and commercial knowledge bases held by Ingenuity Systems (Ingenuity Pathway Analysis, IPA) and GeneGo (Metacore).

At last, a clear difference between network and pathway can be elucidated by this paragraph [15] :

“A network represents a static image of all possible physical and/or regulatory interactions between biological entities, while a pathway represents how the information propagates through the network. Because information propagation is a directional process, a pathway must have entry nodes where the information flow starts and terminal points where the information flow ends.”

Several methods evolved with the aid of these knowledge bases and they will be reviewed in the next two sections. These methods differ mainly in the databases incorporated, however, what as well cannot be left out of consideration are the statistical methodology they utilized and the extent they exploit transcriptomic data.

1.3.3 Pathway analysis

Pathway analysis is actually gene set analysis only to focus on pathways. It is expected to not only find pathways with significant differential expression but also detect consistent yet subtle expression changes among members of a pathway. A review on various pathway analysis methods evaluating the involvement of pathways in different phenotypes under study is available in [19, 20].

A simplest procedure is to assess the significance of overlap between preselected

genes (differentially expressed genes) and predefined annotation groups (pathways) using Fisher Exact test (hypergeometric distribution) or Chi-squared test, which is then followed by adequate multiple hypothesis adjustments. This over-representation method is intuitive, easy to implement and computationally efficient, thus it dominates current commercial and public software, such as IPA [21], Metacore [22] and DAVID [23].

An alternative approach needs no prior filtering of genes. It assigns a score to each gene set and assesses p -value by re-sampling procedure, which is to compare the score with its null distribution. BRB-ArrayTools [24] use the LS statistic and Kolmogorov-Smirnov (K-S) statistic to test if the single-gene p -values in a gene set are of a uniform distribution. In gene set enrichment analysis (GSEA) [25, 26], an enrichment score is obtained by considering the distribution of pathway genes in the entire list of genes, which in spirit is a weighted K-S statistic. Tian *et al.* [27] designed a statistical framework to determine perturbed pathways. In their work the overall objective is to “*test whether a group of genes has a coordinated association with a phenotype of interest.*” After each gene set is assigned a score by averaging the test statistics of its member genes, p -values regarding two different hypotheses are estimated by permuting class labels and gene orders. Eventually, gene sets with significant p -values under both hypotheses are considered differentially expressed

across phenotypes.

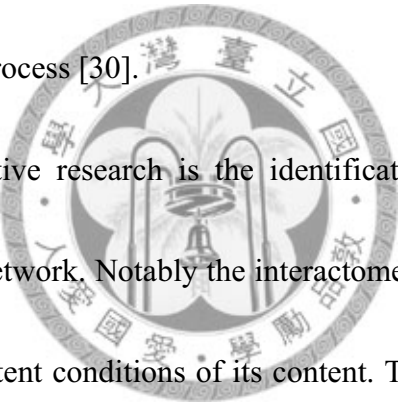
The latter approach is more biologically reasonable than over-representation method since it reserve expression information and does not depend on an artificial filtering of genes, in which the results may depend strongly on the cutoff chosen to make the significant gene list [28].

However, most pathway analysis methods suffer from some weaknesses. First, due to the fact that only a limited number of well-studied genes are classified into pathways, the remained many genes with unknown functions are left out of considerations. Second, most pathway annotations contain only labels of pathway members and with no information about the interplay between them, therefore the results obtained from these methods could not help to give direct understanding of cellular processes at molecular level. Also, pathway is actually a dynamic model and the component activated is specific to conditions under study, which is usually not emphasized in most methods, however, GSEA did specify a list of core members, named leading edge subset, that are the main contributors of the pathway's enrichment score.

In all, pathway analysis can successfully interpret data at pathway level but fails to elucidate the interplay between members within and provide no information about genes not involved in known pathways.

1.3.4 Network analysis

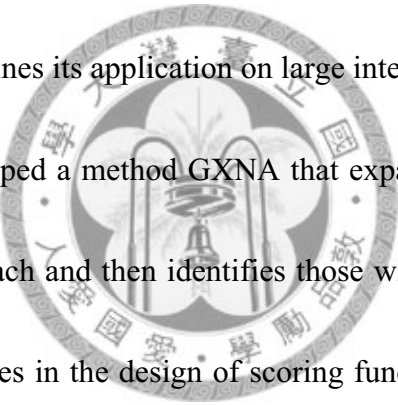
Extensive works had been done seeking to characterize the design principles of biomolecular network structure using graph theory. An interesting finding shows that its connectivity distribution follows the typical power-law distribution for scale-free network and it shows small-world properties [29] in terms of network diameter and clustering. It is believed that this feature, which indicates the existence of hubs, enables the biological network to be robust against occasional removal of arbitrary network elements during evolution process [30].



Another subject of active research is the identification of relevant modules or subnetworks in the global network. Notably the interactome data cannot alone complete this task due to the inconsistent conditions of its content. That is, these interactions are usually temporal, spatial or dependent on conditions and tissue types. Therefore, additional annotations providing condition-specific information is required, such as combining microarray data to identify connected sets of nodes based on their coherent expression patterns at mRNA level. In this regard, several methods of identifying responsive modules are reviewed in [31].

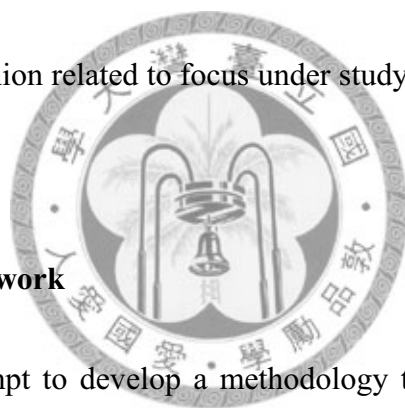
Ideker *et al.* [32] are among the first groups to extract active subnetworks based on both interactome and transcriptomic data. They searched the global network for

high-scoring subnetworks via simulated annealing algorithm, where the score is in spirit an aggregation of z-scores measuring differential expression of individual genes in the subnetwork and calibrated against the null distribution generated by randomly sampled groups of genes. Interestingly, the result contained many examples of genes individually with low score but are required to connect together several high-scoring genes. Genes with such character agree with the behavior of some transcription factors (TFs) and will be referred to as “key nodes” in this article. Despite the exciting observation, the time-consuming nature confines its application on large interaction network.



Nacu *et al.* [33] developed a method GXNA that expands subnetworks from seed nodes using a greedy approach and then identifies those with significantly high scores. One of their contributions lies in the design of scoring functions that correct for biases due to subnetwork size if necessary. They are generally divided into two scoring schemes $-\sum T$ and $T \sum$. What subsequent to the evaluation of subnetwork score is the estimation of p -value that assesses its significance against the null distribution derived from scores of subnetworks under different phenotype permutations. Eventually, these p -values are adjusted for family-wise error rate and used to select relevant subnetwork. GXNA is fast and it focuses on small modules differentially expressed between phenotypes, however, it allows no such key nodes since the greedy approach is adopted.

In general, pathway can be viewed as the assembly of causal sequences of molecular events and with all building blocks available in the biomolecular network. Ideally it seems promising to infer new pathways directly by network analysis, but in reality such pathway inference are complex and less validated due to the noise and incompleteness of current biological networks, especially when coupling with the small-world property of its topology. Additionally, the identified module is prone to be associated with multiple canonical pathways, which makes it hard to reveal a unifying scene and interpret in a fashion related to focus under study, as in the example of Figure A-2.



1.3.5 Methodology in this work

In this study we attempt to develop a methodology that goes a step further than current pathway analysis by bringing in the advantages of network analysis. Practically speaking, it can be achieved by processing microarray data through pathway and network analyses in series, however, it is worthwhile mentioning that the incorporation of both pathway and network knowledge into microarray data analysis is yet to be widely realized. Based on results from pathway analysis, the methodology here aims to enhance researchers' understanding by stepping from pathway level into molecular level, with its main objectives specified below:

1. Extract module most relevant to the pathway's dysregulation.

It is without doubt that pathway is in fact a dynamic model and the members perturbed differs between conditions under study. Although GSEA specifies a leading edge subset and suggests their being responsible for pathway's dysregulation, the subset is of limited biological meaning since it provides no information for elucidating the roles they play. To put it simply, the leading edge subset is statistically meaningful more than biologically meaningful. Therefore, methodology here attempt to complement pathway analysis in this regard by making use of biomolecular networks.

After dysregulated pathways are identified, concept of network analysis is then imposed to extract modules most relevant to dysregulation of interested pathways. It is mainly due to these sequential events that make the pathways deemed dysregulated.

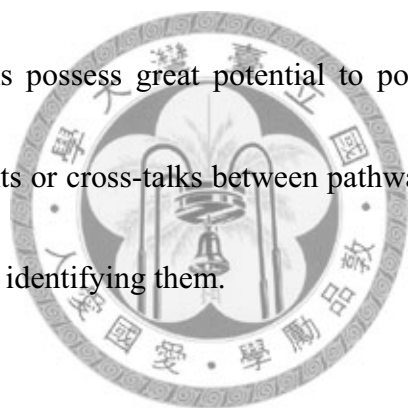
On the other hand, the module obtained here differs from that yielded by other network analyses in its ability to interpret microarray data at known-pathway level and to investigate interested pathways in greater details.

2. Other focus-oriented strategies enabling exploratory survey on pathways of interest.

Current understanding of biological functions at the pathway level is far from being thorough. In one way, some of known pathways are actually incomplete. In the other, although recorded in a separate fashion, they are not truly isolated at the level of

global interaction network. In fact they should not be so as in the example of cross-pathway inhibition, which allows an activated pathway to refocus cellular resources to respond to stimuli it received by competing against other pathways. As a matter of fact, a considerable degree of cross-talks between pathways are revealed by various small-scaled biological studies.

Using the methodology here, a task-oriented investigation is promised by exploiting the valuable information imbedded in biomolecular networks. These yet well-understood interactions possess great potential to point an investigator to either missing pathway components or cross-talks between pathways and help in the design of appropriate experiments for identifying them.



In previous study physical interaction data were integrated with genetic interaction information to uncover the mechanisms underneath [34], which can further be used to find cross-talks between two pathways where the two interacting genes reside.

The interaction data was also coupled with pathway information to provide clues of cross-talks between pathways [35], however, it is done by merely calculating whether physical connections between two pathways are higher than random using Fisher Exact Test.

In two successive works of Ideker *et al.* [36, 37], subnetwork and condition-responsive genes (CORGs) were defined in dysregulated pathway as the part delivering optimal discriminative power for the disease phenotype using $T \Sigma$ evaluation. Nonetheless, they were used as prediction markers rather than to discuss the underlying mechanisms.

Before going to the next chapter, the methodology framework is summarized in four parts.

1. Construction of database for storage of molecular interaction network, canonical pathway collections and gene annotations.
2. Statistical analysis of microarray data at pathway level: in this part Tian's algorithm is adopted yet with slight but crucial modification.
3. Algorithms allow navigating the global network on the region of interested pathways: the scoring function is based on GXNA and modified to tolerate key nodes. In addition, a merging step is developed to complement the scoring function.
4. Visualize the module by laying genes/gene products according to their subcellular localizations.

Chapter 2 Materials

2.1 Lung cancer datasets

1. NTUH Lung Cancer Dataset

This dataset is created by Bioinformatics and Biostatistics Core, National Taiwan University Research Center for Medical Excellence - Division of Genomic Medicine.

Matched normal and tumor samples of 31 female patients with non-small cell lung cancer (NSCLC) in National Taiwan University Hospital (NTUH) are collected for DNA microarray analysis using Affymetrix Human Genome U133 Plus 2.0 Array.

2. Public Dataset GSE7670 (TVGH Lung Cancer dataset)

A public dataset [38] created by Taipei Veterans General Hospital (TVGH) and deposited in NCBI's Gene Expression Omnibus (GEO) [39] (also available at EBI's ArrayExpress: www.ebi.ac.uk/arrayexpress) under GEO series accession number GSE7670 is downloaded as CEL files. Of all the 66 Affymetrix Human Genome U133A Array data, those from 21 female NSCLC patients with paired normal-tumor arrays are used to create this dataset.

2.2 Databases

1. Pathways

Predefined gene sets performing tasks of metabolic functions or signaling

transductions are recorded in several public databases.

The Molecular Signature Database (MSigDB) v2.5 maintained by Broad Institute collects curated gene sets information from several online databases or literatures and records them with a unified format. Its “Canonical Pathway (CP) collection” in “Functional Sets (C2) category” [9] is the main source of pathway information in this work. The collection of 639 gene sets is released in a file with filename extension “gmt”.

In the following work, gene sets are simplified into pathways.

2. Protein interaction network

Protein-protein interactions (PPI) detected by high-throughput methods are recorded in several PPI databases. The NCBI's Entrez Gene database containing curated interaction information from BIND [16], BioGRID [40], EcoCyc [41] and HPRD [17] is utilized as the main source to construct protein interaction network in this work.

3. Target genes of probe sets on microarray

The information of genes targeted by probe sets designed on Affymetrix GeneChip arrays comes from annotation files in Affymetrix website (www.affymetrix.com).

4. Gene annotation

Unique gene identifiers (Entrez ID or HUGO gene symbol) and historical aliases of genes are retrieved from NCBI's Entrez Gene database.

2.3 Program environment and public/ commercial tools

With database built under MySQL, this methodology was realized with the help of Matlab[®]. Besides, some public or commercial tools are used to process the dataset for different purposes. Following are some brief descriptions about them.

1. Partek[®] Genomics Suite [42]

It is a commercial software that enables various statistical analysis of microarray data. In the work here it is used simply to complete preprocessing steps of microarray data, which summarizes expression value for each probe set and applied normalization algorithm to remove potential systematic biases.

2. Gene set enrichment analysis (GSEA) [25]

GSEA evaluates the probability a gene set is differentially expressed across phenotypes and defines a leading-edge subset comprising the core members activated in the gene set.

At first, genes are ordered by their correlation between expression values and phenotype classes, then an enrichment score (ES), corresponding to a weighted Kolmogorov-Smirno-like statistic, is calculated for the gene set. A p -value representing significance level of the ES score is determined against null distribution of ES estimated by permuting class labels. After p -value for each gene sets is obtained, false discovery

rate (FDR) method is used to adjust for multiple hypothesis testing.

3. Database for Annotation, Visualization and Integrated Discovery (DAVID) [23]

DAVID uses an EASE score, a modified Fisher Exact p -value, to measure the enrichment of gene sets in the gene list specified by user. Furthermore, to reduce the redundant nature of annotations that might dilute the focus of the result, which is especially inevitable when associating with Gene Ontology (GO) terms, DAVID provides the option to classify significantly associated gene sets into different clusters and order the clusters according to their significance. According to manual on the website, it is achieved by integrating the same techniques of Kappa statistics to measure the degree of the common genes between two gene sets, and fuzzy heuristic clustering to classify the groups of similar annotations according to kappa values.

4. Cytoscape [43]

Cytoscape is a JAVA application which provides basic functionality to layout and query networks, overlay nodes with expression data, or link genes/gene products to databases of functional annotation. Notably, it is featured in its extensibility through a straightforward plug-in architecture, which enables additional computational analyses to be incorporated. In this study, a plug-in - Cerebral v2.0 [44] is used to layout the network according to the subcellular location of each node.

5. Kyoto Encyclopedia of Genes and Genomes (KEGG) - Color Objects in KEGG Pathways [45]

It is an on-line tool that provides a personalized pathway map by allowing the assignment of different colors to font, border or background of each particular node.

6. European Bioinformatics Institute (EBI) - QuickGO [46]

QuickGO is a web-based browser that enables the extraction of branches from the entire hierarchy of Gene Ontology according to a list of specified GO terms.



Chapter 3 Methods

Figure 3-1 illustrates the flow chart of this methodology and in the subsequent sections we will describe each step in detail.

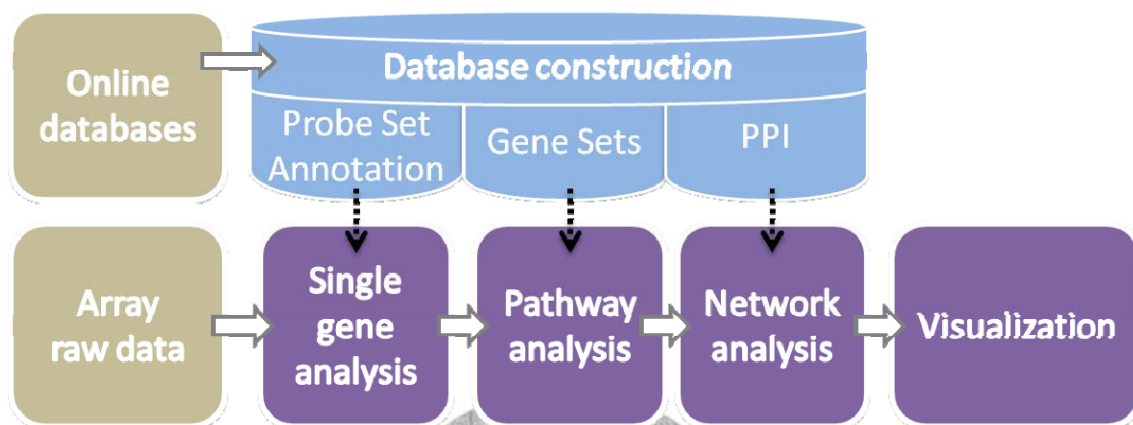


Figure 3-1. Flow chart of methodology in this work.

PPI is the abbreviation of protein-protein interaction data.

3.1 Database construction

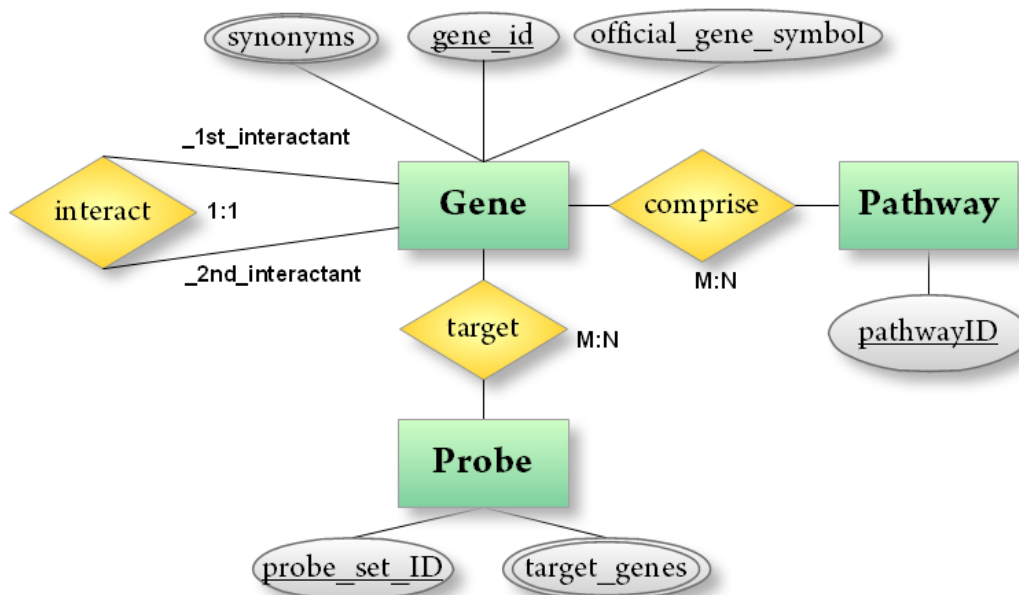


Figure 3-2. Entity Relationship Diagram (ERD) for database constructed here.

In this ERD rectangles stand for entities, ellipses for attributes and diamonds for relationships.

As the first step of integrating genomic data at different levels, a relational database is required in order to bridge between information from various sources, to speed up data retrieval and to correct for ambiguously recorded data. The structure of database constructed here is elucidated in Figure 3-2 as a simplified entity relationship diagram, where genes, pathways and array probes are considered as different entity types.

While constructing the database, some records with ambiguous information should be corrected:

1. Importing gene sets from MSigDB.

The gmt file records gene set members in the form of “synonym”, which is alias instead of official name. The task here is to convert each of them into corresponding unique identifier: “gene_id”. This is done by comparing them to “official gene symbol”, or to “synonym” of genes in the case when there were no “official gene symbols” matched.

2. Importing target genes of each probe set on microarray.

Some “gene_id” recorded in NetAffx annotation files had been changed or discontinued. New “gene_id” should be updated based on information from NCBI’s Entrez database.

3.2 Single gene analysis

1. Probe sets filtering

As one might note in chapter 2, the two datasets were assayed on different versions of Affymetrix GeneChip array. In fact, U133 Plus 2.0 comprises probe sets in both U133A and U133B [47], thus in following works only those probe sets common in both versions are utilized. Note that this step could be skipped when datasets to be compared are of the same version.

2. Summarizing expression value for each gene

The analysis of Affymetrix array data starts with CEL files recording fluorescence signals at probe level, from which probe-set level intensities are derived using robust multi-chip average (RMA) method [48]. In this method probe level data undergo background correction, quantile normalization [49] and median-polish summarization [50]. The RMA process is completed under the commercial software Partek[®] [42] and the resulting values are log-transformed expression values.

3. Hypothesis testing

Hypothesis testing methods are used to measure the degree of association between response/covariate (either numerical or categorical factor, *e.g.* phenotypes) and random variables (expression levels of probe sets/genes). In the simple but most common case,

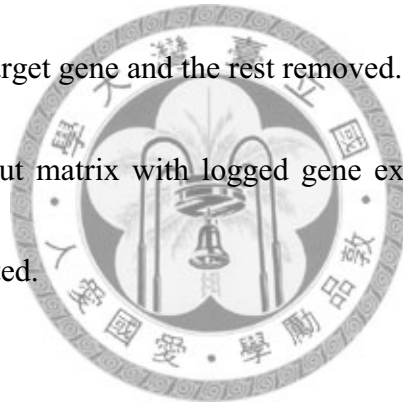
given a covariate, its association with each gene is estimated by univariate hypothesis testing method such as two-sample t test or Mann-Whitney statistics for binary phenotype, F -statistic for polytomous phenotype, to name but a few.

Here two-sample paired t test is applied and the concluded p -value functions as an index of degree of differential expression in terms of a probe set between phenotypes.

4. Select representative probe set for each gene

Among all probe sets targeting the same gene, the one with the smallest p -value is selected to represent their target gene and the rest removed.

Until this step, an input matrix with logged gene expression values in rows and arrays in columns is generated.



3.3 Pathway analysis

Both Tian method and modified Tian method are applied on the datasets. Conceptually the procedure of pathway analysis starts with evaluating a pathway score by employing a scoring function, which is the major difference between the two methods. The score is then to be normalized and assigned with p -value according to its null distribution that could be generated in two different ways of permutation. Finally pathways are ordered by the addition of rankings under two permutation types. The detailed procedures of both methods are described below.

1. Scoring function

For each pathway $S_m = \{g_1, \dots, g_{k_m}\}$, its score is calculated by either

$$f_0(S_m) = \frac{1}{k_m} \sum_{i=1}^{k_m} T_{g_i} \quad (\text{Tian's method})$$

or

$$f_1(S_m) = \frac{1}{k_m} \sum_{i=1}^{k_m} |T_{g_i}| \quad (\text{modified Tian's method}),$$

where k_m is the size of S_m and T_{g_i} the t statistic of gene i . To note, the latter scoring function comes from equation 2 in Nacu *et al.*[33].

2. Significant level of the score

A one/two-sided p -value representing significance of a pathway is estimated from the f_1/f_0 score's null distribution that could be generated in different ways depending on the null hypothesis to be tested. Tian *et al.* [27] proposed two ways to choose from: either to test if genes inside a set show significantly higher associations with phenotypes than that outside a set, or to test if a set does contain genes differentially expressed between phenotypes. The former is achieved by permuting members of a pathway, and the latter by randomly shuffling phenotype labels on each paired samples.

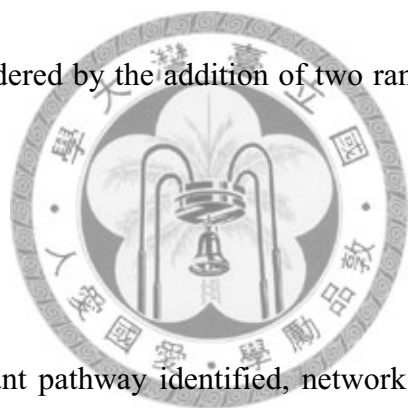
Since all pathways are tested simultaneously, multiple testing problems can no longer be ignored. The p -values are either adjusted by Bonferroni method [51] or converted to q values [52].

3. Pathway score normalization

To overcome the hurdle that pathway scores are not able to be compared directly due to its dependency on the size and unique correlation structure of each pathway, Tian proposed that normalization of the observed scores can be achieved by replacing them with their quantiles. Here f_0 and f_1 scores are normalized in the same principle.

4. Ranking the pathways

Under each permutation procedure, an adjusted p -value and a normalized score are obtained and from which a ranking is summarized. With descending importance, all pathways are eventually ordered by the addition of two rankings derived from separate permutation procedures.



3.4 Network analysis

Based on the significant pathway identified, network analysis tries to investigate the pathway, looking for connected subgraphs that are either essential for differential expression of the pathway or related to genes outside the defined pathway boundary. In short, a candidate subnetwork is generated from each root and then being merged into several main components. This algorithm follows Ideker's idea and some methods of GXNA, and will be described dividedly in six steps.

1. Starting points and the search space

In the beginning, a background interaction network is constructed and afterward

referred to as the “search space”, within which the algorithm searches for main components. As the objective here is to show the applicability of this approach before any further explorative investigations, the search space is at first confined to genes within the dysregulated pathways. After then a version of searching under global interaction network was demonstrated.

Suppose there are N genes (nodes) in the pathway, each of them will be considered as a root and thus N candidate subnetwork would be generated.

2. Extension

Starting from a root node, a candidate subnetwork is generated by an assigned number of extensions within the search space. In each time of extension, the node yielding maximal score of the new subnetwork is incorporated from those directly neighboring the current subnetwork.

3. Scoring

Two ways of evaluating current subnetwork are adopted. One is identical to $f_1(S)$, only when applied here, S represents a subnetwork instead of a pathway. The other, $f_2(S)$, is similar but with slight modifications in order to increase the tolerance of key nodes mentioned in chapter 1.

Considering a subnetwork of size k , the algorithm first rearrange its members in

ascending order of p -values, so that for $S = \{g_1, \dots, g_k\}$, $|T_{g_i}| \leq |T_{g_{i-1}}|$ for all $i \in \{2..k\}$,

where T_{g_i} is the t statistic for gene i . The score of S is then obtained by

$$f_2(S) = \frac{1}{k-m} \sum_{i=1}^{k-m} |T_{g_i}|,$$

where the setting of m is flexible to users (default “1”). The equation suggests that the m members with least contribution would be left aside from the scoring function.

Note the default ΣT scoring function in GXNA (eq.6) [33] is not applied here, for more details please refer to chapter 6.

4. Stopping criterion

There are two criteria in GXNA for stopping the extension of a subnetwork. One is when predefined size is met and the other is when the new subnetwork score does not surpass the current one. The former criterion is used here due to the same reason that f_2 score is dependent on subnetwork size and thus not comparable to each other. Nonetheless, to make up for artificial restrictions in fixed-size search, a merging step is developed to produce subgraphs of different sizes.

5. Merging

Until this step, a candidate pool has been formed by the N candidate subnetworks derived from the N roots. The merging process is a decisive step. It ends up with at most h main components as final results where h is a user-specified parameter.

For each $i \in \{1..h\}$ the merging process starts a main component C_i with an empty set \emptyset . Step by step, the algorithm merges it with the highest-scored candidate subnetwork sharing overlap with it. Ever since a candidate subnetwork has been chosen from the candidate pool and merged with C_i , it is excluded from the pool. The merging process stops when certain criteria are met, which varies depending on the user's concern. Here a handleable size of main component within the pathway is to be found, so the algorithm stops when it reaches an amount approximately r percent of the search space size, or, stops at predefined min/max size in the case of small/large search space. However, the whole process could break off anytime the candidate pool is emptied.

6. Visualization

Main components found in this methodology are visualized using Cytoscape [43].

Moreover, they can also be mapped on the pathway figure using KEGG's online tool [45] when the significant pathway is retrieved from KEGG database.

3.5 Results demonstration

In the end, attempts were made to reveal the biological scenes underlying the results of this methodology by associating Gene Ontology terms with members of main components. This is done with the help of DAVID [23] for GO terms association and clustering, and QuickGO [46] for GO hierarchy visualization.

Chapter 4 Results

4.1 Database

Table 4-1 lists some characteristics of the database constructed here, and informative relationships of its content data are shown in Figure 4-1.

Table 4-1. Statistics of current database

Current Database	Number of Records
NCBI - Entrez Gene database	40234
signaling / metabolic pathways	691
biomolecular interactions	31340
genes with interaction information (A)	8787
genes involved in pathways (B)	5596
genes targeted by U133A probe sets (C)	13799

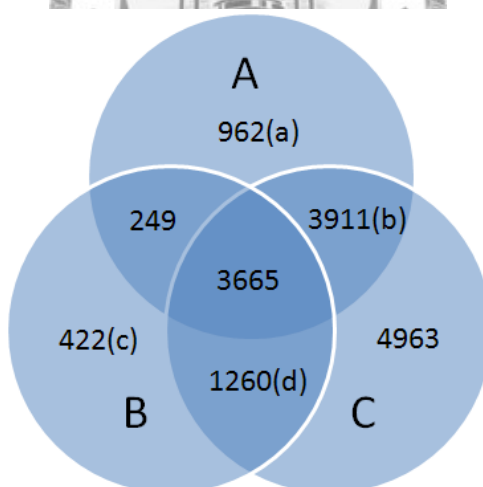


Figure 4-1. It shows overlaps between array, pathway and interaction data.

Notably there are a large number ($a+b=4,873$) of genes within the global interaction network that are currently not assigned to any predefined pathways, which implies the potential to exploit functions of unknown genes when integrating interaction information. On the other hand, there are also 1,682 ($c+d$) genes in pathways that show

no related interaction information, which foretells some obstacles to be met in this methodology. However, this could be overcome by constructing a database that includes more complete interaction data, although it may take lots of time to manually record them from pathway databases and scientific literatures.

4.2 Single gene analysis

In this step, probe level data in CEL files are transformed into an input matrix which will be fed into pathway/network analysis. Figure 4-2 describes the single gene analysis procedure and Figure 4-3 shows the *t*-score distribution of probe sets before and after representative filtering. When a Bonferroni adjusted *p*-value < 0.05 criterion was applied, there were 1,489/1,345 significantly up-/down- regulated probe sets in NTUH dataset after representative filtering, which was much greater than that in GSE7670 dataset as shown in the lower panel in Figure 4-3.

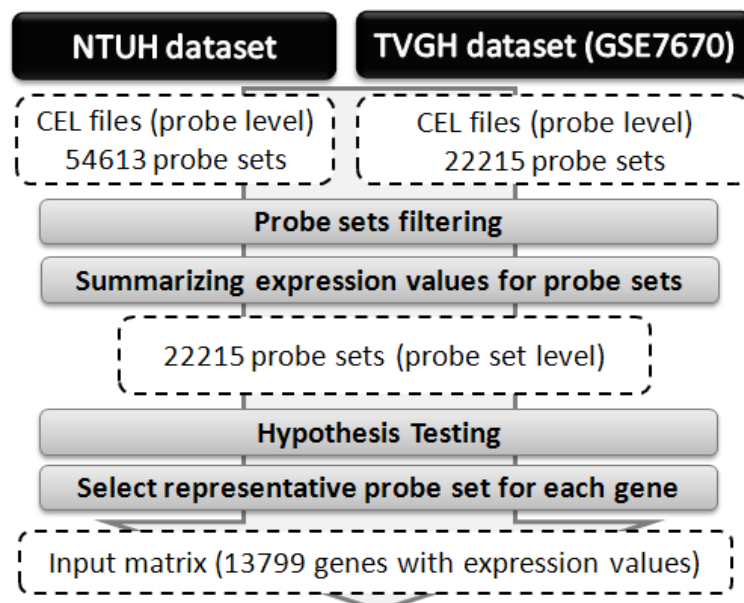


Figure 4-2. Single gene analysis procedure. *It describes how the input matrix was produced from CEL files.*

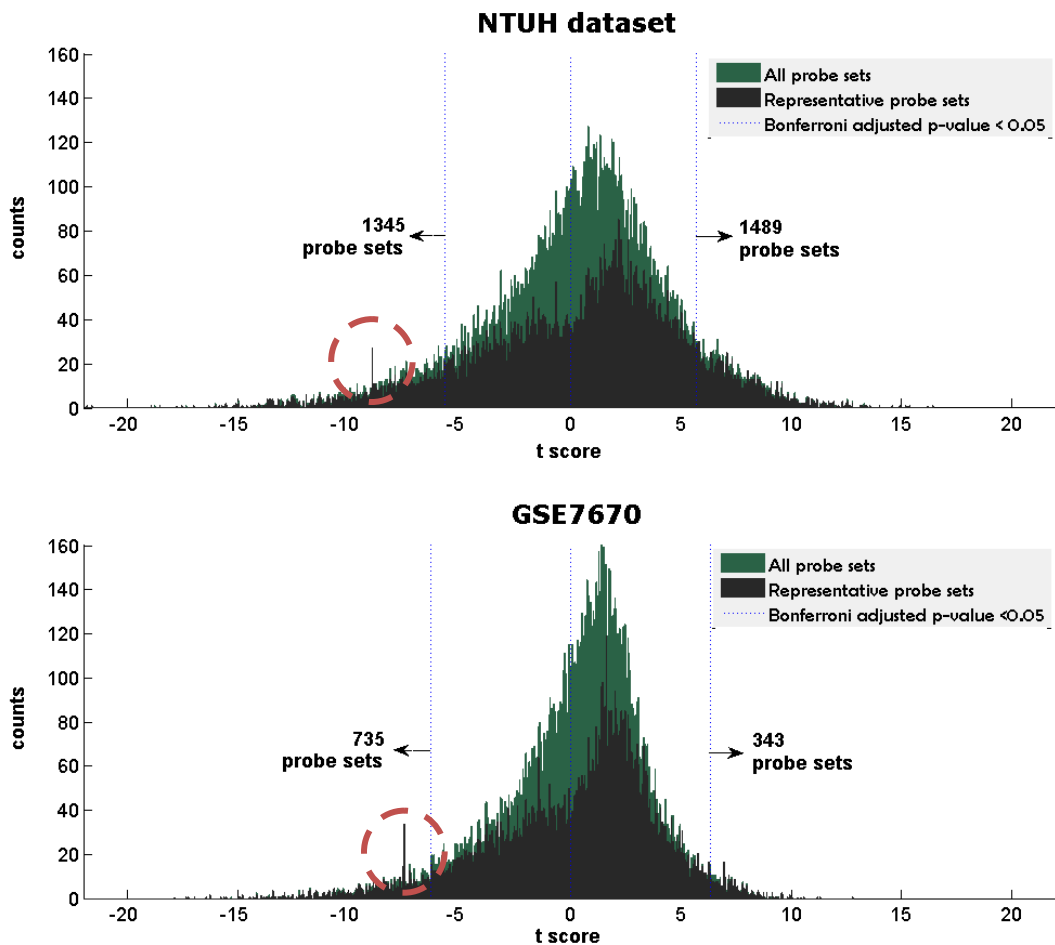


Figure 4-3. t-score distribution of probe sets before and after representative filtering.
The total probe set number decreases from 22215 to 13799. Probe sets with t-scores greater/smaller than the Bonferroni thresholds are significantly up-/down-regulated.

In both datasets the original left-skewed distribution turned into a bimodal distribution after filtering, and in the bimodal distribution the positive and negative sides show unequal peak heights and variances. This might be due to the different correlation structures between genes. More detailed statistics of *t*-score distribution is shown in Table A-1. In addition, an unusual peak highlighted in the Figure 4-3 comes from a nonspecific high-scoring probe set 209079_x_at that targets at 22 related genes.

4.3 Pathway analysis

Three pathway analysis methods were applied on the datasets, namely Tian scoring method (using scoring function f_0), modified Tian scoring method (using scoring function f_1) and GSEA, and the third one was applied only for result comparisons.

A total of 560 pathway information with size ranging from 10 to 500 were used for pathway analysis; each pathway has corresponding null distributions generated by permuting either gene order or class label for 10,000 times. The unadjusted p -values of the 560 pathways were corrected for multiple hypotheses test using Bonfferoni correction or q -value conversion in f_0 scoring function and f_1 scoring function, respectively. Detailed parameter settings in each method were provided in Table A-2.

Histograms of unadjusted p -values for 560 pathways under different permutation types or scoring schemes were displayed in Figure 4-4 and among them a number of pathways were deemed significant under both scoring functions. This phenomenon was different from that in GSEA (Table A-3) and suggested an elevated statistical power in the methods applied here. Detailed significant pathways with adjusted p -value passing the criteria of 0.05 in both datasets were listed in Table A-3,A-4.

After all pathways had undergone pathway score evaluation and p -value adjustment, network analysis was applied further to investigate molecular mechanisms

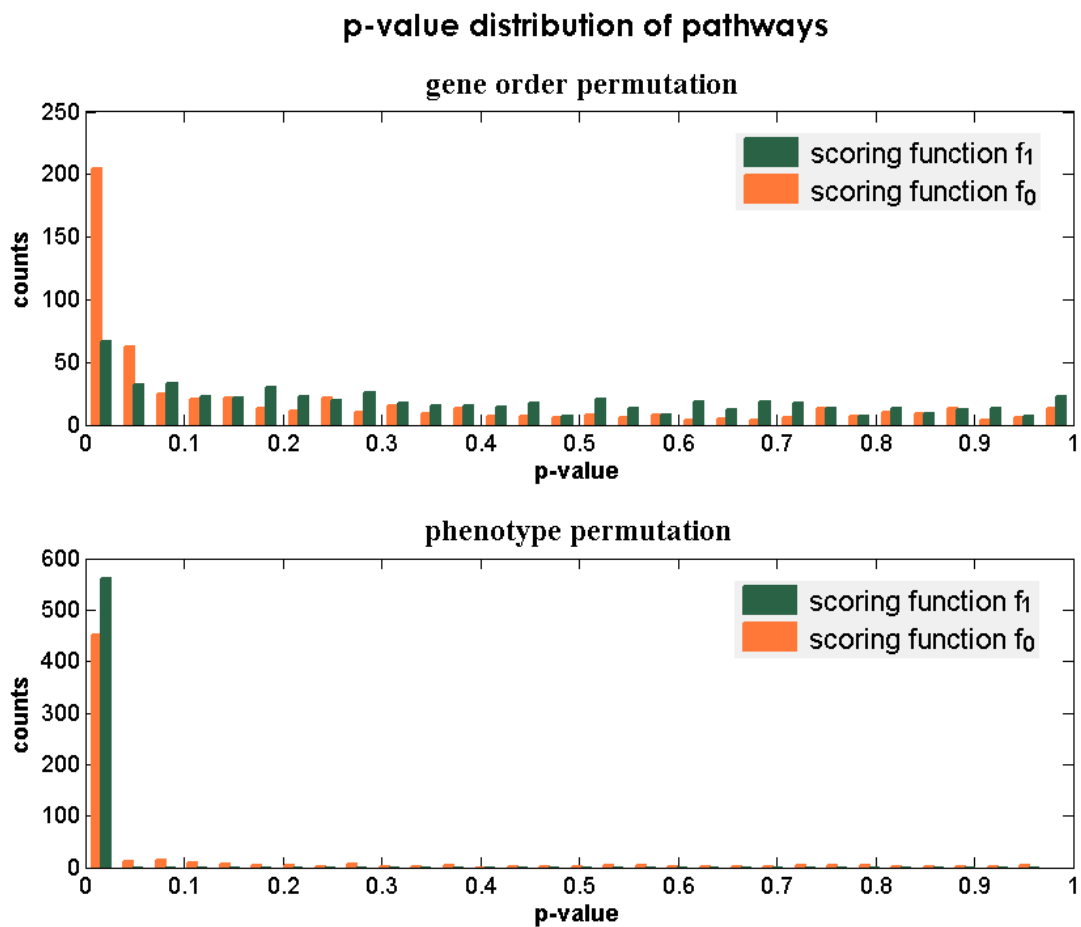


Figure 4-4. Histograms of unadjusted p-values for 560 pathways in database.

Pathway p-values in the upper panel were derived by permuting gene orders and the lower ones by permuting phenotypes. Bars in green color stand for the result when applying f_1 scoring function and that in orange color come from result of f_0 scoring function.

based on the interested pathways. These two scoring methods identified significant pathways for further analysis. For example, “Cell cycle” pathway was identified by f_0 scoring, whereas “Focal adhesion” pathway was selected by the f_1 scoring method. In principle, cancer-associated pathways with larger pathway size were chosen out for examples from those with adjusted p -value lower than 0.05 in both datasets.

4.4 Network analysis - within pathways

In this and the next section, network analysis was applied on significant pathways, and the focus will be on extracting main components from pathways that take the interplay of member genes into consideration.

Within the search space containing all nodes in the pathway, candidate subnetworks each with size 8 were formed using f_1 scoring. They were then merged into one main component with size basically 0.75 times the space size, or bounded by the size of 15, 25 for small, large search spaces respectively. The whole process was summarized in Table 4-2.

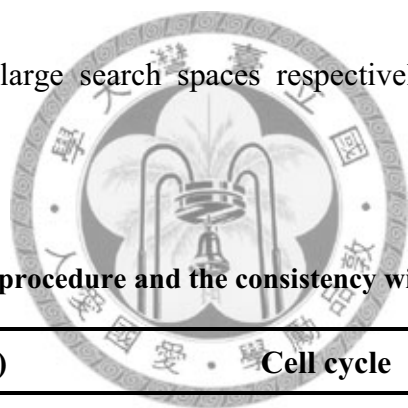


Table 4-2. Network analysis procedure and the consistency with GSE7670.

NTUH (m=0)	Cell cycle	Focal adhesion
search space (entire pathway)	90	199
candidate subnetworks	60	111
top subnetworks merged	10	10
1 st main component size	19	25
leading-edge subset size	35	63
consistency with GSE7670	74%	60%
Fisher's exact test p-value	9.4E-08	4.6E-10

This table showed that the search spaces of two pathways were dramatically reduced from 90 and 199 to 19 and 25 of their 1st main components, respectively.

Furthermore, the 1st main component sizes were also smaller than the sizes of leading

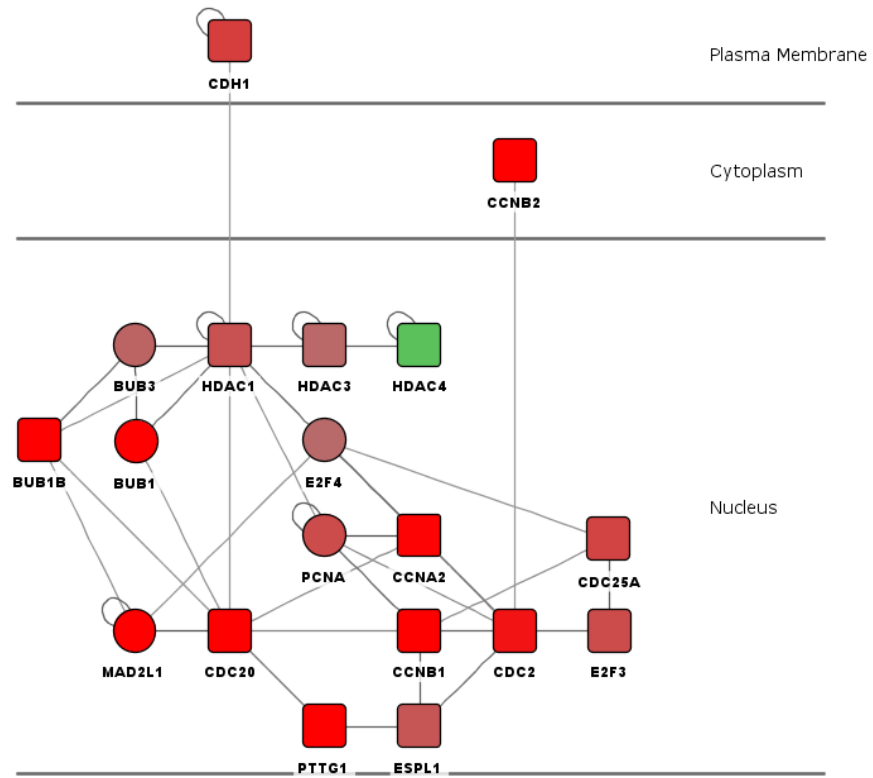
edge subsets obtained by GSEA that equaled to 35 and 63.

To examine the robustness of this method, members of main components derived from NTUH dataset were compared with that obtained from GSE7670 dataset under same analysis procedures. The result of comparison was also shown in the table that the 1st main component from NTUH dataset had a significant overlap (74% and 60% overlap and the overlaps were all with p -value $\ll 0.05$) with that derived from GSE7670 dataset.

These main components obtained by network analysis are displayed in Figure 4-5, where gene/gene products are arranged according to their cellular locations. For detailed expression level and annotations of each gene in Figure 4-5, please refer to Table A-5.

Evidently the major difference between main components obtained from the two pathways is that a coherently up-regulation in tumor samples is observed in Figure 4-5A, while such coherence does not appear in Figure 4-5B. This observation once again reveals the main difference of Tian method and modified Tian method: the former emphasizes pathways with moderate but concordant changes and the latter focuses on pathways with significant degree of overall changes, regardless of up- or down-regulation.

(A)



(B)

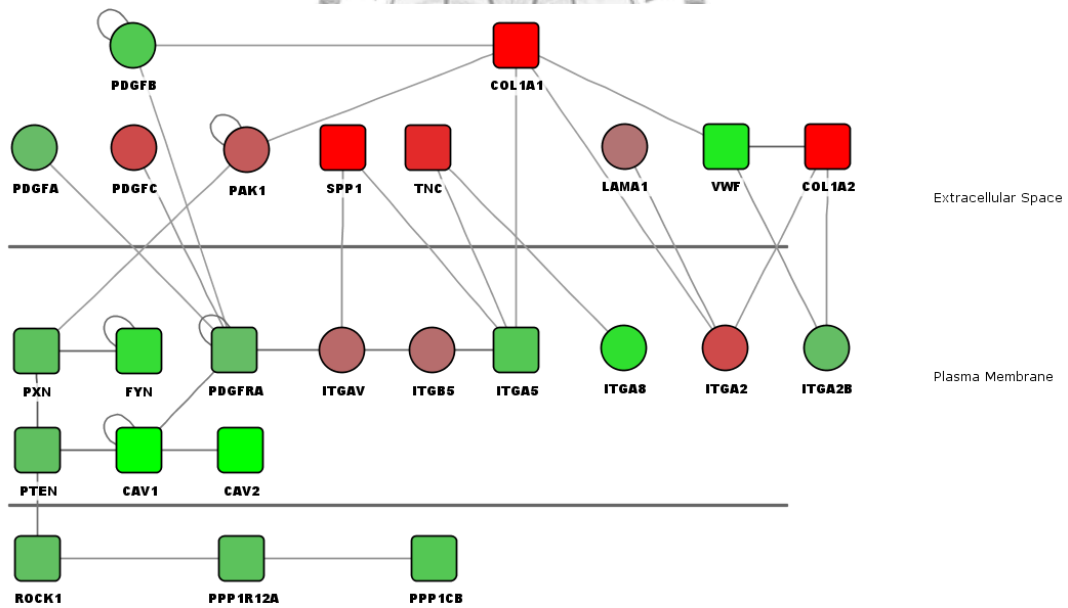
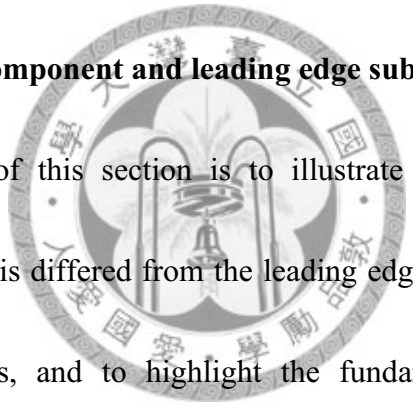


Figure 4-5. Main components extracted from two pathways. After network analysis, main components were derived from (A) cell cycle pathway and (B) focal adhesion pathway. Genes were arranged by their subcellular locations and colored by expression fold change values (red for up-regulation and green for down-regulation in cancer phenotype). Rectangles indicate that genes were found in GSE7670 and NTUH dataset in common, whereas circles mean that genes were identified only in NTUH dataset.

In addition, we also applied network analysis using f_2 scoring with $m=1$ and the rest settings remained unchanged on the two pathways. The main components yielded by this analysis share approximately 70-90 percent similarities with previous ones. Some nodes playing essential roles of maintaining the integrity of main components did exist, however, Figure A-3 and Table A-6 indicate that an expected situation of catching clues of key nodes bridging separately connected subgraphs did not show apparently.

4.5 Result demonstration and comparison

4.5.1 Mapping the main component and leading edge subset on KEGG pathways



Two main purposes of this section is to illustrate how the main component obtained by network analysis differed from the leading edge subset obtained by GSEA from biological viewpoints, and to highlight the fundamental difference between information stored in pathway databases and interaction databases, thus to present the benefit it brought to integrate both information with microarray data.

Figure 4-6 was produced by an on-line tool provided by KEGG [45], where members of leading edge subset and main component for each pathway were simultaneously mapped onto simplified pathway figures. It seemed that most members of main components overlap with leading edge subsets, however, they actually share an overlap of about 60-80 percent as summarized in details in Table A-7.

In the simplified figure, rectangles represent genes, complexes or families, etc. in (A) cell cycle pathway and (B) focal adhesion pathway. Members of main components were filled in red/green colors standing for up-/down-regulation in cancer phenotype. Members of leading edge subset were colored in grey. Note that those nodes with black border and filled with color were common members of both sets, while those with white borders were found in main components only. In addition, dashed red lines were added artificially implying interaction relationship not recorded in KEGG databases.

In Figure 4-6, the members of the leading edge subsets were colored in grey.

Although these genes show statistical significances, they located in the pathway in a scattered manner. In contrast, main components are tightly interconnected subset of genes because it took topology into consideration. It represents the most significant module in the pathway, which would be more biologically meaningful when compared to the leading edge subset that only takes gene significances into account.

Furthermore, the dashed red lines represent possible interactions which do not appear in the predefined pathway may illustrate the additional information one would get when incorporating interaction data with pathway information. The benefit of bringing in the additional information would become more valuable later in section 4-6.

4.5.2 GO term enrichment analysis

In this section, we try to manifest the biological meanings of these main components. It was achieved by using DAVID to compare the GO terms enriched in both the entire pathway and the main component extracted from the pathway. As

mentioned in section 2.3, DAVID evaluates the randomness of each GO term being associated with a user-specified gene list and assigns it with both an EASE score and a false discovery rate (FDR) reflecting the significance of the EASE score. After that, similar GO terms were sorted into a cluster and gave the cluster a new enriched score by summarizing member term EASE scores. This enriched score was then used to rank relative importance of clusters.

DAVID was used to separately evaluate GO terms/clusters associated with gene lists containing the entire pathway and the main component. Then the GO terms or clusters that were contained terms with significant association ($FDR < 0.05$) with the gene list were visualized by a pie chart. Each portion in the pie chart stands for a GO term/cluster, and the proportion of gene list members involved in this cluster/term is showed as a percentage. In addition, the rankings of relative importance of clusters are specified on the pie chart. Note that different GO term/cluster may probably contain overlap information, so the overall percentage in the pie chart would not be exactly 100. Figure 4-7, 4-8 display the result obtained in focal adhesion pathway and Figure 4-9, 4-10 for that in cell cycle pathway.

In Figure 4-7, GO terms/clusters enriched in focal adhesion was identified in terms of cellular component category. It was obvious that Cluster C turned to occupy a larger

proportion in the main component than in the entire pathway. It suggested that when reducing the list of genes from the entire pathway to the main component, the functions that the list of genes able to play had gradually specialized. It might also imply that certain functions are more differentially regulated than other functions in the original pathway, and that these important functions could be revealed by methodology. Furthermore, when member terms of Cluster C were individually considered in Figure 4-7C, each of them showed a consistent trend of such function specialization in both leading edge subset and main component. In contrast, this specialization was not observed in a random subset that was randomly chosen from the entire pathway and with the same size of the main component. In terms of molecular function category in Gene Ontology, several significant GO terms not grouped into clusters (Term B-F) also showed this consistent trend of function specialization in Figure 4-8C.

Such trends did not appear only in focal adhesion pathway. In Figure 4-9 cell cycle pathway was analyzed using GO biological process terms, where Cluster A showed also a dramatic increase from sharing 36% of the entire pathway to sharing 63% of the main component. When analyzing cell cycle pathway using GO cellular component in Figure 4-10, the function specialization also existed since the proportion Cluster C,D,E occupied were all amplified in Figure 4-10B.

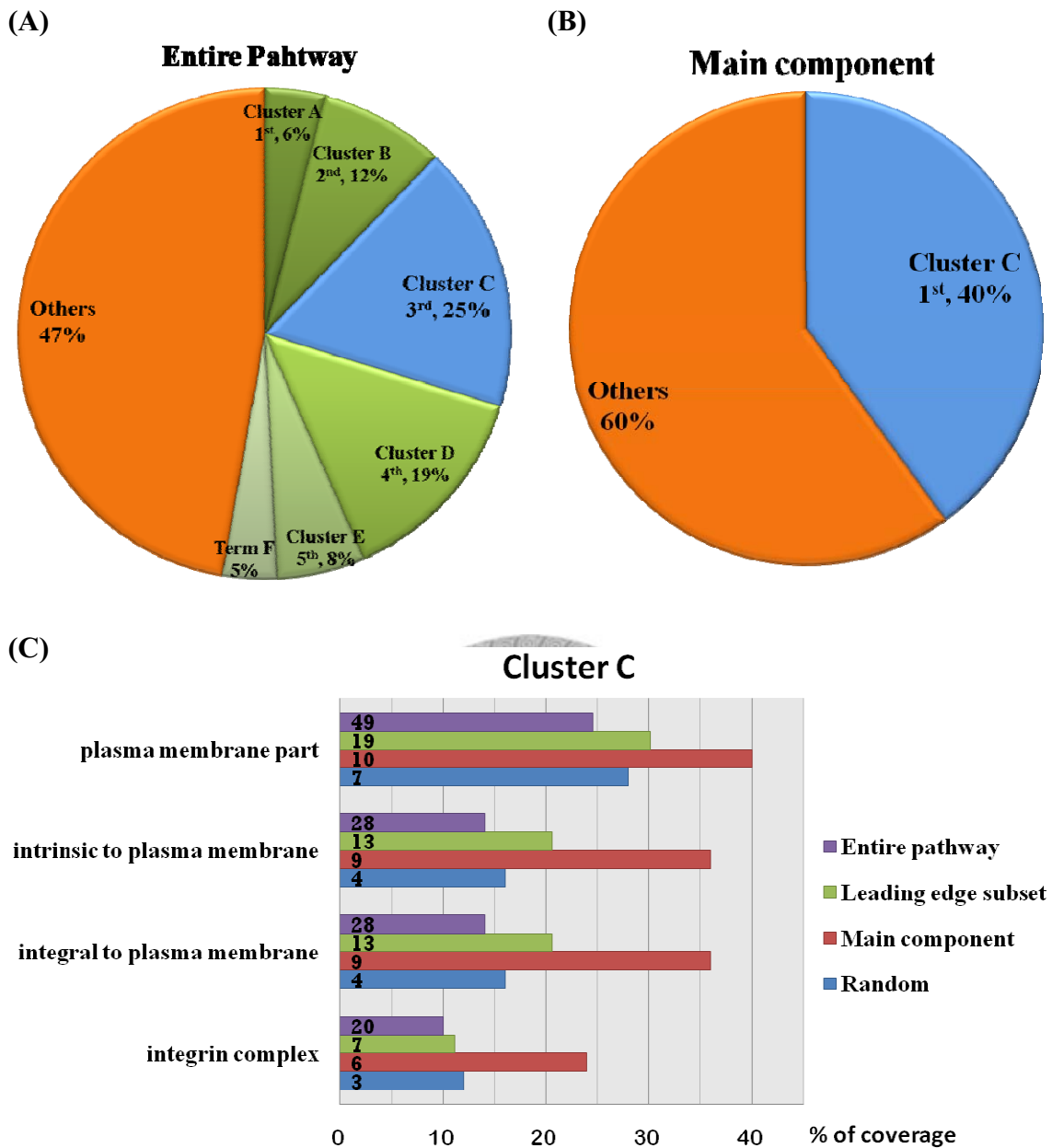


Figure 4-7. GO terms (cellular component 5) enriched in focal adhesion. Pie charts were used to illustrate how GO clusters share the members of (A) the original pathway and (B) the main component, where the rankings represent relative significances of clusters. Terms in Cluster C were listed in (C) and their hierarchical relationships are available in Figure A-4. In (C), genes involved in each term accounted for different percentages in the four sets and the percentages were illustrated as a bar chart with the digit on the bar specifying the actual number of genes involved in this term. The random set was randomly selected from the original pathway and with the same size as the main component.

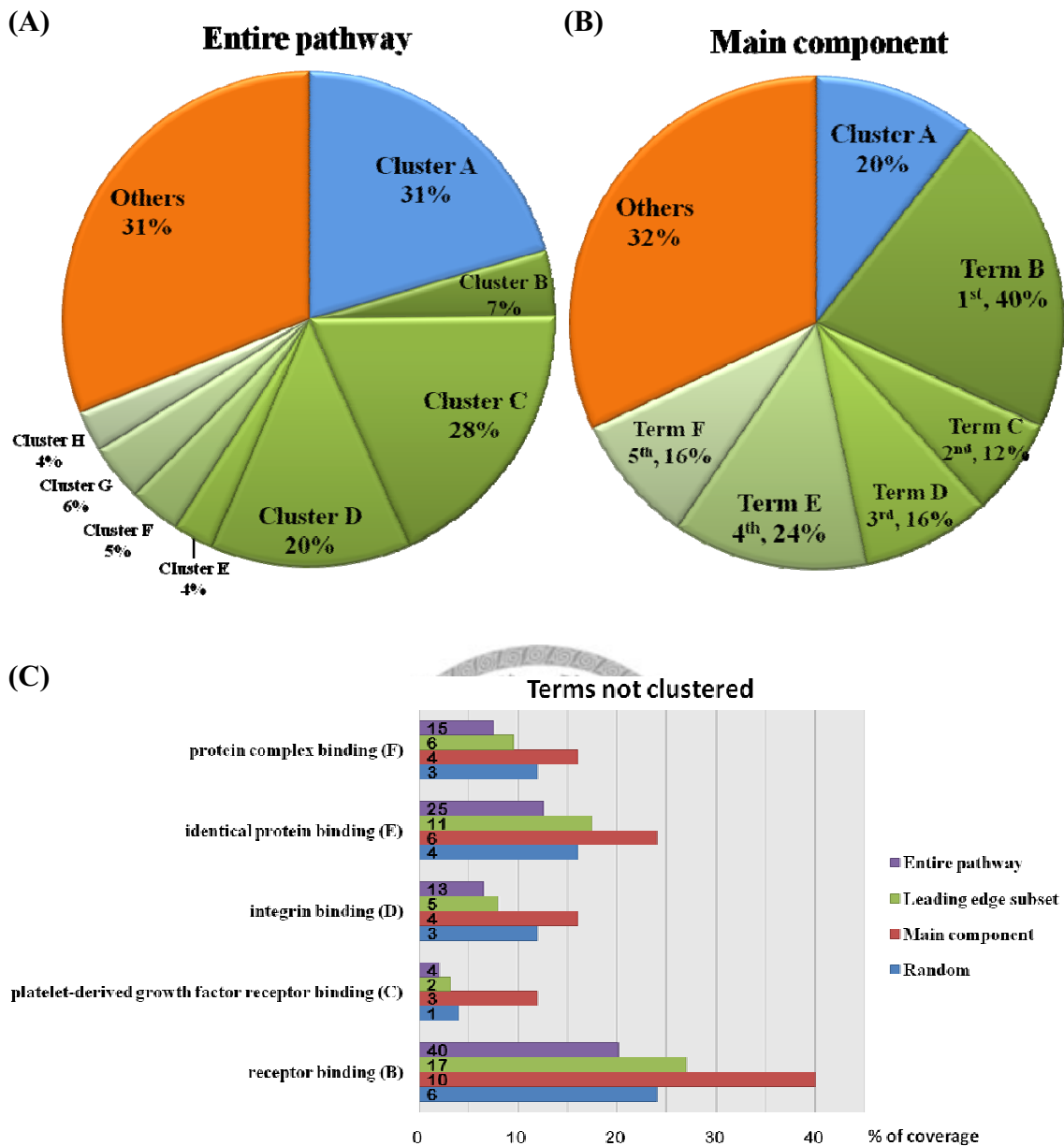


Figure 4-8. GO terms (molecular function 1-5) enriched in focal adhesion. *Pie charts were used to illustrate how GO terms/clusters share the members of (A) the original pathway and (B) the main component. In addition, significant GO terms occupying an amplified proportion in main component were listed in (C). The relationships between these terms are illustrated in Figure A-5. In (C), genes involved in each term accounted for different percentages in the four sets and the percentages were illustrated as a bar chart with the digit on the bar specifying the actual number of genes involved in this term. The random set was randomly selected from the original pathway and with the same size as the main component.*

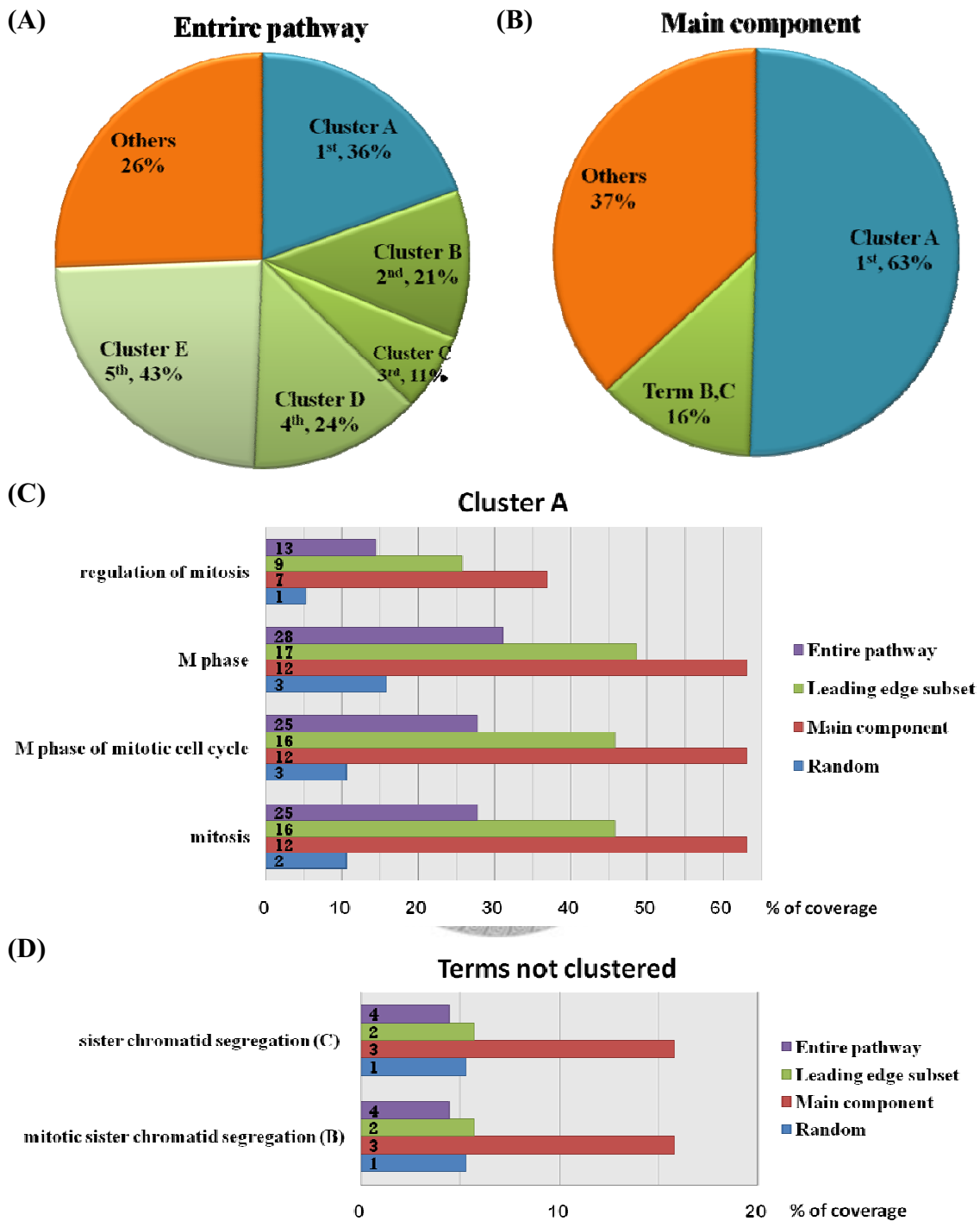


Figure 4-9. GO terms (biological process 5) enriched in cell cycle. *It illustrates how GO terms/clusters share the members of (A) the original pathway and (B) the main component, where the rankings represent relative significances of clusters. Terms in cluster A were listed in (C) and their hierarchical relationships are illustrated in Figure A-6. In (C) and (D), genes involved in each term accounted for different percentages in the four sets and the percentages were illustrated as a bar chart with the digit on the bar specifying the actual number of genes involved in this term.*

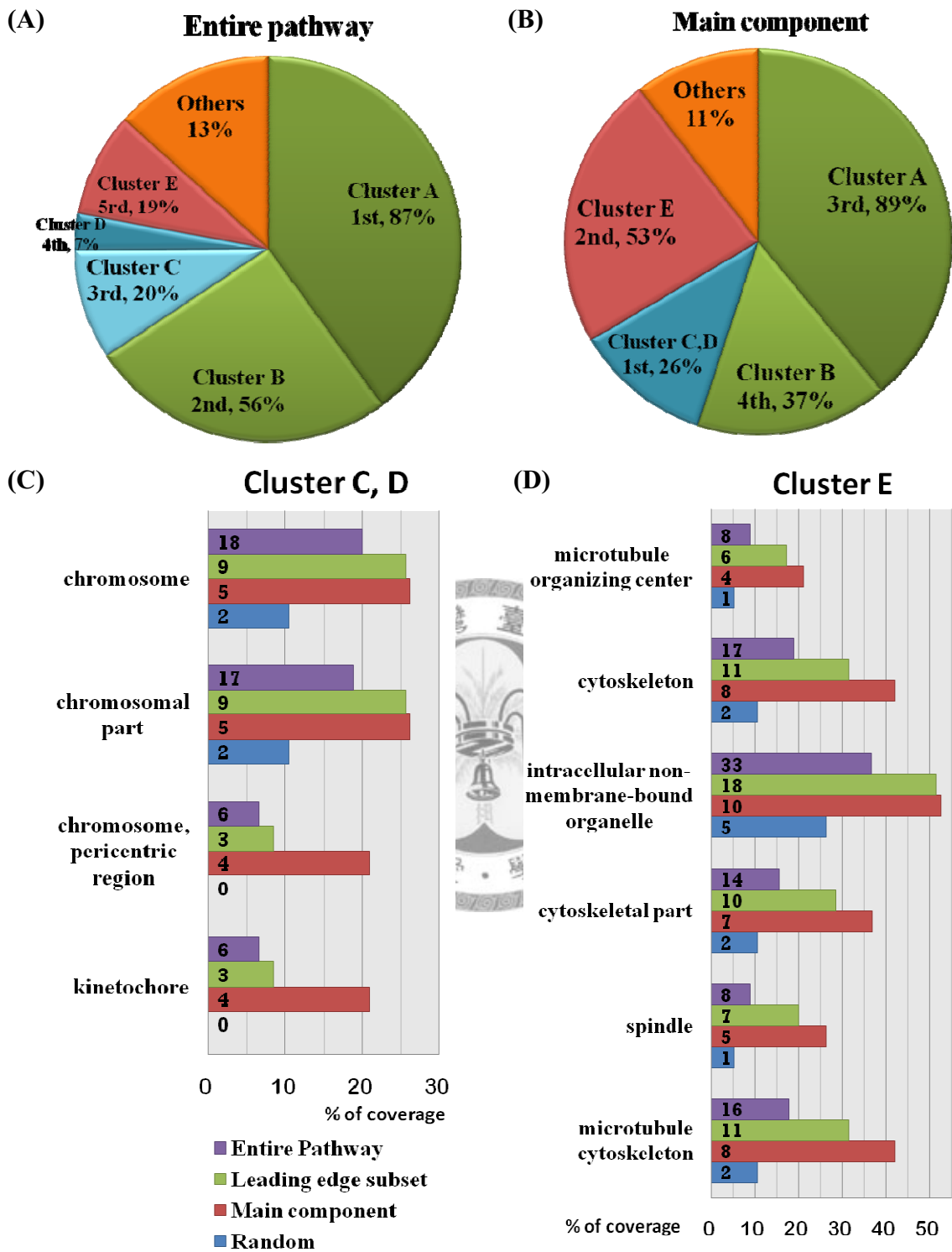


Figure 4-10. GO terms (cellular component 5) enriched in cell cycle. *It illustrate how GO clusters share the members of (A) the original pathway and (B) the main component, where the ranking represents relative significance of clusters. In (C) and (D), member terms in cluster C,D,E were listed and their relationships are illustrated in Figure A-7. Genes involved in each term accounted for different percentages in the four sets and the percentages were illustrated as a bar chart with the digit on the bar specifying the actual number of genes involved in this term.*

By showing that the main component focuses on specific functionality than the original pathway does, the analyses in this section manifest the biological meanings of these subnetworks. More specifically, the subnetwork suggests that a group of interconnected genes probably performing a specific function that is most dysregulated in the original pathway.

4.6 Network analysis – protruding pathways

This section shows a preliminary attempt to extend searches to outside the pathway. During network analysis, root nodes remained to be each member of the pathway in order to focus on the dysregulated pathways. Almost all procedures were unchanged, the only setting different from that in section 4.4 is that the search space changed to be the global interaction network. It means that we aim to find subnetworks containing genes potentially interacting with these predefined pathway members. By doing so, it allowed an exploratory analysis relating the pathway through giving researchers hints to important potential interactions that are related to known pathways but are not defined inside them. The results were displayed in Figure 4-11 and annotations of these genes were listed in Table A-8. The subnetworks in Figure 4-11 actually contain genes not defined in the pathways. These genes are represented as circles and those genes being the member of original pathways are displayed by rectangles.

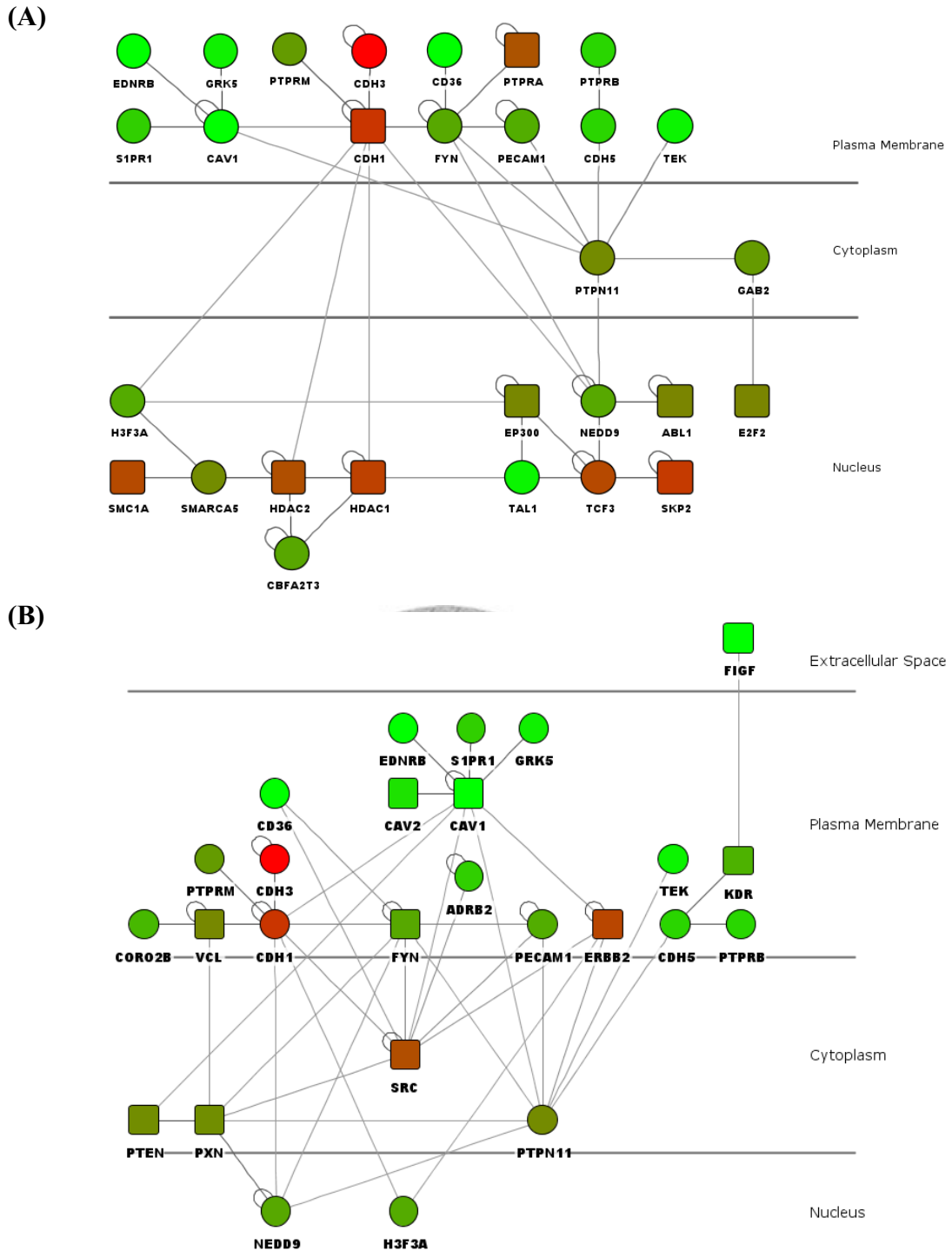
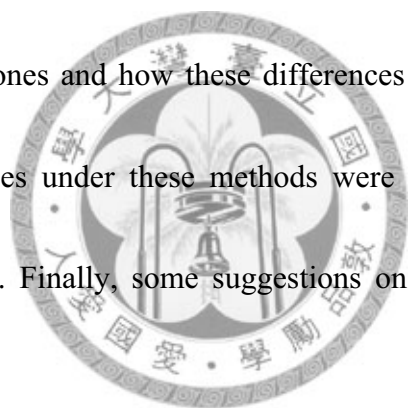


Figure 4-11. Extend subnetwork search to the global interaction network. Based on (A) cell cycle pathway and (B) focal adhesion pathway, subnetworks containing potential interacting gene neighbors of the pathways are obtained by extending network analyses to the global interaction network. In these subnetwork, genes were arranged by their subcellular locations and colored by expression fold changes (red for up-regulation and green for down-regulation in cancer phenotype). Rectangles represent genes' being a member of the interested pathway while circles stand for the neighboring genes interacting with the pathway.

Chapter 5 Discussion

It is worthwhile to emphasize that the value of this methodology lies in the motivation to integrate both analyses, since it may complement each other and make use of both advantages. Although our methodology followed Tian's pathway analysis method and GXNA, they can be substituted with comfort by any other pathway/network analysis method possessing the same capability.

In this chapter, it is explained in separate sections why the methods adopted here differed from the original ones and how these differences might influence the results. Furthermore, the weaknesses under these methods were revealed and corresponding suggestions were proposed. Finally, some suggestions on future perspectives will be mentioned



6.1 Input matrix creation

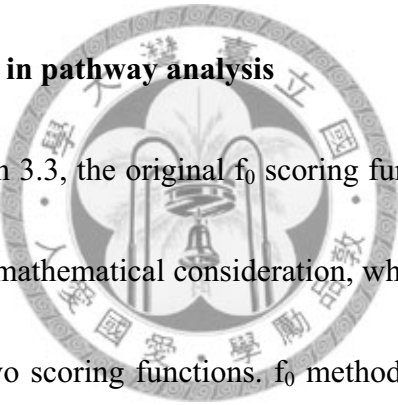
To obtain an input matrix prerequisite for pathway/network analysis, redundant probe sets were removed by eliminating non-representative probe sets for each gene. Three general approaches were usually adopted: to represent each gene by its maximal or median probe set in terms of differential expression or by the average of all its probe sets.

To prevent the gene's significance level from being affected by potentially

ineffective probe designs, maximal probe set was selected in section 4.1. However, when making such a decision, the trade-off is to have possibly amplified the noise which may be produced by a high-scoring probe set targeting at several different genes. The peak illustrated in Figure 4-3 is an example for this situation. Certainly, it could be avoided by simply truncating those non-specific probe sets; however, they were preserved here in order to reserve as many information as possible.

6.2 Pathway analysis

✧ Two scoring functions in pathway analysis



As mentioned in section 3.3, the original f_0 scoring function is slightly altered into f_1 . Doing this is not for any mathematical consideration, whereas different types of gene sets were targeted by the two scoring functions. f_0 method aims to find a set of genes with concordant changes while they might not show individually significant differential expression; f_1 targets to select sets that contain a proportion of significant genes higher than that outside the sets regardless of their concordance in terms of direction of changes. This difference was revealed in Figure 4-5, where a consistent up-regulation in cancer phenotype was observed in Figure 4-5A but not in Figure 4-5B. The decision of scoring scheme to use during analysis is indeed dependent on one's purpose. For example, if one focuses on downstream targets of a transcription factor, f_0 would just fit;

in contrast, if one is searching for chains of signaling transductions or regulatory circuits that involve various activation/inhibition relationships and lead to an indefinite overall direction of change, f_1 might be more close to the need.

✧ **Permutation method and pathway score normalization**

Significance level of a pathway score was derived by its null distribution and served as the major index to assess importance of a pathway. Tian *et al.* [27] suggested two types of permutation methods that correspond to different biological questions: one is to permute gene order and the other is to permute phenotypes.

In the case of phenotype permutation, it is inadequate to directly shuffle all class labels as it usually did because paired normal-tumor arrays were utilized here. It is because that doing so, one is further assuming the invariance of expression profiles among patients, which is obviously not the truth. Alternatively, phenotype permutation is achieved by randomly deciding whether to exchange each pair of tumor and normal class labels.

The effect of this modification did not show apparently because both ways of shuffling yield mostly significant results. In fact, cancer tissues usually exhibit great differences from normal ones and thus, it was not surprising to identify so many pathways passing the significant criteria in Figure 4-4 when comparing their scores with

null distributions assuming no differences exist between phenotypes. Therefore, phenotype permutation is accordingly much less discriminative than gene order permutation.

Furthermore, a limited resolution problem evolves due to the incapability of a null distribution to cover a broad-enough range of pathway scores. Inevitably, the weakness is derived from the essence of resampling procedures. It occurs in the situation where insufficient permutations are performed and becomes especially evident when using a dataset whose genes showed dramatically altered expression, and this is exactly the case here and leads to a lot of pathways with same extremely small significance level.

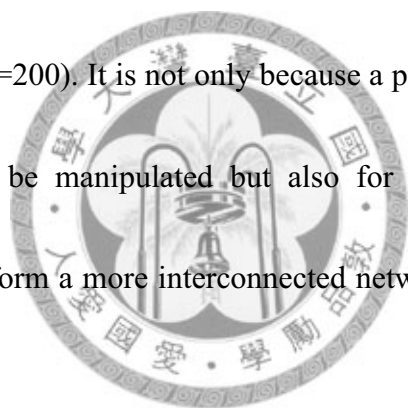
In such situations where significance level is unable to discern pathways, normalized pathway score serves as a further index to compare their importance. Tian *et al.* [27] normalize pathway scores by using the following principle: if the score falls within its null distribution it is replaced with its quantile, and those falling far from the null distribution are converted into corresponding z-scores.

However, z-scores might not be directly comparable to each other since null distributions differ from one another in different datasets. Thus, normalized scores obtained by this method should always be used with notice, especially when many of them are derived from z-score transformation. This is because it might fail to be reliable

when z-scores depend largely on the features of their null distributions. Therefore, when the pathway analysis indicates it is a significant pathway, it really is, while if it suggests that one pathway is the most dysregulated among these significant results, users should always be more careful.

✧ **The roles of this methodology in relatively large and small pathways**

The ability of this methodology to extract modules within pathways is both applicable and profitable, especially in large pathways such as the focal adhesion pathway selected here (size=200). It is not only because a pathway with handful amount of members are easier to be manipulated but also for a larger group of pathway members would generally form a more interconnected network by using information in interaction databases.



At present, manually curated pathways remain the most reliable source for pathway analysis, yet many of the public databases, such as BioCarta [7], tend to be more conservative, since relatively small number were recorded when they were compared with the actual size, which was believed to contain a few hundred or even thousands of molecules. An overall concept of pathway size distributions in the database were illustrated in Figure A-8. It shows that most pathways are with a size smaller than 50.

However, one might question what this methodology can actually provide in terms of small pathways? In fact, the contents of individual pathway are expected to be improved if more biological data are gathered. With the aid of a computational tool that enables the extension of modules to genes, which locate outside a predefined pathway, it has great potential to point researchers to those interacting neighbors which are suspicious to be the missing components in the existing pathways with relatively small size. This applicability is yet to be widely realized by other computational analysis tools.

6.3 Network analysis

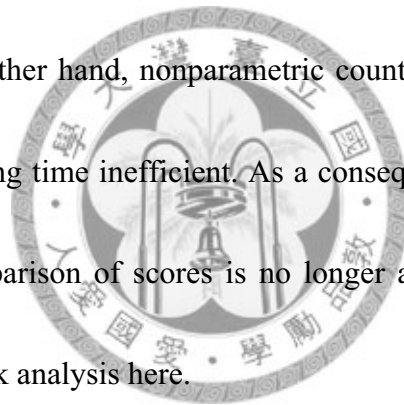
✧ Two scoring functions in network analysis

Specifically, to assess whether a group of genes (either a pathway or a subnetwork) is related to a study, two indices are the major concern: a score independent of group size and a significance level of the score.

When scores do not depend on size it means that they are directly comparable to each other. In terms of network analysis it implies the ability to conduct flexible-size subnetwork search which may identify modules with indefinite size. It can be achieved by directly implementing $T\Sigma$ scoring scheme (equation 8 in Nacu *et al.* [33], Ideker *et al.* [36]) or parametric ΣT scoring scheme (equation 6 in Nacu *et al.* [33], default

scoring function in GXNA). Otherwise, it is required to eliminate dependency on group size before comparison. In the example of nonparametric ΣT scoring scheme (f_0 and f_1 scoring functions), scores are normalized by using the reference null distribution.

However, as illustrated in Figure 4-3, the filtered t-scores do not follow a normal distribution as expected. This situation was not improved when median probe set was used to represent a gene. The parametric assumption was thus failed to be established and this is the major reason why we did not to apply GXNA's scoring function in our network analysis. On the other hand, nonparametric counterpart requires large amount of resampling and thus being time inefficient. As a consequence, a fixed-size approach is adopted where the comparison of scores is no longer an issue, and thus f_1 scoring method was used in network analysis here.

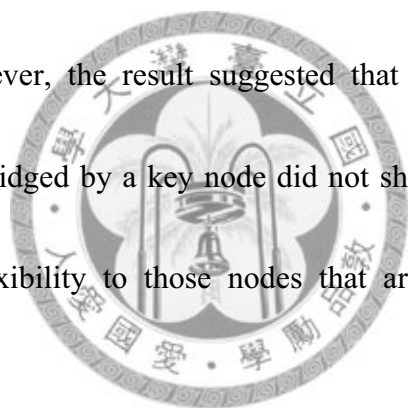


In addition, it is known that regulations mechanisms spread from DNA/mRNA level to protein level, which implies the probability of certain proteins being key players to the connection of significant components, but they may show no differentially expression at mRNA level, as mentioned in chapter 1.

In GXNA it filtered out probe sets with small variances, and doing this might lose tract of these key nodes. However, in the work of Ideker *et al.* [32] such problem did not exist because it utilized simulated annealing. In fact, their main objectives are

different. GXNA aimed to identify subnetworks where all members show certain degree of differential expression and Ideker *et al.* [32] tried to adapt the algorithm to the events actually happening in biological systems. Unfortunately, simulated annealing costs too much time, so we followed GXNA's approach. In order to compensate it, devised f_2 scoring function was utilized. The new scoring function is able to tolerate key nodes as shown as in Figure A-3.

As in Table A-6, the key nodes found in the two pathways did not pass the significance criteria. However, the result suggested that the two groups of densely connected genes may be bridged by a key node did not show apparently. Nonetheless, this idea still reserves flexibility to those nodes that are not identified by mRNA microarrays studies.



◇ **Starting condition : root nodes and search space**

In terms of searching algorithm, we basically follow the greedy approach in GXNA, while some modifications were made in the determination of root nodes and search space.

Different from GXNA, which always chooses random root nodes and searches under global interaction network, the starting condition in this methodology is relatively much more flexible. The search space and root node determination depends on the

purposes. The root nodes can be members of specific and interesting pathways, and the search space can be the global interaction network or its subsets by functional or positional groupings, and this provides full flexibility to meet biologists' interests.

In section 4.4, we aimed to obtain the most important module in a pathway, so the root nodes were pathway members and the search space was defined within the pathway.

In section 4.6, since the purpose was to explore genes interacting with known pathway members, the root nodes were pathway members and the search space was the global interaction network.

Within the most significant subnetwork obtained by GXNA (visualized in Figure A-2), few members hint to a common pathway. In contrast, the results in Figure 4-5, Figure A-3 and Figure 4-10 obtained by our methodology were much more focused on specific pathways. This advantage to conduct focus-oriented analyses evidenced that these modifications make our approach much more useful.

✧ **Merging process**

GXNA allows both fixed-size and flexible-size subnetwork search. The reason why the flexible-size approach was not applied was discussed in the previous section. In this methodology we only allows for fixed-size search; however, to compensate this disadvantage, a merging process was developed in this methodology.

It is hypothesized that once an information flow is triggered, the signal propagates sophisticatedly along its pre-designed paths including various interactions between molecules. Once a region existed strong evidence of such information flow, which amounted to the existence of a group of connected genes showing significant differential expression, its neighboring genes would follow the gradient of evidence strength and finally reach the region during the greedy extension algorithm. Thus in the methodology here such a region would be implied within several candidate subnetworks. Once the region with the strongest evidence is identified, the merging process is used to reshape the region. However, there might be more than one such informative region and this is the reason why we accept to specify multiple main components.

A potential alternative solution is to apply the clustering method in DAVID. It proposes to cluster the candidate subnetworks to identify overlapped regions, and the clusters are ordered in a fashion that each of them can be viewed as a main component.

✧ **Incompleteness of biomolecular interaction information**

Although a pathway is an integration of interacting genes that shall be also seen in biomolecular networks, it is observed that, small pathways are prone not to be connected into an integral component because of the incompleteness of interaction information. In commercial databases the knowledge base are constructed by hiring an

army of experts to curate information from public databases or scientific literatures. On the other hand, teams maintaining public databases might also have transformed their pathway information into corresponding formatted data such as the KEGG Markup Language (KGML). These data enable automatic pathways drawing and provide facilities for computational analysis. The incompleteness of interaction data can be improved by incorporating such formatted data from these pathway databases.

6.4 Future perspectives

✧ Future perspective on pathway analysis methods

Among various methods developed for pathway analysis, most of them do not take the topology of pathways into consideration. While in contrast, it is of tremendous need to take into account the pipelines (physical structure) that enables the information flow during pathway activation. It is especially important in a cancer-related study because pathways are with increased probability to contain more significant genes than random selection.

Draghici *et al.* [53] detected pathway dysregulation by an impact analysis that take into account some crucial factors of genes such as differential expression, interactions and positions in the pathway, etc. It was done by incorporating all upstream information as well as measures of expression change such as fold-change into the scoring function,

which would reduce to traditional statistics when the additional information are forcefully ignored. Such a scoring function is actually designed for weighting the chained activations in gene signaling network, while it lost attention to the existence of genes acting in coordinated fashion. Moreover, a prior filtering for significant genes is also a prerequisite for this analysis. Although the idea in their work seems intuitive, they had indeed made a step ahead to practically realize the incorporation of upstream information. In the future, solutions shall gradually emerge and we expect a biologically reasonable method for pathway analysis including topology evaluation and without prerequisite filtering.

✧ **Future perspectives on this methodology**

Two strategies in terms of searching pathway cross-talks are suggested here.

1. Focusing on cross-pathway inhibition.

Pathways suspicious to interact with each other are suggested to be identified previous to applying network analysis. A possible approach begins with identifying main components in separate dysregulated pathways. By calculating canonical correlations between these modules, those pathways with potential correlated gene expression would be then identified. Based on these modules rather than the entire pathways the cross-talks are to be found between them using network analysis. While in

terms of small pathways, the component extraction step could be ignored.

2. Focusing on upstream/downstream targets of pathways

Another task-oriented approach is to find the upstream or downstream targets of pathways. Other than cross-pathway inhibition that contains many interactions in the transverse direction in terms of cell structure, pathway cross-talks in the vertical direction is also interesting to many biologists. It might be, for instance, the binding of ligands and receptors or the regulation of TFs on target genes. Such events could be highlighted by extending modules to ligands/TFs outside the pathway.



Chapter 6 Conclusions

By integrating advantages in both pathway analysis and network analysis, we developed a methodology that is able to perform deep investigations in dysregulated pathways and to perform exploratory analyses based on these pathways. This methodology was applied to our own dataset of lung cancer microarrays and the results were consistent with that of a public lung cancer dataset (GSE7670). A knowledge database was constructed in the very beginning, and all needed information during analysis is available within this database.

In section 4.3, both Tian method and modified Tian method were applied on the datasets to identify dysregulated pathways. Table A-3 showed the better statistical power of these two methods than that of another pathway analysis method GSEA. In network analysis, one dysregulated pathway in common to both dataset was selected by each method, respectively: cell cycle pathway was selected by Tian method and focal adhesion pathway was selected by modified Tian method.

In section 4.4, we attempted to find the most differential component inside dysregulated pathways from the viewpoint of biomolecular interaction network, and this component was then referred to as a module or a main component. The main component in cell cycle and focal adhesion pathway, which were presented in section 4.4, found in

our dataset were consistent with that in the GSE7670 dataset.

In section 4.5.1, members of main components and leading edge subsets obtained by GSEA were simultaneously overlaid on the conceptualized pathway map. In addition, potential interactions absent in predefined pathways were complemented by information in interaction database. Figure 4-6 revealed the advantage of incorporating biomolecular interaction network during analysis: it showed that despite the members of these two sets overlapped to some degree, the main component was topologically more connected than the leading edge subset did.

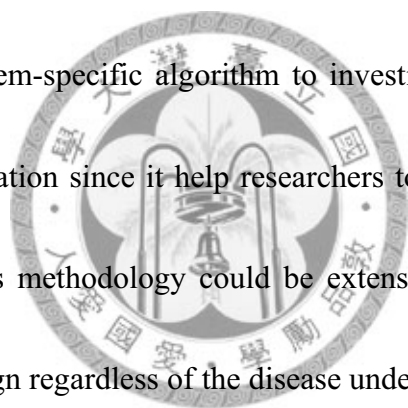
Furthermore, these modules were analyzed by DAVID to elucidate the underlying biological meaning in terms of different gene ontology categories. It was shown in section 4.5.2 that compared to the original pathways, the main components indeed show a specialized functionality and such trend of specialization appeared consistently in both leading edge subset and main component of these two pathways. However, the main component is more advantageous than the leading edge subset in two aspects: the size of leading edge subset is much larger than that of main component and makes the main component seemed much easier to be further investigated; the leading edge subset does not take interaction between genes into consideration as is done in this methodology.

Therefore, the main component would be more biologically meaningful in terms of

analysis procedures.

With the confidence to extract biologically meaningful modules by this methodology, further focus-oriented investigations would be easier to be conducted. For example, a preliminary attempt was made in section 4.6 to search for possible missing components in pathways or cross-talks between pathways by extending the search space to outside the dysregulated pathways.

Although it is in spirit an ad-hoc procedure, this methodology provides an adequate tool that implements problem-specific algorithm to investigate topics of interest. It is valuable in terms of application since it help researchers to highlight on their research interests. Undoubtedly, this methodology could be extensively applied to other array experiments of similar design regardless of the disease under study.



REFERENCES

1. 行政院衛生署-統計資訊網. 2009;
Available from: http://www.doh.gov.tw/CHT2006/index_populace.aspx.
2. Subramanian, J. and R. Govindan, *Molecular genetics of lung cancer in people who have never smoked*. *Lancet Oncol*, 2008. **9**(7): p. 676-82.
3. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray*. *Science*, 1995. **270**(5235): p. 467-70.
4. Yongchao Ge, S.D., and Terence P. Speed, *Resampling-based multiple testing for microarray data analysis*. *TEST*, 2003. **12**(1): p. 1-77.
5. Sandrine Dudoit, J.P.S.a.J.C.B., *Multiple Hypothesis Testing in Microarray Experiments*. *Statistical Science*, 2003. **18**(1): p. 71-103.
6. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes*. *Nucleic Acids Res*, 2000. **28**(1): p. 27-30.
7. *BioCarta pathways database*.
Available from: <http://www.biocarta.com/genes/index.asp>.
8. Dahlquist, K.D., et al., *GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways*. *Nat Genet*, 2002. **31**(1): p. 19-20.
9. *Category C2, Curated canonical pathways*. April, 2008;
Available from: <http://www.broad.mit.edu/gsea/msigdb/collections.jsp#C2>.
10. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource*. *Nucleic Acids Res*, 2004. **32**(Database issue): p. D258-61.
11. Mani, R., et al., *Defining genetic interaction*. *Proc Natl Acad Sci U S A*, 2008. **105**(9): p. 3461-6.
12. Daraselia, N., et al., *Extracting human protein interactions from MEDLINE using a full-sentence parser*. *Bioinformatics*, 2004. **20**(5): p. 604-11.
13. von Mering, C., et al., *Comparative assessment of large-scale data sets of protein-protein interactions*. *Nature*, 2002. **417**(6887): p. 399-403.
14. Hart, G.T., A.K. Ramani, and E.M. Marcotte, *How complete are current yeast and human protein-interaction networks?* *Genome Biol*, 2006. **7**(11): p. 120.
15. Yuryev, A., *Introduction to pathway analysis*, in *Pathway analysis for drug discovery - computational infrastructure and applications*, A. yuryev, Editor. 2008, Wiley. p. 1-20.
16. Bader, G.D., D. Betel, and C.W. Hogue, *BIND: the Biomolecular Interaction Network Database*. *Nucleic Acids Res*, 2003. **31**(1): p. 248-50.
17. Peri, S., et al., *Human protein reference database as a discovery resource for*

- proteomics*. Nucleic Acids Res, 2004. **32**(Database issue): p. D497-501.
18. Zanzoni, A., et al., *MINT: a Molecular INTERaction database*. FEBS Lett, 2002. **513**(1): p. 135-40.
 19. Curtis, R.K., M. Oresic, and A. Vidal-Puig, *Pathways to the analysis of microarray data*. Trends Biotechnol, 2005. **23**(8): p. 429-35.
 20. Sivachenko, A.Y., *Pathway analysis of high-throughput experimental data*, in *Pathway analysis for drug discovery - computational infrastructure and applications*, A. yuryev, Editor. 2008, Wiley. p. 103-117.
 21. *Calculating and Interpreting the P-values for Functions, Pathways, and Lists in Ingenuity Pathways Analysis*. 2008.
 22. Ekins, S., et al., *Pathway mapping tools for analysis of high content data*. Methods Mol Biol, 2007. **356**: p. 319-50.
 23. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*. Nat Protoc, 2009. **4**(1): p. 44-57.
 24. Richard Simon, A.P.L., *BRB-ArrayTools Version 3.7 - User's Manual*. 2007. p. 69-73.
 25. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
 26. Mootha, V.K., et al., *PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes*. Nat Genet, 2003. **34**(3): p. 267-73.
 27. Tian, L., et al., *Discovering statistically significant pathways in expression profiling studies*. Proc Natl Acad Sci U S A, 2005. **102**(38): p. 13544-9.
 28. Pavlidis, P., D.P. Lewis, and W.S. Noble, *Exploring gene expression data with class scores*. Pac Symp Biocomput, 2002: p. 474-85.
 29. Watts, D.J. and S.H. Strogatz, *Collective dynamics of 'small-world' networks*. Nature, 1998. **393**(6684): p. 440-2.
 30. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization*. Nat Rev Genet, 2004. **5**(2): p. 101-13.
 31. Wu, Z., X. Zhao, and L. Chen, *Identifying responsive functional modules from protein-protein interaction network*. Mol Cells, 2009. **27**(3): p. 271-7.
 32. Ideker, T., et al., *Discovering regulatory and signalling circuits in molecular interaction networks*. Bioinformatics, 2002. **18 Suppl 1**: p. S233-40.
 33. Nacu, S., et al., *Gene expression network analysis and applications to*

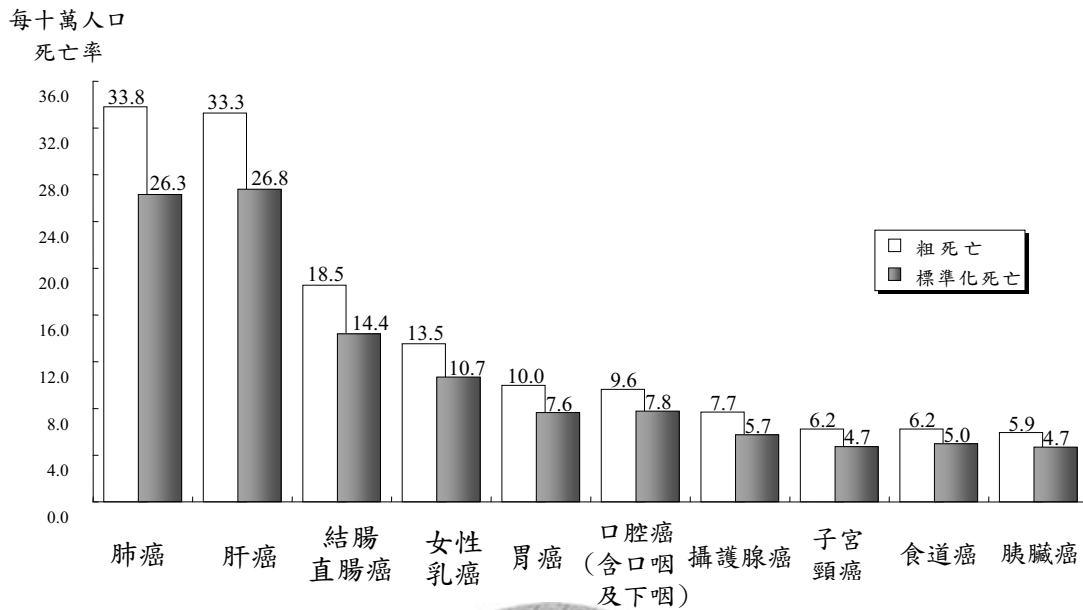
- immunology*. Bioinformatics, 2007. **23**(7): p. 850-8.
34. Kelley, R. and T. Ideker, *Systematic interpretation of genetic interactions using protein networks*. Nat Biotechnol, 2005. **23**(5): p. 561-6.
 35. Li, Y., P. Agarwal, and D. Rajagopalan, *A global pathway crosstalk network*. Bioinformatics, 2008. **24**(12): p. 1442-7.
 36. Lee, E., et al., *Inferring pathway activity toward precise disease classification*. PLoS Comput Biol, 2008. **4**(11): p. e1000217.
 37. Chuang, H.Y., et al., *Network-based classification of breast cancer metastasis*. Mol Syst Biol, 2007. **3**: p. 140.
 38. Su, L.J., et al., *Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme*. BMC Genomics, 2007. **8**: p. 140.
 39. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository*. Nucleic Acids Res, 2002. **30**(1): p. 207-10.
 40. Stark, C., et al., *BioGRID: a general repository for interaction datasets*. Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.
 41. Karp, P.D., et al., *The EcoCyc Database*. Nucleic Acids Res, 2002. **30**(1): p. 56-8.
 42. Inc., P., *Partek Genomics Suite*. 2009, Partek Inc.: St. Louis.
 43. Cline, M.S., et al., *Integration of biological networks and gene expression data using Cytoscape*. Nat Protoc, 2007. **2**(10): p. 2366-82.
 44. Barsky, A., et al., *Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation*. Bioinformatics, 2007. **23**(8): p. 1040-2.
 45. *Color Objects in KEGG Pathways*.
Available from: http://www.genome.jp/kegg/tool/color_pathway.html.
 46. *QuickGO : a web-based browser for Gene Ontology terms and annotations*.
Available from: <http://www.ebi.ac.uk/QuickGO/>.
 47. *GeneChip[®] Human Genome Arrays*.
 48. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data*. Nucleic Acids Res, 2003. **31**(4): p. e15.
 49. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*. Bioinformatics, 2003. **19**(2): p. 185-93.
 50. Daniel Holder, R.F.R., V. Bill Pikounis, Vladimir Svetnik, Keith Soper.

Statistical analysis of high density oligonucleotide arrays: a SAFER approach.
in *Proceedings of the ASA Annual Meeting 2001*. 2001.

51. Bland, J.M. and D.G. Altman, *Multiple significance tests: the Bonferroni method*. *BMJ*, 1995. **310**(6973): p. 170.
52. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies*. *Proc Natl Acad Sci U S A*, 2003. **100**(16): p. 9440-5.
53. Draghici, S., et al., *A systems biology approach for pathway level analysis*. *Genome Res*, 2007. **17**(10): p. 1537-45.
54. 圖 20. 97 年主要癌症死亡率, 97 年死因統計結果分析.doc, Editor, Department of Health, Executive Yuan, R.O.C. (Taiwan).



APPENDIX



附註：標準化死亡率係以 W.H.O.2000 年世界標準人口數為基準。

Figure A-1. Top cancer killers in Taiwan in 2008. See [54] for the source of this figure.

Table A-1. Statistics of t-scores in two datasets.

(A)

All probe sets	Overall			Bonferroni adjusted p-value < 0.05		
	t score <=0	t score >=0	in total	t score <=0	t score >=0	in total
NTUH	9125	13099	22215	1653	1621	3274
GSE7670	8708	13507	22215	804	313	1117

(B)

Representative probe sets	Overall			Bonferroni adjusted p-value < 0.05		
	t score <=0	t score >=0	in total	t score <=0	t score >=0	in total
NTUH	5492	8307	13799	1345	1489	2834
GSE7670	5374	8425	13799	735	343	1078

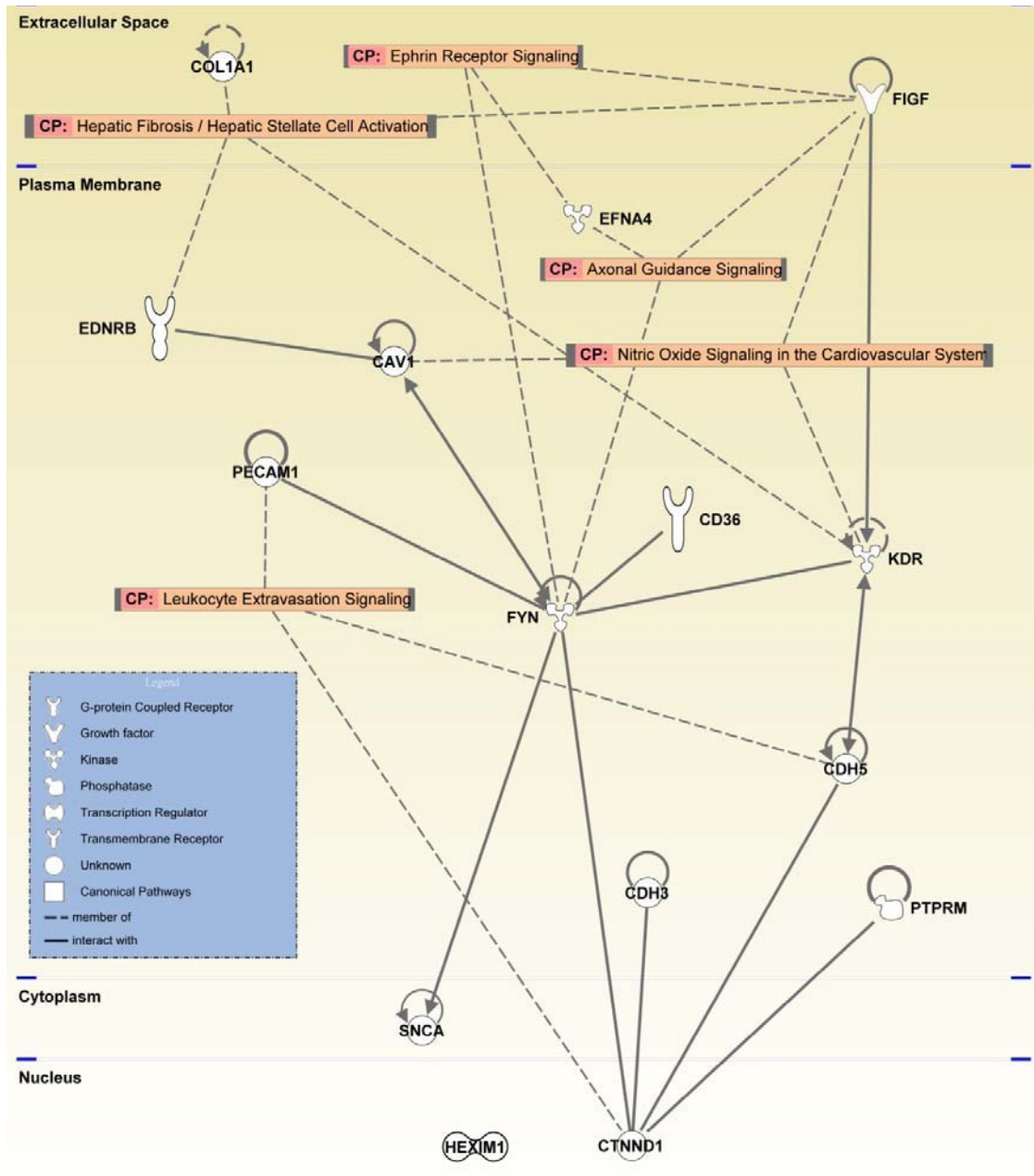


Figure A-2. The most significant subnetwork identified by GXNA. The result obtained by GXNA program was visualized using Pathway Designer in IPA .

Table A-2. Parameters set during pathway analysis.

GSEA	NTUH_real_t (f_0)	NTUH_abs_t (f_1)
set_min	10	10
num	100	100
nperm	10000	10000
plot_top_x	200	200
set_max	500	500
chip	HG_U133A.chip	HG_U133A.chip
gmx	c2.cp.v2.5.symbols.gmt	c2.cp.v2.5.symbols.gmt
mode	Max_probe	Max_probe
sort	real	abs
median	FALSE	FALSE
norm	meandiv	meandiv
rnd_type	no_balance	no_balance
permute	phenotype	phenotype
metric	tTest	tTest
rnd_seed	timestamp	timestamp
collapse	TRUE	TRUE
make_sets	TRUE	TRUE
scoring_scheme	weighted	weighted

Table A-3. Significant pathways identified by different methods.

Tian	NTUH			GSE7670			NTUH \cap GSE7670
Permutation type	PG	PC	PG \cap PC	PG	PC	PG \cap PC	
f_0 Bonferroni	49	340	49	16	138	16	12
f_0 q-value	247	521	243	132	439	132	too much
f_1 Bonferroni	2	558	2	1	548	1	too less
f_1 q-value	14	---	---	23	---	---	6

※ PG stands for permute gene order; PC stands for permute class label; --- means it is not available.

GSEA		NTUH				TVGH			
		Enriched sets	FDR < 25 %	nominal p < 1 %	nominal p < 5 %	Enriched sets	FDR < 25 %	nominal p < 1 %	nominal p < 5 %
f_0	normal	317	0	13	41	331	6	13	70
	cancer	180	0	2	21	166	0	0	13
f_1	normal	7	2	1	2	5	0	0	0
	cancer	490	0	19	61	492	0	8	61

※ GSEA suggests FDR < 25 % as the criteria for pathway dysregulation.

Table A-4. Lists of significant pathways identified by f_0 and f_1 scoring function.

Pathways with q-value < 0.05 under f_0 scoring scheme		
Pathway Name	Size	Average t-score
GLUTAMATE_METABOLISM	23	3.80
PYRIMIDINE_METABOLISM	59	3.32
HSA00240_PYRIMIDINE_METABOLISM	74	3.28
CELL_CYCLE_KEGG	85	3.09
integrin signaling	175	-1.03
HSA04010_MAPK_SIGNALING_PATHWAY	240	-1.13
HSA04810_REGULATION_OF_ACTIN_CYTOSKELETON	190	-1.18
SMOOTH_MUSCLE_CONTRACTION	138	-1.58
HSA04650_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	124	-2.05
G_PROTEIN_SIGNALING	91	-2.20
PPARAPATHWAY	54	-2.64
PROSTAGLANDIN_SYNTHESIS_REGULATION	29	-4.85

Pathways with Bonferroni adjusted p-value < 0.05 under f_1 scoring scheme		
Pathway Name	Size	Average t-score
CARDIACEGFPATHWAY	18	5.949233
PROSTAGLANDIN_SYNTHESIS_REGULATION	29	5.3233
HSA04512_ECM_RECEPTOR_INTERACTION	82	4.744201
BREAST_CANCER_ESTROGEN_SIGNALING	101	4.694804
HSA04520_ADHERENS_JUNCTION	72	4.631791
HSA04510_FOCAL_ADHESION	190	4.516165

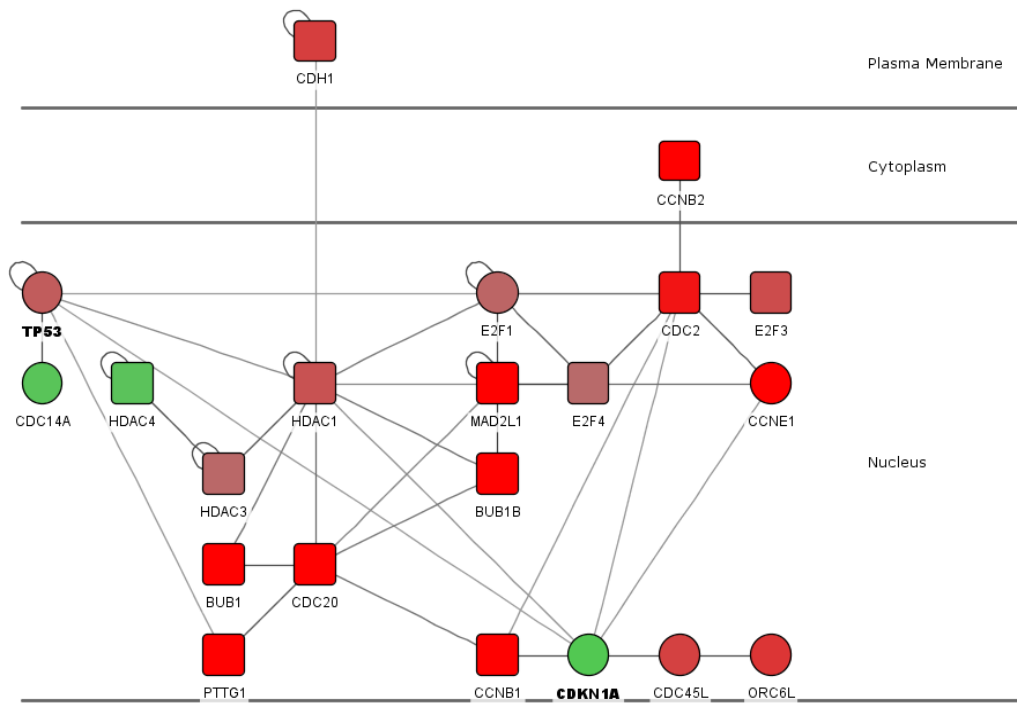
Table A-5. Annotations of genes present in Figure 4-5.

Genes in the main component of cell cycle pathway. (Fig. 4-4A)				
Gene Symbol	Fold Change	<i>p</i>-value	Normal Expression	Tumor Expression
CDH1	2.4	1.26E-13	1994.5±381.8	4752.1±1690
CCNB2	5.6	5.75E-11	69.8±23.3	388.4±401.5
BUB1	4.6	8.02E-09	31.6±9.9	144.9±144.1
BUB1B	6.0	1.56E-10	57.3±20.8	344.6±354.3
CDC2	3.5	1.33E-08	155.5±45.5	540.5±521.6
BUB3	1.4	4.94E-08	693.2±98.2	957.9±211.7
CCNA2	3.9	1.69E-07	33.1±7.8	129.6±148.5
CCNB1	6.3	1.32E-08	59.4±20.7	371.6±482.8
CDC20	7.1	5.01E-10	58.9±15.4	415.8±460.9
MAD2L1	5.1	4.78E-08	56.3±23.3	284.9±315.1
PCNA	2.0	5.86E-07	890.8±196.1	1787.4±1005.2
ESPL1	1.8	5.29E-07	67.9±16.3	119.7±65.6
CDC25A	2.2	7.12E-07	51.3±12.3	115.2±92.8
E2F3	2.0	1.28E-07	270±87.5	543.4±295.3
E2F4	1.2	4.20E-07	606.6±140.3	739.5±213.7
HDAC1	1.9	1.20E-07	1181±222.6	2198.6±1099.5
HDAC3	1.3	7.30E-06	467.5±45	604.7±156.6
HDAC4	-1.6	1.05E-07	260.4±75.8	161.8±80.4
PTTG1	4.8	1.27E-09	476.2±166.2	2262.6±2763.6

Genes in the main component of focal adhesion pathway. (Fig. 4-4B)

Gene Symbol	Fold Change	p-value	Normal Expression	Tumor Expression
SPP1	29.7	1.02E-15	310.8±476.6	9227.8±5885.9
PDGFA	-1.2	5.43E-04	96.8±28	80±24.5
PDGFB	-1.8	2.25E-09	264.9±80.4	147.2±59.4
PDGFC	2.1	4.57E-07	1006±221.2	2145.3±1055
COL1A1	13.5	1.50E-13	125.4±96.9	1698.6±1360.4
COL1A2	4.6	1.07E-10	2443.1±1279.9	11195.9±6145
LAMA1	1.1	5.91E-04	43.7±6	48.5±6
TNC	2.9	0.0012183	862.4±535.4	2537.8±2580.6
VWF	-3.1	2.53E-11	4647.6±1467.5	1479.5±1131.8
FYN	-2.6	2.71E-10	421.8±158.3	165±115.1
PDGFRA	-1.3	7.42E-04	2110.7±752.4	1682.4±989.7
CAV1	-6.4	2.17E-14	9715.3±2470.9	1509±1142.2
CAV2	-5.1	1.12E-11	4455.4±1139.8	876.3±658.2
ITGA2	2.1	7.72E-08	269.1±90.6	576.7±364.1
ITGA5	-1.7	1.40E-07	880.6±498	517.1±360.2
ITGA8	-2.7	1.60E-09	552.3±203.9	203.5±157.2
ITGAV	1.4	4.10E-04	2059.6±491.8	2815.4±1051.6
ITGB5	1.3	0.0018841	519.2±88.9	659.2±190.2
ITGA2B	-1.3	7.16E-04	35.3±11.7	27.3±5.1
PAK1	1.7	1.05E-05	47.7±13.3	82.5±51.8
ROCK1	-1.4	2.15E-06	1952.1±387.7	1425.7±394.7
PXN	-1.5	1.47E-08	624.4±126.9	425±122.2
PPP1CB	-1.7	4.68E-12	1207.2±291.1	708.6±220.1
PPP1R12A	-1.5	1.63E-08	1162.9±225	776.2±215.6
PTEN	-1.4	7.32E-07	163.8±54.5	116.9±43.2

(A)



(B)

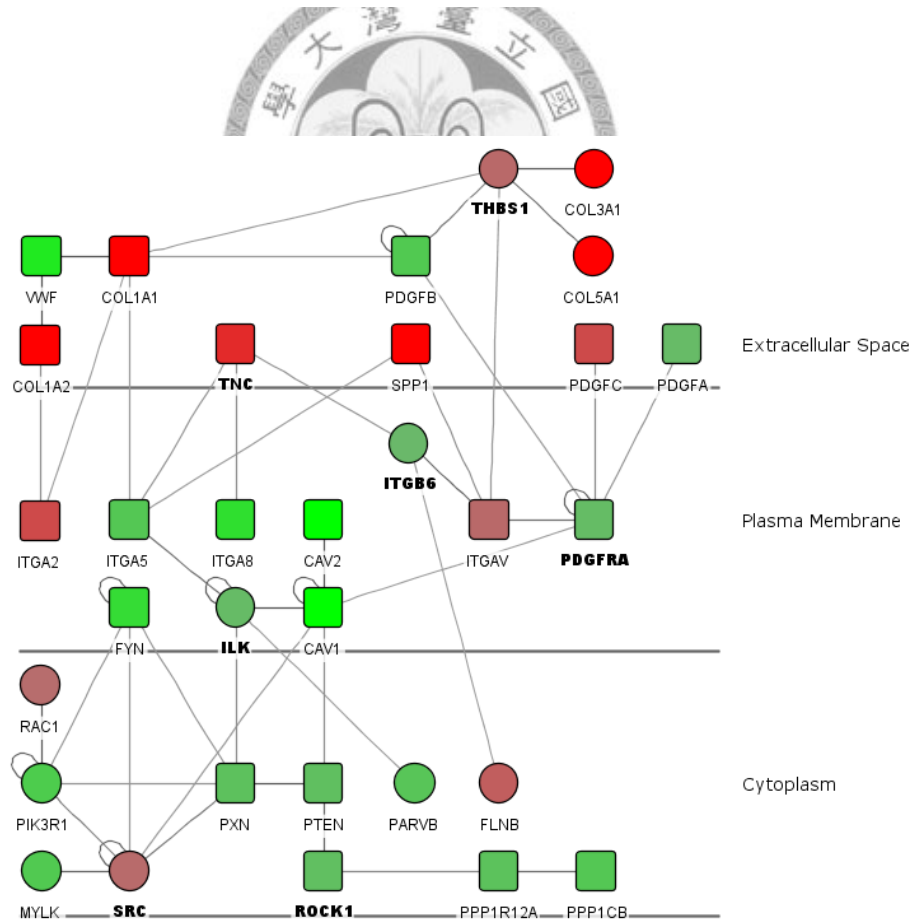


Figure A-3. Main component obtained by f_2 scoring method. Genes with symbol in bold face represent key nodes. Rectangles indicate genes also found by f_1 scoring method.

Table A-6. Annotations of genes present in Figure A-3.

Genes in the main component of cell cycle pathway. (Fig. S3A)					
Gene Symbol	Fold Change	p-value	Normal	Tumor	Key node
CDH1	2.4	1.26E-13	1994.5±381.8	4752.1±1690	no
CCNB2	5.6	5.75E-11	69.8±23.3	388.4±401.5	no
CDKN1A	-1.8	9.71E-05	2557.9±1746.5	1384.7±886.7	yes
BUB1	4.6	8.02E-09	31.6±9.9	144.9±144.1	no
BUB1B	6.0	1.56E-10	57.3±20.8	344.6±354.3	no
CDC2	3.5	1.33E-08	155.5±45.5	540.5±521.6	no
ORC6L	2.6	8.60E-09	125.2±33.6	327.4±219	no
MAD2L1	5.1	4.78E-08	56.3±23.3	284.9±315.1	no
CDC45L	2.3	5.14E-09	63.1±19.6	145.9±108.8	no
CCNB1	6.3	1.32E-08	59.4±20.7	371.6±482.8	no
CDC20	7.1	5.01E-10	58.9±15.4	415.8±460.9	no
CDC14A	-1.7	4.11E-08	37.3±11.2	22.3±7.6	no
E2F1	1.4	3.45E-05	114.2±15.3	155.1±71	no
E2F3	2.0	1.28E-07	270±87.5	543.4±295.3	no
E2F4	1.2	4.20E-07	606.6±140.3	739.5±213.7	no
HDAC1	1.9	1.20E-07	1181±222.6	2198.6±1099.5	no
TP53	1.6	8.12E-05	198.9±53.5	314.8±129.4	yes
HDAC3	1.3	7.30E-06	467.5±45	604.7±156.6	no
CCNE1	5.5	7.38E-06	99.2±24	548.6±1757	no
PTTG1	4.8	1.27E-09	476.2±166.2	2262.6±2763.6	no
HDAC4	-1.6	1.05E-07	260.4±75.8	161.8±80.4	no

Genes in the main component of focal adhesion pathway. (Fig. S3B)					
Gene Symbol	Fold Change	p-value	Normal	Tumor	Key node
SPP1	29.7	1.02E-15	310.8±476.6	9227.8±5885.9	no
PDGFA	-1.2	5.43E-04	96.8±28	80±24.5	no
PDGFB	-1.8	2.25E-09	264.9±80.4	147.2±59.4	no
PDGFC	2.1	4.57E-07	1006±221.2	2145.3±1055	no
COL1A1	13.5	1.50E-13	125.4±96.9	1698.6±1360.4	no
COL1A2	4.6	1.07E-10	2443.1±1279.9	11195.9±6145	no

COL3A1	4.5	1.24E-11	1266.8±787.7	5732.3±2430.5	no
COL5A1	4.2	1.56E-10	591±290.9	2470.4±1754.9	no
TNC	2.9	1.22E-03	862.4±535.4	2537.8±2580.6	yes
THBS1	1.4	1.97E-02	402.5±317.4	552.2±351.5	yes
VWF	-3.1	2.53E-11	4647.6±1467.5	1479.5±1131.8	no
FYN	-2.6	2.71E-10	421.8±158.3	165±115.1	no
ILK	-1.2	5.99E-05	1962.9±481.1	1609±364.9	yes
PDGFRA	-1.3	7.42E-04	2110.7±752.4	1682.4±989.7	yes
ITGA2	2.1	7.72E-08	269.1±90.6	576.7±364.1	no
ITGA5	-1.7	1.40E-07	880.6±498	517.1±360.2	no
ITGAV	1.4	4.10E-04	2059.6±491.8	2815.4±1051.6	no
ITGB6	-1.1	6.31E-02	121.9±136	109.9±201.1	yes
ITGA8	-2.7	1.60E-09	552.3±203.9	203.5±157.2	no
CAV1	-6.4	2.17E-14	9715.3±2470.9	1509±1142.2	no
CAV2	-5.1	1.12E-11	4455.4±1139.8	876.3±658.2	no
RAC1	1.3	1.21E-07	6237.6±726.4	7932.8±1485.3	no
MYLK	-1.8	2.76E-08	3154.5±1181.1	1746.1±1214.2	no
PIK3R1	-1.9	3.88E-09	1025.9±382.9	538.1±273.8	no
ROCK1	-1.4	2.15E-06	1952.1±387.7	1425.7±394.7	yes
SRC	1.3	2.08E-04	331.7±63.5	439.3±191.3	yes
FLNB	1.6	1.54E-08	331.7±93.7	546.5±222.3	no
PARVB	-1.6	1.48E-08	287.9±100.9	177.9±73.9	no
PXN	-1.5	1.47E-08	624.4±126.9	425±122.2	no
PPP1R12A	-1.5	1.63E-08	1162.9±225	776.2±215.6	no
PPP1CB	-1.7	4.68E-12	1207.2±291.1	708.6±220.1	no
PTEN	-1.4	7.32E-07	163.8±54.5	116.9±43.2	no

Table A-7. How the main components overlap with the leading edge subset.

NTUH	Cell cycle			Focal adhesion		
	size	overlap	percentage	size	overlap	percentage
main component(m=0)	19	15	0.79	25	16	0.64
main component(m=1)	21	15	0.71	32	20	0.63
leading edge subset	35	---	---	63	---	---

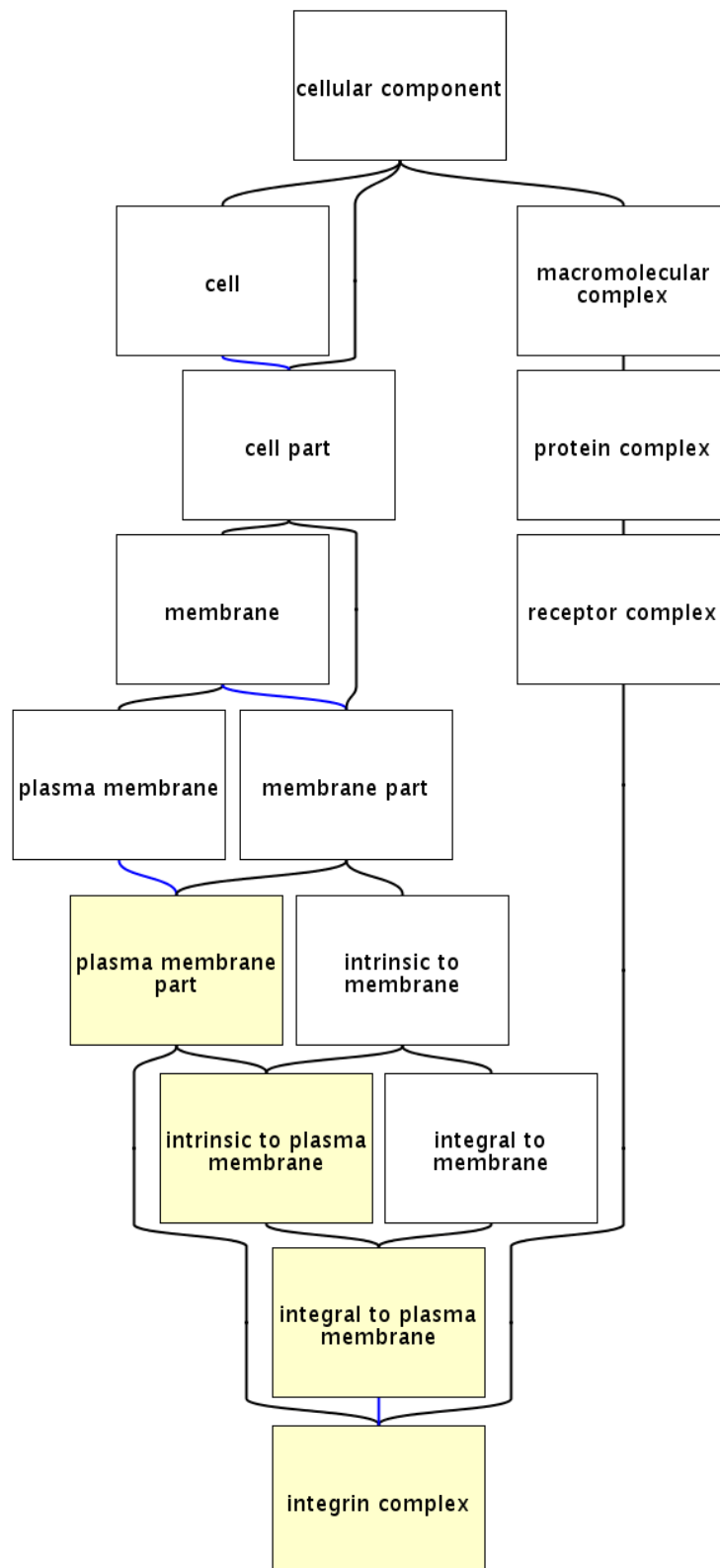


Figure A-4. GO term hierarchy for cluster C in Fig. 4-6. Rectangles filled with color represent the cluster enriched in focal adhesion pathway in terms of cellular component category in GO.

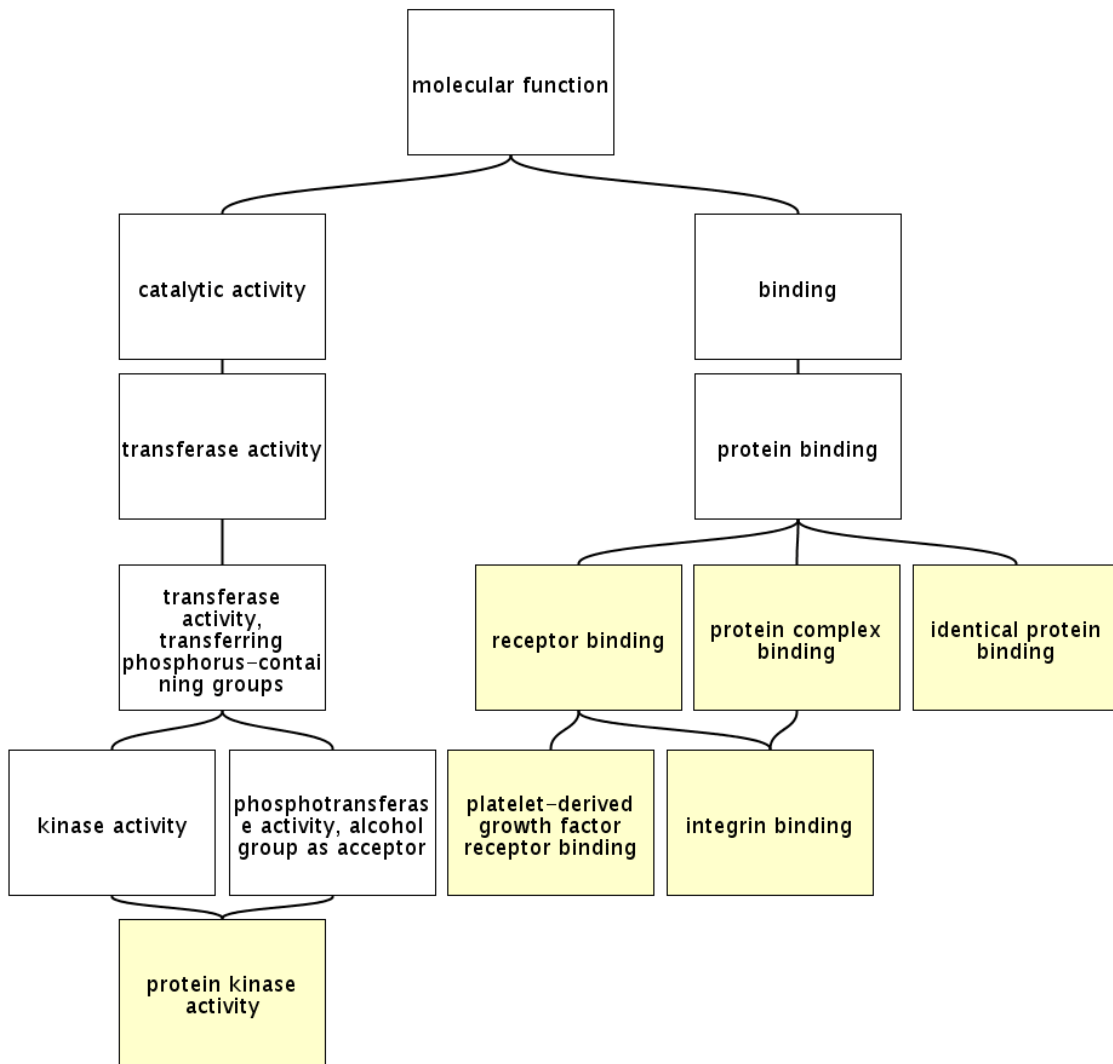


Figure A-5. GO term hierarchy for terms involved in Fig. 4-7.

Rectangles filled with color represent the terms enriched in focal adhesion pathway in terms of molecular function category in GO.

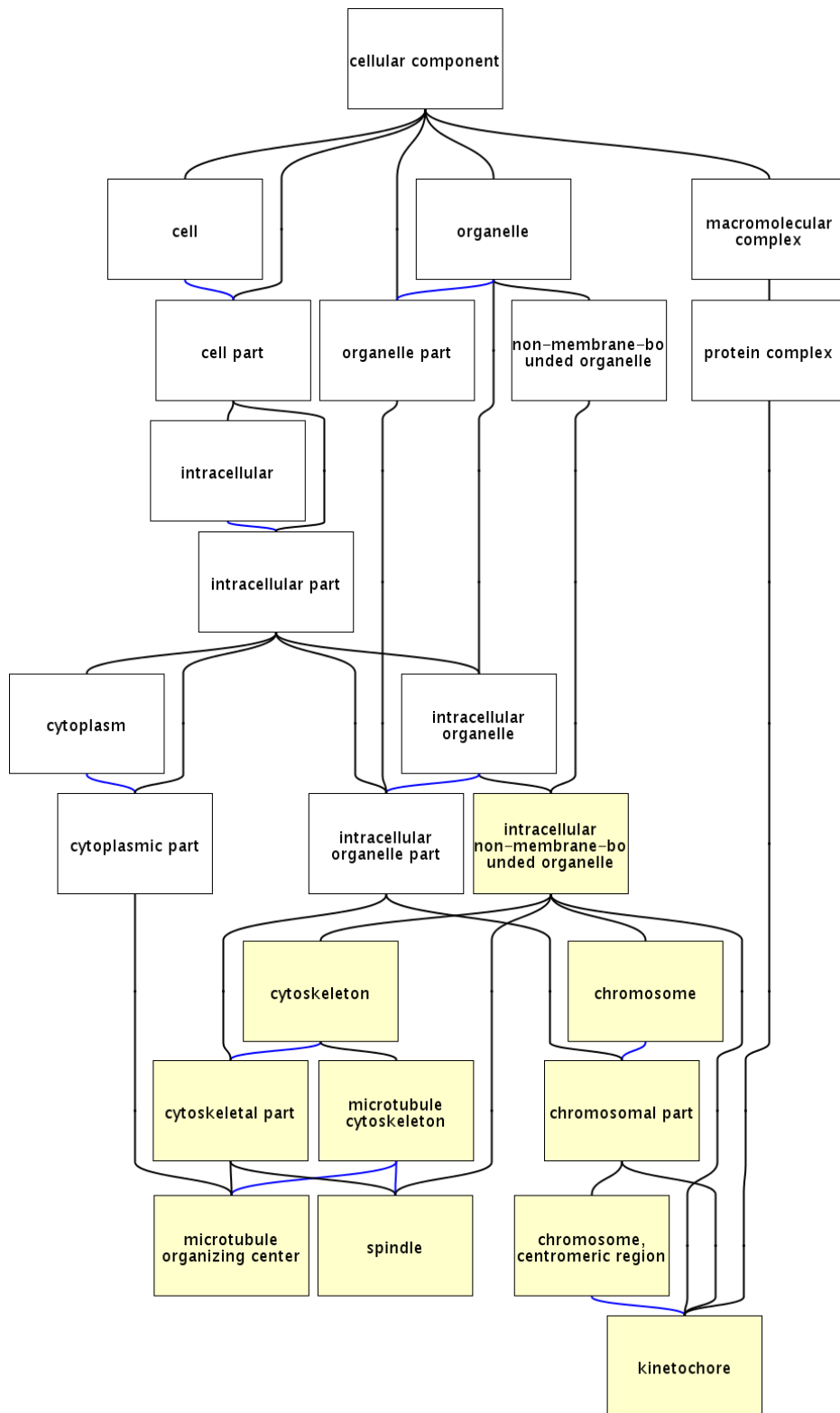


Figure A-7. GO term hierarchy for cluster C,D,E in Fig. 4-9.

Rectangles filled with color represent the the cluster enriched in cell cycle pathways in terms of cellular component category in GO.

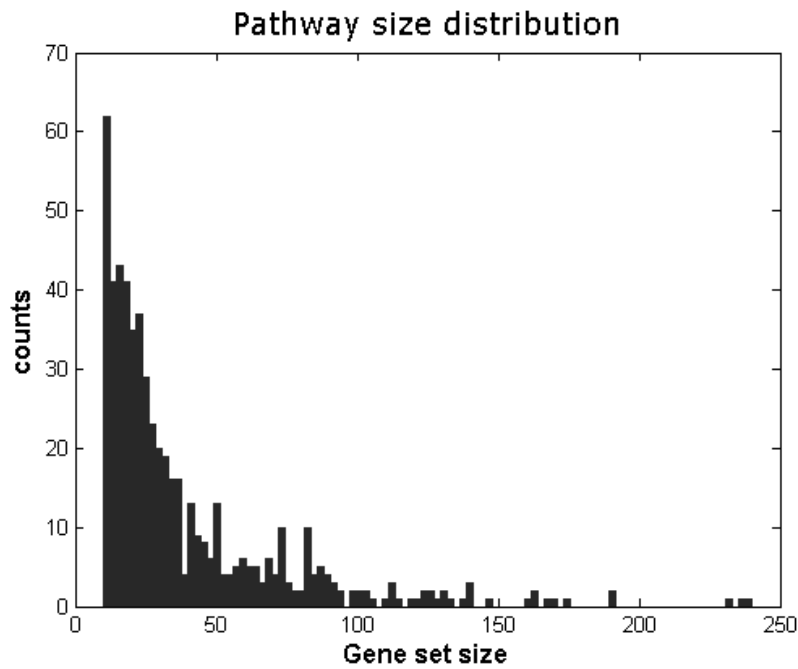


Figure A-8. Histogram of pathway sizes in our database.

Table A-8. Annotations of genes present in Figure 4-11.

Genes in the main component of cell cycle pathway. (Fig. 4-10A)					
Gene Symbol	Fold Change	<i>p</i> -value	Normal	Tumor	Member of Cell Cycle Pathway
S1PR1	-4.3	1.06E-14	1028.9±368.1	240.1±210.4	no
EDNRB	-7.4	3.80E-14	724±440	98.2±90.3	no
FYN	-2.6	2.71E-10	421.8±158.3	165±115.1	no
GRK5	-5.7	5.90E-17	2126±496.4	372.7±263.8	no
TEK	-5.9	8.99E-15	902.3±293.9	154.1±135.8	no
CAV1	-6.4	2.17E-14	9715.3±2470.9	1509±1142.2	no
CDH1	2.4	1.26E-13	1994.5±381.8	4752.1±1690	yes
CDH3	6.1	7.44E-13	156.3±80.1	949.4±683.9	no
CDH5	-4.5	1.78E-14	1356.6±453.8	299.7±214.9	no
PECAM1	-2.8	1.07E-13	7238.7±1866.3	2586.2±1354.8	no

PTPRA	1.1	7.21E-02	457.9±127.8	507.2±171.3	yes
PTPRB	-4.5	6.49E-16	435.5±194.9	95.7±68.1	no
PTPRM	-2.0	1.93E-13	878±140.9	437.5±148.9	no
CD36	-7.7	7.05E-15	1762.2±865.5	228.1±179.2	no
GAB2	-2.0	9.98E-11	1014.2±158.4	518.8±180.3	no
PTPN11	-1.3	1.65E-10	2514±333.8	1905.2±343	no
ABL1	-1.1	4.25E-01	681.6±260.8	636.2±205.1	yes
H3F3A	-2.7	3.03E-14	1700.5±410.7	632.7±261.5	no
NEDD9	-2.6	2.03E-12	1030.6±359.5	399.4±222	no
SKP2	2.2	8.03E-05	240.5±66.3	521.9±454	yes
TAL1	-5.9	6.97E-16	494.7±281.7	83.5±87.9	no
CBFA2T3	-2.6	1.74E-13	172±54.8	67.3±29.9	no
E2F2	-1.1	6.38E-02	38.5±5.3	36.6±5.8	yes
EP300	-1.2	2.35E-03	124.4±31.4	106.3±29.4	yes
HDAC1	1.9	1.20E-07	1181±222.6	2198.6±1099.5	yes
HDAC2	1.3	8.48E-02	910.7±136.8	1150.9±548.1	yes
TCF3	1.6	5.96E-11	120.6±19	187±48.3	no
SMARCA5	-1.4	7.64E-10	1337.3±179.8	968.8±223.4	no
SMC1A	1.5	2.48E-06	1249.1±152.1	1858.2±710.7	yes

Genes in the main component of focal adhesion pathway. (Fig. 4-10B)

Gene Symbol	Fold Change	<i>p</i>-value	Normal	Tumor	Member of Focal Adhesion Pathway
FIGF	-6.6484	6.54E-13	1370.9±422.7	206.2±194.5	yes
VCL	-1.1844	9.24E-06	32.1±7.9	27.1±5	yes
ADRB2	-4.2686	3.33E-14	1006.5±387.9	235.8±160.3	no
S1PR1	-4.2861	1.06E-14	1028.9±368.1	240.1±210.4	no
EDNRB	-7.3741	3.80E-14	724±440	98.2±90.3	no
ERBB2	1.6518	5.70E-07	904.2±175.7	1493.7±638.1	yes
FYN	-2.5571	2.71E-10	421.8±158.3	165±115.1	yes
GRK5	-5.7049	5.90E-17	2126±496.4	372.7±263.8	no
KDR	-3.0059	8.41E-13	1263.7±462.6	420.4±229.3	yes
TEK	-5.8555	8.99E-15	902.3±293.9	154.1±135.8	no
CAV1	-6.4383	2.17E-14	9715.3±2470.9	1509±1142.2	yes
CAV2	-5.0843	1.12E-11	4455.4±1139.8	876.3±658.2	yes
CDH1	2.3826	1.26E-13	1994.5±381.8	4752.1±1690	no
CDH3	6.0747	7.44E-13	156.3±80.1	949.4±683.9	no
CDH5	-4.5269	1.78E-14	1356.6±453.8	299.7±214.9	no
PECAM1	-2.799	1.07E-13	7238.7±1866.3	2586.2±1354.8	no
PTPRB	-4.5494	6.49E-16	435.5±194.9	95.7±68.1	no
PTPRM	-2.0071	1.93E-13	878±140.9	437.5±148.9	no
CD36	-7.7251	7.05E-15	1762.2±865.5	228.1±179.2	no
SRC	1.3244	2.08E-04	331.7±63.5	439.3±191.3	yes
PXN	-1.4692	1.47E-08	624.4±126.9	425±122.2	yes
PTEN	-1.4013	7.32E-07	163.8±54.5	116.9±43.2	yes
PTPN11	-1.3195	1.65E-10	2514±333.8	1905.2±343	no
H3F3A	-2.6879	3.03E-14	1700.5±410.7	632.7±261.5	no
NEDD9	-2.5803	2.03E-12	1030.6±359.5	399.4±222	no
CORO2B	-3.2862	1.53E-14	144.8±54	44.1±21.6	no