臺北醫學大學醫學資訊研究所

碩士論文

Gene Expression Analysis for Cancer-Related Genes Using Public Microarray Databases

分析公開微陣列資料庫中癌症相關基因的基因表現

指導教授：CHIU, HUNG-WEN 邱泓文

研究生： CHEN, LILLIAN YU-HSUAN 陳宇瑄 撰

中華民國九十五年一月

January, 2006

Acknowledgements

Two and half years passed without even noticed. Although there were some setbacks and frustrations along the way that almost made me to give up my study, I am glad that I did try to hang in there, or else I would definitely regret in the future.

When I looked back the days I spent to completing my master degree, I cannot ignore the greatest support from my beloved family members, including Daddy, Mommy, Brother Cliff, Aunties and Grandma. Without your warm care and continuous encouragement, I would not make to this point at all.

I cannot fully express my gratitude to the advisor, Dr. Hung-Wen Chiu, for his superb guidance and generosity throughout the past 2.5 years. Without Dr. Chiu's generous assistance, I would not complete this master thesis at all.

My appreciation also to the dearest friends and colleagues, including Kiss, YiFen, Cool8, Ted, amebajoe, FHHUNG and ncrain from GIMI; Debby, Mike, Paul, Elisa and Jack from UBC; David, Sylvia, Grace, Kevin, Dr. Shau and Dr. Luke Lin from GSK. Lastly, very special thanks to Linus, who believes in me and is there for me all the time to share ideas and offer constant support.

Written on 27 Jan 2006

At Graduate Institute of Medical Informatics, Taipei Medical University

Table of Contents

List of Figures

List of Tables

論文摘要

論文名稱：分析公開微陣列資料庫中癌症相關基因的基因表現

臺北醫學大學醫學資訊研究所

研究生姓名： 陳宇瑄

畢業時間： 94 學年度 第 1 學期

指導教授： 邱泓文 臺北醫學大學醫學資訊研究所 副教授

背景：

癌症是現今世界共通的疾病，發展有效治療癌症的藥物是當下醫藥學界主要致力的目標。微陣列技術的應用已成為近幾年生醫研究的主流，許多研究利用微陣列技術來觀測人類癌症細胞株或器官組織的基因表現，藉此有效率得找出癌症相關基因。

材料和方法：

我們從 OMIM 資料庫中做資料探勘癌症相關的基因，整理得到一個癌症相關基因名單。另一方面從 GEO 和 SMD 資料庫中蒐集人類癌症相關的微陣列數據資料，並根據不同種類型癌症做分類，再利用統計原理篩選出可能的基因名單。分析這些基因名單的生化反應路徑，和比較不同資料庫來源

所得到結果的差異。

結果與討論：

我們將由 OMIM 所得到十種不同癌症的基因名單做交集，得到三個共通的癌症相關基因，分別為 APC、CDKN2A 和 PTEN。針對乳房、前列腺、肝、肺、胰臟、胃等六種器官的癌症組織相關基因名單做交集，觀察到 APC、CDKN2、APTEN、TP53 和 BRAF 是共通的癌症相關基因。另外，利用微陣列資料庫中所得到的癌症相關基因名單做交集，寡核柑酸微陣列的資料中我們得到九個基因，分別是 FOXM1、HNRPDL、BIN1、BUB3、CCNI、PMS1、PRKCBP1、PURA 和 RPA3。在 cDNA 微陣列部分，得到六個共通的癌症相關基因，分別為 ARGBP2、CD53、FCGBP、JUN、MME 和 VBP1。在這些我們所觀察到的癌症共通基因都在 *Wnt signaling pathway*.中扮演重要的角色。

結論：

將 OMIM 十種主要的癌症基因名單作交集，我們得到 3 個共通的癌症相關基因。另外再交集六種癌症基因名單得到五個共通的癌症相關基因，並比較這些基因在微陣列中的表現是否有所差異。結論是在 OMIM 中找到的癌症相關基因與微陣列實驗數據的結果不一定會相符合。

關鍵字：癌症、癌症相關基因、微陣列、OMIM、GEO、SMD

Abstract

Title of Thesis：Gene Expression Analysis for Cancer-Related Genes Using Public Microarray Databases

Author：Lillian Yu-Hsuan Chen

Thesis advised by ： Hung-Wen Chiu

Taipei Medical University,

Graduate Institute of Medical Informatics

Introduction:

As cancer has drawn much of the attention worldwide these days, development of effective drugs is definitely the focus in today’s medical research field. Since microarray technologies have become a biological research trend over the last few years, using the microarray data to monitor gene expression in human cell lines and tissues is certainly the most efficient way to identify cancer-related genes.

Materials and Methods:

The cancer-related gene lists were obtained by reviewing literatures on the OMIM database. The microarray expression datasets were downloaded from the GEO and SMD websites. After having collected the cancer-related genes and microarray expression data, we would classify them according to each datum’s specific cancer-causing nature.

Results and Discussion:

When having intersected ten OMIM cancer-related gene lists, APC, CDKN2A and PTEN were resulted as the three common cancer-related genes; when having performed the intersection of breast, prostate, liver, lung, pancreatic and stomach tissues, APC, CDKN2A, PTEN TP53 and BRAF were obtained. Based on microarray gene expressions, intersections of cancer-related genes among oligonucleotide arrays have found nine common genes, which are the FOXM1, HNRPDL, BIN1, BUB3, CCNI, PMS1, PRKCBP1, PURA and RPA3. Intersections of cancer-related genes among cDNA arrays have got six common genes, which are the ARGBP2, CD53, FCGBP, JUN, MME and VBP1. Many of those defined cancer-related genes were found to play important roles in *Wnt signaling pathway*.

Conclusion:

Three OMIM cancer-related genes across ten cancer types were defined while five OMIM cancer-related genes were obtained as a result of intersection of six cancer types. OMIM mentioned cancer-related genes are not necessarily supported by microarray gene expression patterns.

Keywords: Cancer, Cancer-Related Genes, Microarray, OMIM, GEO, SMD

I. Introduction

1.1 Background

Nowadays cancer has become one of the deadly diseases affecting people's life worldwide. Based on the statistical reports published by World Health Organization (WHO), cancers in lung, colorectal and stomach are the three major cancer types that affected lives in both sexes for many years globally. Lung and stomach cancers are the life-threatening factors in male populations as to that of breast and cervical cancers in female populations. WHO has estimated that approximately over ten million people will be diagnosed with cancer on an annual base. By year 2020, however, approximately fifteen million of new cancer patients will be reported annually (Yang et al., 2002).

According to the published data for top ten leading causes of death by Taiwanese Department of Health for year 2004, cancer is once again ranked the top leading cause of death for a consecutive of twenty-two years. Among various cancer types, cancers in lung (19.67%), liver (19.42%), colorectal (19.73%), breast (3.68%, which is only calculated based on female population) and stomach (6.88%) have ranked the top five deadly diseases in Taiwan. In addition to the humiliate statistics, on average approximately every fifteen minute would a Taiwanese lose his/her life due to cancer (DOH Website, 2004).

As cancer has drawn great attention and focus in the medical field due to its life-threatening nature, many researchers around the world have dedicated their time to look for a better cancer treatment. Until today, however, there have been no perfect medications being developed yet. Effective treatments on cancer patients basically rely on a better understanding of the tumour genes in relation to the specific cancer type. Scientists have used high-throughput methods to define the relationship between cancer and genes. Gel electrophoresis, microarray technology and serial analysis of gene expression (SAGE) are the most popular ones known today. Among them all, microarray technology is most well-known for its ability to determine the cancer gene expressions in related to the cancer types (Liotta et al., 2000; Nelson et al., 2000; SEER's Training Website, 2005).

1.2 Motivation

The current cancer research trend favours the idea that genetic mutation has driven the initial formation of malignant tumours. It is generally believed that cancer begins at the cellular level, in which the disease actually initiates in a single cell that will eventually pass its acquired abnormality onto its progeny (Lu et al., 2003). Based on Aranda-Anzaldo's view, those initiated cells must contain a few "caner-causing genes" in their DNA. It is very possible that those caner-causing genes may have remained in latent stages for a long time, and are waiting to be triggered by any cancer-promoting agents. Even if caner-causing genes are not activated at all, or do not transform into lethal cancerous cells, they still possess certain degrees of dangerous factors that might affect people's life (Aranda-Anzaldo et al., 2001).

Within the OMIM database, many literatures have been reviewed and categorized into different groups based on their research topics and contents by scientists at John Hopkins University. Articles that include information on genes that have caused cancers can easily be sorted out by having limited the search result to "cancer", "carcinoma" and "tumor". Moreover, a list of cancer-related genes can be resulted from reading through these articles (OMIM, 2000). On the other hand, microarray technologies have become a biological research trend over the last few years for monitoring gene expression in human cell lines and tissues. Previous understanding of gene expression levels in different cancer types by microarray hybridization have provided an idea that this is indeed a useful and eventually will be an essential method to identify possible biomarkers as well as drug targets.

Nowadays, most microarray gene expressions are used by worldwide researchers to categorize different cancer types. Moreover, literatures found in OMIM database do reveal that different cancer types have possessed different microarray gene expressions. Based on those two understandings, we would like to find out whether there will be one or more genes that are related to various types of cancers at once by using OMIM literatures as our evident cancer-gene finders and combining with microarray gene expressions to confirm our thoughts.

1.3 Objective

Our goal is to focus on the cancer-related genes mentioned in the literatures. We would like to identify the gene-tissue relationship as well as how those genes are expressed in normal and cancer tissues. Based on the cancer-related gene list obtained from OMIM database, we would match those genes with the gene expressions from the microarray datasets. At the same time, we would also determine cancer-related gene lists for microarray expression datasets. Following the collection of all the relevant data, including cancer-related genes from both OMIM database and publicly accessible microarray databases, and microarray expression datasets, we would further analyze and look for any relationships of those cancer-related gene lists with the biological pathways via KEGG database. When the determination of the relationship between cancer-related genes and pathways completed, we would like to see if there is one or more genes that are located in the upper stream of the pathway. By this means, medical researchers can develop both prevention and more effective cancer treatments that are specifically targeted on those genes.

II. Literature Review

2.1 Cancer

In every healthy human body, ten million cells will undergo normal cell division every minute. When a cell has undergone mutations in its deoxyribonucleic acid, or DNA, the genetic material which carries the hereditary codes for human body, it will become a cancerous cell which will reproduce without restraint. In other words, cancerous cells not only divide faster than that of normal cells, but also grow indefinitely and immaturely (Affymetrix et al., 2001; Lu et al., 2003).

As time passes, a single cancerous cell eventually grows into a microscopic collection of cells and ultimately begins to invade surrounding tissues. Each cancer has its own distinctive course. For example, in leukemia, the abnormal cells disperse throughout the body via blood streams and bone marrow. Most of the other cancer types, however, a mass of cancer cells called tumours grow freely in their rate. Some tumours may double their size in a month while others may require two months or even more than a year to double (Nelson et al., 2000; SEER’s Training Website, 2005).

Tumours can be categorized into two types: benign and malignant tumours. Benign tumours remain localized to the tissue where they arise; they may grow large but will not spread to other parts of the body. If they are diagnosed in earlier stages, they can be cured by surgical removals or by radiation therapy. On the other hand, malignant, or cancerous tumours are a more serious matter. Some of their cells might break off, invading and destroying surrounding tissue or traveling through the blood or lymph streams to distant parts of the body, where new tumours might form. From these new tumours, malignant cells could break off again and establish even more colonies, in which the invasive process is known as metastasis. For example, breast cancer and lung cancer have possessed many different characteristics. However, when metastatic breast cancer in the lungs is observed, the lung cancer characteristics are not easily observed under a microscope. The cancer in lung acts just like a cancer originated in the breast. Thus, it is worth noted that it is important to understand that cancer originating in one body organ takes its characteristics with it even if it spreads to another part of the body (Liotta et al., 2000; Nelson et al., 2000; SEER’s Training Website, 2005).

2.2 OMIM

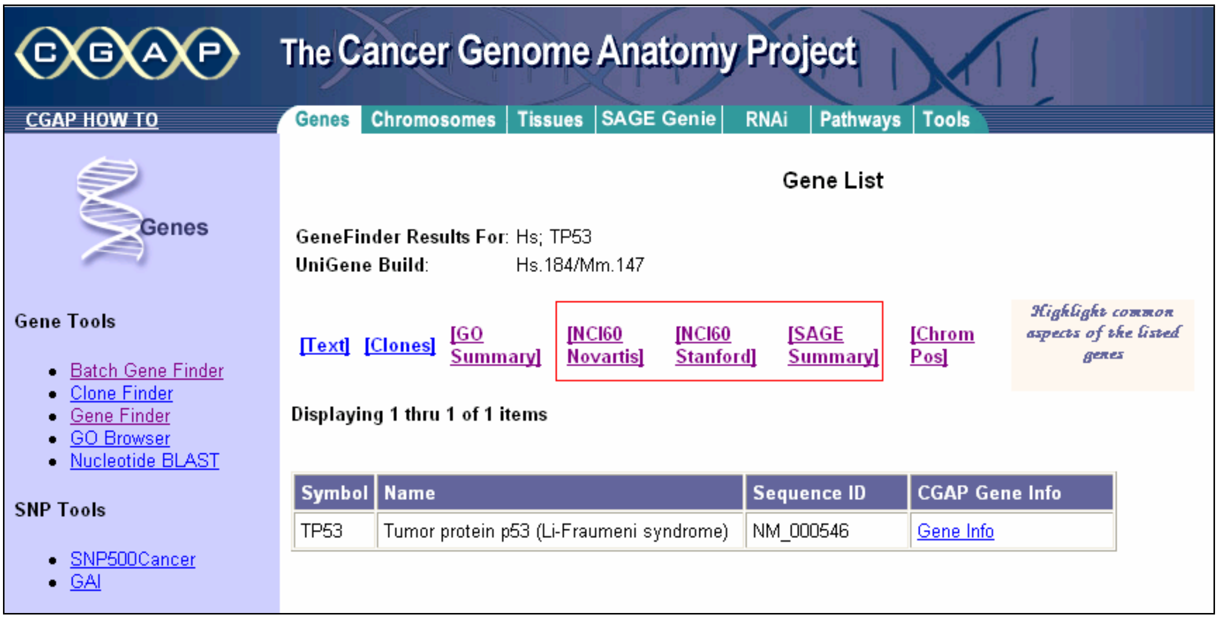
OMIM, short for Online Mendelian Inheritance in Man, is a constant updated catalog of human genes and inherited, genetic disorders authored and edited by Dr. Victor A. McKusick and his coworkers at John Hopkins University. The database, provided by the National Center for Biotechnology Information, can be publicly accessible through the World Wide Web at: http://www.ncbi.nlm.nih.gov/omim/. The OMIM contains not only a variety of textual information and references, but also links to records in the Entrez system and relevant resources at MEDLINE plus the NCBI databases. As to 25 Jun 2005, OMIM has included a total of 16,115 entries and 9,288 loci entries for the synopsis of the Human Gene Map (OMIM, 2000).

Each OMIM entry is assigned to a unique six-digit number in which the first number indicates the inheritance mode of the gene involved. For example, 100000- and 200000- both refer to autosomal loci or phenotypes created before 15 May 1994. Numbering of 300000- means the x-linked loci or phenotypes while 400000- means the y-linked loci or phenotypes. Mitochondrial loci or phenotypes are given the numbering of 500000- while autosomal loci or phenotypes created after 15 May 1994 are numbered starting with 600000-. The allelic variant is named after its parent entry, followed by a decimal point and a distinct four-digit variant number. For example, beta-globin locus (HBB) is numbered 141900 as the sickle hemoglobin is numbered 141900.0243.

2.3 Cancer Genome Anatomy Project

The Cancer Genome Anatomy Project (http://cgap.nci.nih.gov/), also known as CGAP, is a huge task sponsored by the U.S. National Institutes of Health. CGAP is aimed not only to determine, catalog and annotate genes that are expressed during the cancer developmental process, but also to eventually improve detection, diagnosis and treatment for the cancer patients. With the cooperative work from researchers worldwide, CGAP wants to both increase the scientific expertise and enlarge its databases so that all cancer researchers can be benefited from it (CGAP Website, 2005).

CGAP has incorporated various searching tools, including tools to find genes, cDNA libraries, single nucleotide polymorphisms (SNPs), and tools to examine gene expressions and chromosomes, in order to meet each researcher's need. For example, if we would like to check gene expression profile for TP53, we can first go to the "GeneFinder" function and key in "TP53" as my search item. The search results have shown to have three choices for which gene expressions can be viewed visually (Figure 1). The expression data has displayed in different array formats: "NCI60_Novartis" is the gene expression data of NCI 60 cell lines on oligonucleotide array; "NCI60_Stanford" is the gene expression data of NCI 60 cell lines on spotted arrays; "SAGE Summary" data is a 2-dimentional display of a common tissue and histology, such as brain cancer vs. brain normal, lung cancer vs. lung normal. Having clicked on any of the three choices will give the gene expression data for TP53 in different array format. Figure 2 has shown the visualization of TP53 gene expression in "NCI60_Stanford". The colouring scales have indicated that higher expression levels will be in red colour while lower expression level will be in blue instead.



Gene List

GeneFinder Results For: Hs; TP53
UniGene Build: Hs.184/Mm.147

[Text] [Clones] [GO Summary] [NCI60 Novartis] [NCI60 Stanford] [SAGE Summary] [Chrom Pos]

Highlight common aspects of the listed genes

Displaying 1 thru 1 of 1 items

Symbol	Name	Sequence ID	CGAP Gene Info
TP53	Tumor protein p53 (Li-Fraumeni syndrome)	NM_000546	Gene Info

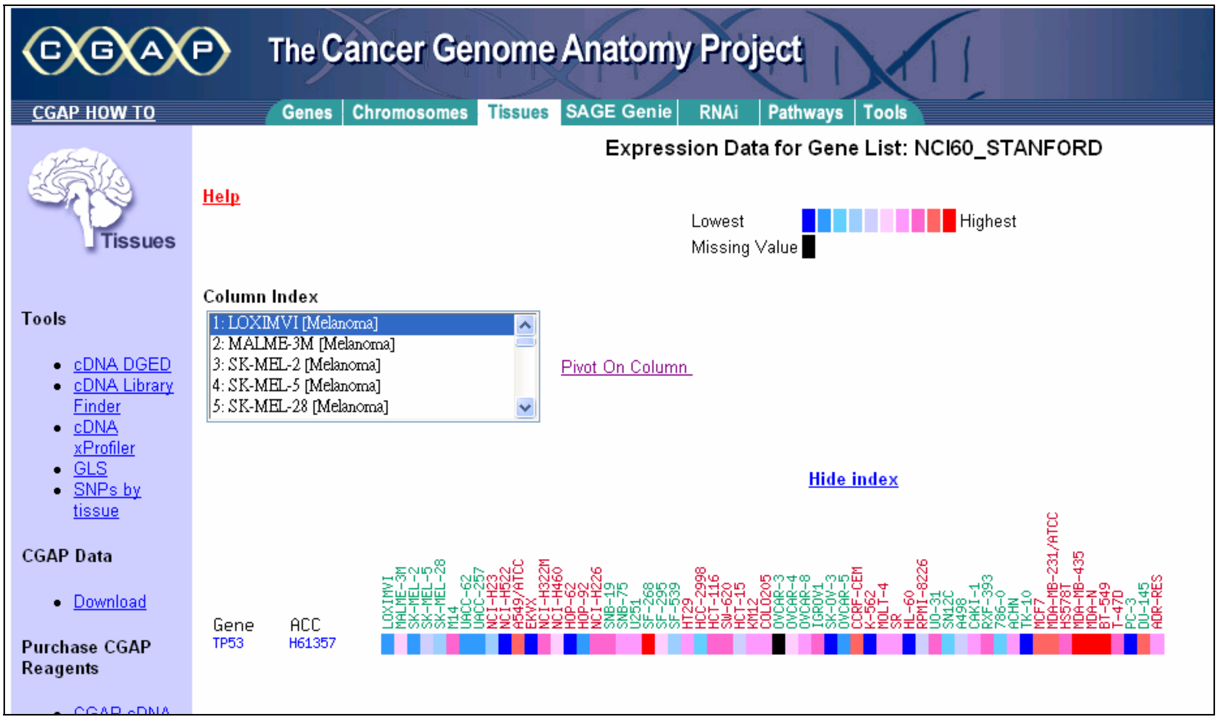
Gene Tools

- Batch Gene Finder
- Clone Finder
- Gene Finder
- GO Browser
- Nucleotide BLAST

SNP Tools

- SNP500Cancer
- GAI

Figure 1: GeneFinder search result for TP53 gene. The red box has indicated the three visualization selections for gene expression of TP53. (CGAP Website, 2005)



Expression Data for Gene List: NCI60_STANFORD

Help

Column Index

1: LOXIMVI [Melanoma]
2: MALME-3M [Melanoma]
3: SK-MEL-2 [Melanoma]
4: SK-MEL-5 [Melanoma]
5: SK-MEL-28 [Melanoma]

Pivot On Column

Hide index

Gene ACC
TP53 H61357

Tools

- cDNA DGED
- cDNA Library Finder
- cDNA xProfiler
- GLS
- SNPs by tissue

CGAP Data

- Download

Purchase CGAP Reagents

- CGAP cDNA

Figure 2: Visualization of TP53 gene expression in NCI60_Stanford. (CGAP Website, 2005)

2.4 Microarray

Transcription of DNA into RNA and the subsequent translation of messenger RNA into protein are the basic mechanisms by which cells mediate their growth, function and metabolism. After the human genome has been sequenced and annotated successfully a few years ago, the next step in functional genomics is to analyze the transcriptome, which can be defined as a complete collection of transcribed elements of the genome. In addition to messenger RNAs (mRNAs), the transcriptome can also represent non-coding regions of RNAs whose main functions are of structural and regulatory purposes. Alterations in the structure or expressions levels of any one of these RNAs or their proteins eventually will contribute to disease occurrences (Nelson et al., 2000). The use of microarray technologies to monitor gene expressions in organisms, cell lines, and human tissues has become very important in today's biological research field (Schadt et al., 2000). The most well-known technologies developed to examine gene expressions of thousands of genes are the cDNA microarrays and oligonucleotide arrays. These two techniques are most famous for their ability to compare and contrast expression levels across various tissue types (Gibson et al., 2002). There are a few major differences between cDNA and oligonucleotide microarrays. One difference is that cDNA microarrays only provide gene expression data in relative values as to that of absolute data values provided by oligonucleotide arrays. Another variation would be the difference in the design of the array since cDNA microarrays uses PCR-amplified cDNA fragments (ESTs) extracted from a sequenced cDNA library compared to oligonucleotide microarrays uses a series of 25-mer oligonucleotides to represent known or predicted open reading frames (Gibson et al., 2002; Lipshutz et al., 1999; Wilson et al., 2003).

2.4.1 cDNA Microarray Technology

cDNA microarrays is designed to monitor relative gene expression levels of thousands of genes in cells simultaneously. In a typical cDNA microarray chips, PCR-amplified cDNA fragments, also known as expression-sequence tags (ESTs), are spotted at high density, usually at 10-50 spots per mm^2, onto a glass microarray slide (Gibson et al., 2002). The two different mRNA samples derived independently will be transcribed into reverse-cDNA and labeled using two different fluorescents, which usually are a red fluorescent dye Cy5 and a green fluorescent dye Cy3. The labeled cDNA populations will then hybridized simultaneously to the glass microarray slide (Yang et al., 2002). Red and green laser beams will scan the microarray slides separately, and the signal intensity values observed from the two scans are calculated for individual cDNA spots by having the intensity levels of the experimental samples (Cy5) divided by the intensity levels of the reference sample (Cy3). As a result, each derived gene expression level is a relative ratio for the cDNA spot in the sample (Figure 3).

The relative ratio obtained from cDNA microarrays has possessed a central idea that it is the change in relative level of expression that is of biological interesting. Genes with greater expression level do not mean that they have higher fluorescence intensities than genes with lower expression levels. The reason is that the fluorescence intensity is dependent on the length of the EST, the amount of label incorporated into the cDNA during the reverse transcription process, the preparation of DNA concentration for the particular clone and the efficiency of hybridization (GEO Website, 2005).

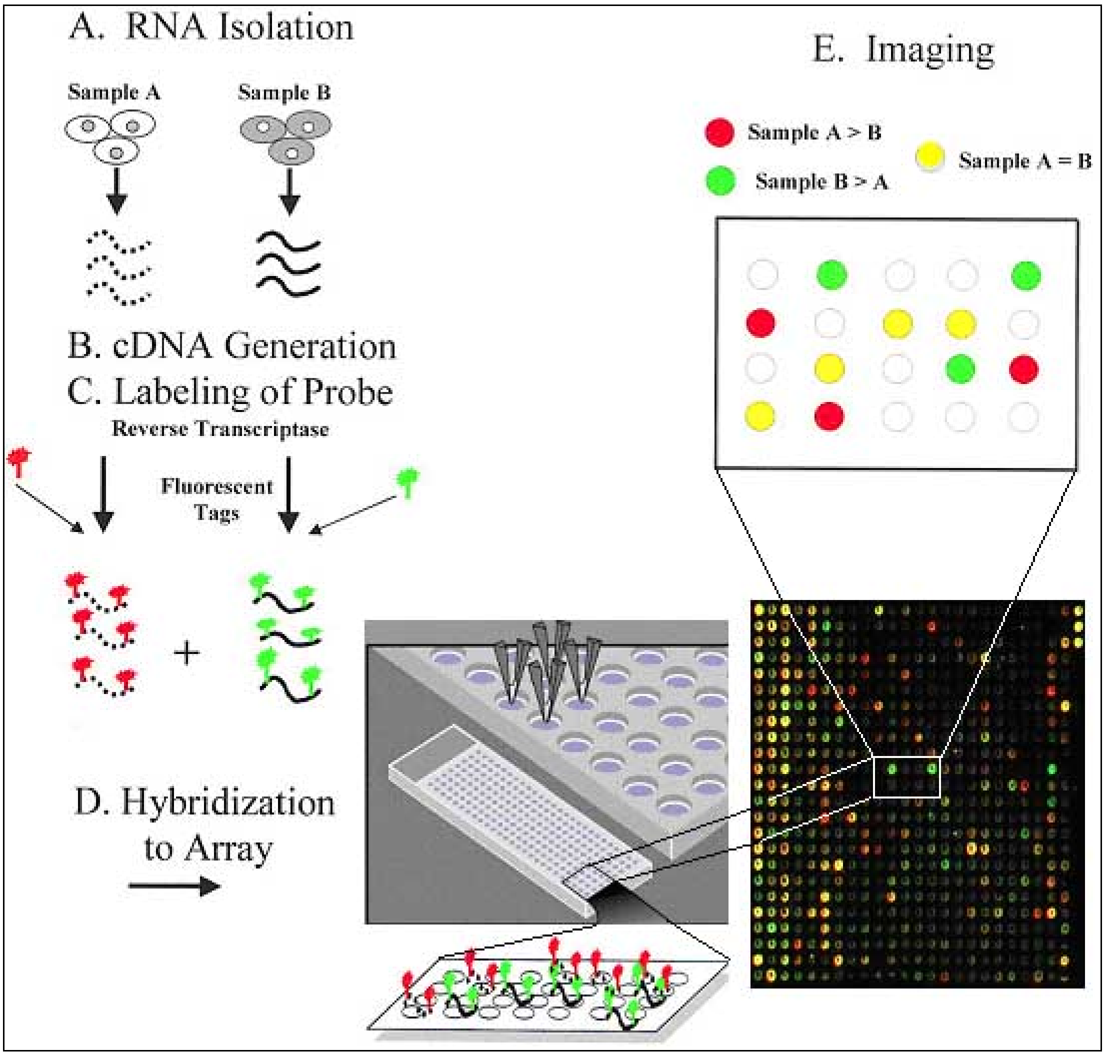


Figure 3: Overview Process of Making cDNA Microarray Chips (www.fao.org).

According to Claverie, the meaningful change in gene expression can be determined by the twofold induction or repression of experimental samples relative to the reference sample. This rule, however, does not meet the standard statistical definitions of significance. As a result, genes in cDNA microarrays will be classified as "differentially expressed" only if they have shown at least a 2-fold change in expression (Claverie et al., 1999).

$$\frac{GeneA}{GeneB} \geq 2 \quad or \quad \frac{GeneA}{GeneB} \leq \frac{1}{2}$$

2.4.2 Oligonucleotide Microarray Technology

High-density oligonucleotide arrays are built, or synthesized *in situ* on a silicon chip by Affymetrix. Each gene is uniquely represented by 10 to 20 different nucleotides on a probe array. Probe synthesis takes place in a parallel fashion, in which an A, T, C, or G nucleotide will be added to multiple growing chains simultaneously. After having undergone through a series of photolithographic and combinatorial chemical process, each probe will reach its particular length of 25 nucleotides (Lipshutz et al., 1999; Schadt et al., 2000) (Figure 4).

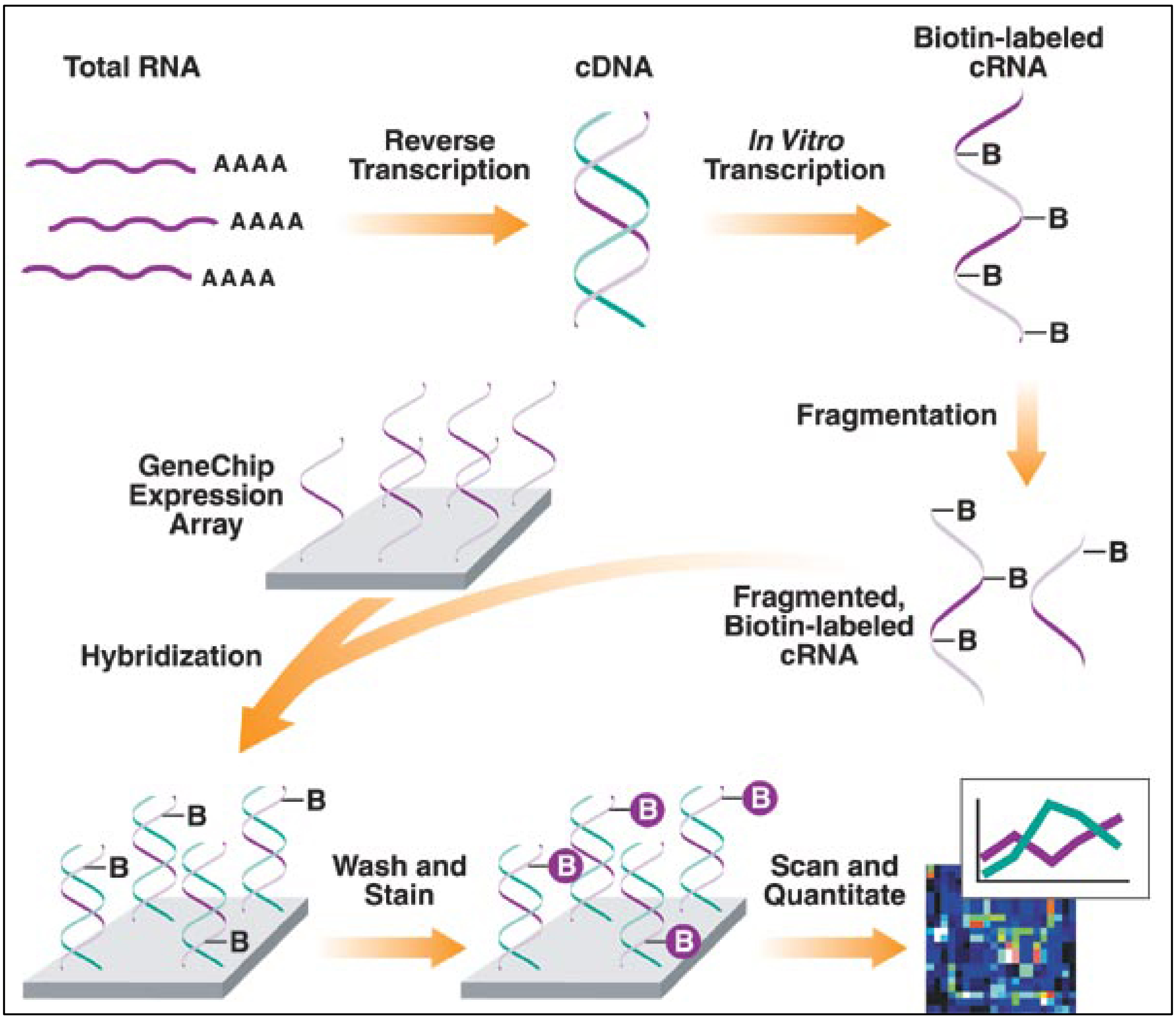


Figure 4: Oligonucleotide microarray technology. (Lipshutz et al., 1999).

In order to prevent the possibility of having cross-hybridization with similar short sequences in transcripts rather than the one being probed, a partner probe is designed to be perfectly complementary to the target probe except that a single base in its centre will be purposely mutated, resulting in a mismatch probe (MM). As shown in Figure 5, each Mismatch (MM) probe, also known as partner probe, will be paired with a complementary Perfect Match (PM) probe, also known as reference probe, and these two probe pairs allow the quantization and subtraction of intensity signals caused by non-specific cross-hybridization (Gibson et al., 2002; Lipshutz et al., 1999; Schadt et al., 2000). In oligonucleotide arrays, the expression level of each gene is calculated based on the average of the differences between PM and MM, which means the derived value of each gene expression level is an absolute amount in oligonucleotide arrays rather than that of relative ratio in cDNA arrays (Schadt et al., 2000).

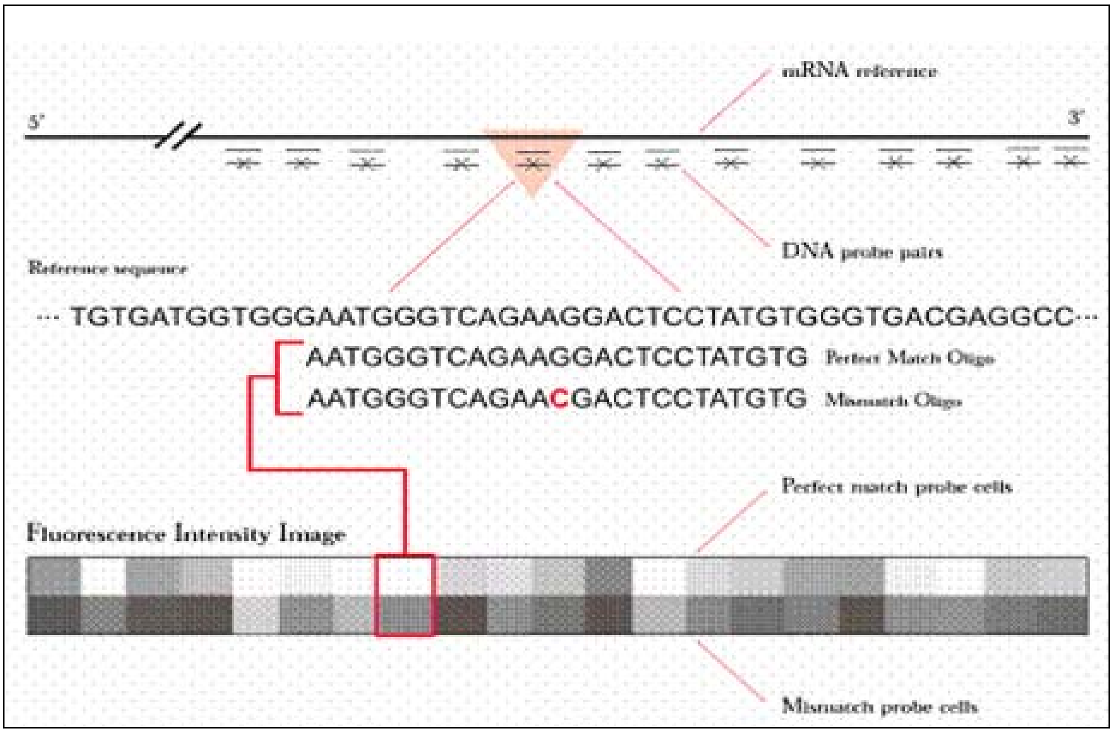

mRNA reference
5'
3'
DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC···
AATGGGTCAGAAGGACTCCTATGTG Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTG Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 5: Oligonucleotide Probe Pair Design. Oligonucleotide probes are chosen based on composition design rules, whereas proves for eukaryotic organisms are chosen particularly from the 3’ end. The use of the PM–MM differences averaged across probe sets has reduced cross-hybridization problems and increased the quantitative accuracy (Lipshutz et al., 1999).

2.5 Microarray Databases

Nowadays, a number of microarray databases are available for public access. Each public microarray database has its unique features and data sources, and two major ones in which most of the microarray data have been incorporated into would be introduced here.

2.5.1 Stanford Microarray Database

The Stanford Microarray Database (http://genome-www.stanford.edu/microarray), also known as SMD, is a research tool designed for scientific people to study biomedical problems using multiple microarray platforms. Today, SMD supports the research of more than 1,000 users in over 260 laboratories worldwide. Those users can input data generated from more than 50,000 microarrays used to study the biology of thirty-four organisms, including but not limited to *Homo sapiens*, *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Escherichia coli*. In addition, over a hundred of papers have already published and referred to data in SMD while complete raw data of more than 7,000 microarrays have become freely accessible via the SMD website. In other word, SMD has offered users to upload or store raw and/or normalized data for the microarray experiments. Moreover, SMD also provides functions such as data retrieval, data analysis and visualization interfaces for viewing gene expression patterns (SMD Database Website, 2004).

2.5.2 Gene Expression Omnibus

The Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo) is designed to serve the scientific community a place to share, browse, query and retrieval the high-throughput gene expression / molecular abundance data repository. The datasets have included single and multiple channel microarray-based experiments measuring mRNA, genomic DNA and protein molecules. Serial Analysis of Gene Expression (SAGE) datasets are also accepted by the GEO even though SAGE is not an array-based high-throughput functional genomics and proteomics technology. As to July 2005, GEO has archived 43,010 publicly released samples, including but not limited to organisms of *Homo sapiens, Mus musculus, Drosophila melanogaster and Rattus norvegicus* across 1,446 publicly released platforms, such as the in situ oligonucleotides, spotted oligonucleotides and DNA/cDNA, etc. (GEO Website, 2005).

2.6 Biological Pathways

Biological pathways are defined as having over thousands of enzyme-catalyzed chemical reactions in cells that are functionally organized into many different sequences of consecutive reactions (Nelson et al., 2000). In other words, the product of one reaction would become the reactant in the next.

Every biological reaction has its own unique feature and distinct role that have all worked together to maintain cellular functions in all living organisms. For example: the main functions of catabolic reaction pathways are to degrade organic nutrients into simple products to produce chemical energy and eventually convert this energy for cell use. On the other hand, anabolic reaction pathways would start with small molecules and convert them to relatively larger and more complex molecules, such as proteins (Nelson et al., 2000). The combination of catabolic and anabolic reaction pathways has formed the major metabolic pathways process in all living organisms (Figure 6).

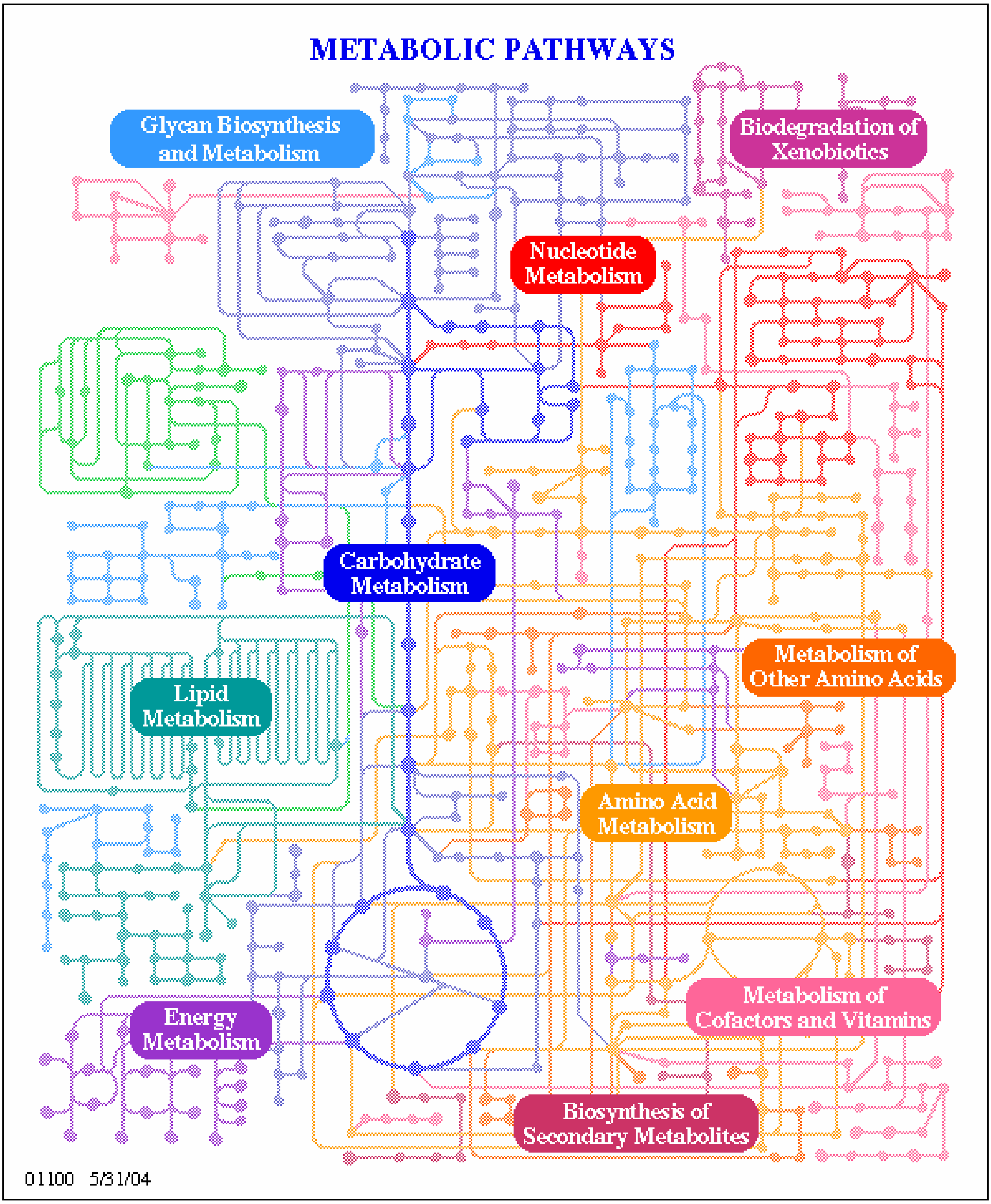

METABOLIC PATHWAYS
Glycan Biosynthesis and Metabolism
Biodegradation of Xenobiotics
Nucleotide Metabolism
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Metabolism of Cofactors and Vitamins
Energy Metabolism
Biosynthesis of Secondary Metabolites
01100 5/31/04

Figure 6: An overview of the metabolic pathways

(*www.genome.ad.jp/kegg/ pathway/map/map01100.html*)

Beside metabolic pathways, regulatory pathways also have significant influences in living lives. The cellular utilization of genetic information is one of the major regulatory pathways known today. The process starts with DNA replication, the copying of double-helix DNA to form daughter DNA molecules with identical nucleotide sequences, followed by transcription, the process where the DNA will be copied into RNA, and ended with translation, whereas the genetic message encoded in messenger RNA is translated into protein (Figure 7).

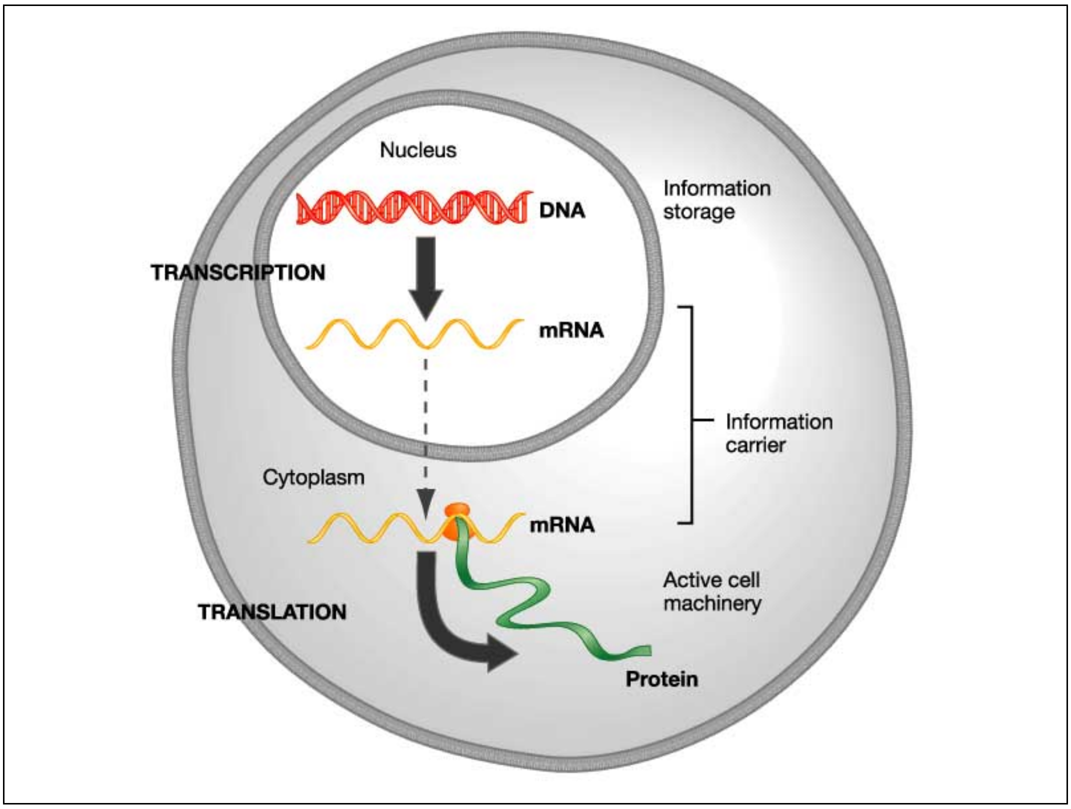


Figure 7: DNA sequences are transcription into RNA sequences in nucleus. The RNA sequences are then moved to the cytoplasm and translated into linear protein chains. (*campus.queens.edu/.../bio103/tests/TEST3Help.htm*).

Another major category of regulatory pathways is the biosignaling process. The ability of cells to receive and act on signals is fundamental to life (Figure 8). If any defective signaling proteins, which are brought along by oncogenes, keep continuing giving the signal for cell division, tumours will be formed as a consequence. Moreover, when abnormal cell development, growth and death occur in the cell regulatory processes, cancer is generally the result of the malfunctions of those fundamental biological processes (SEER's Training Website, 2005).

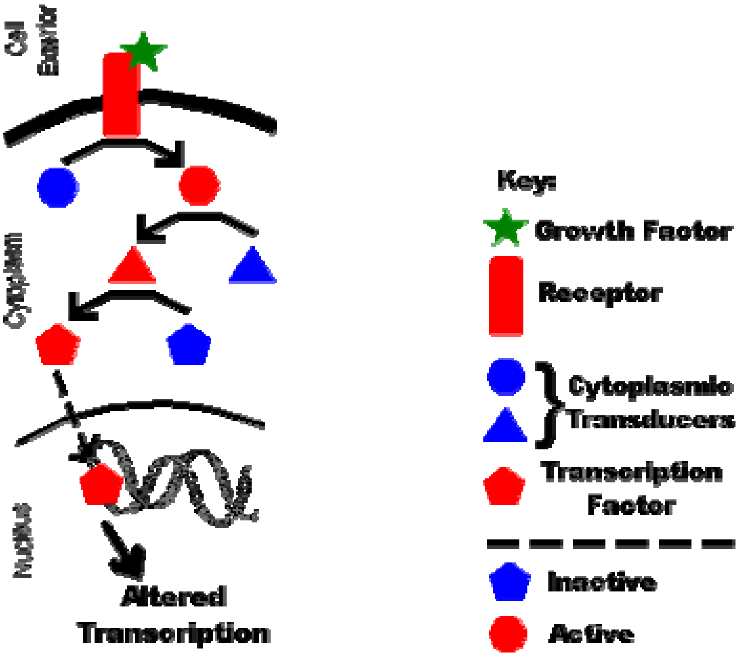


Figure 8: Biosignaling transduction (*http://webhost.bridgew.edu/fgorga/ras/signaling.htm*).

2.7 Gene Ontology

Gene ontology is a set of controlled vocabulary that can explain cell functions and biomedical knowledge of genes or proteins in eukaryotic organisms. Those vocabularies will be updated and changed accordingly as time goes. As to today, biological process, molecular function and cellular component have been developed to represent the three independent sets of vocabularies or ontologies. Molecular function refers to the activities rather than the entities that perform the actual actions at the molecular level. An example of the molecular function can be a hydrolase or enzyme inhibitor activity. Biological process means a biological goal achieved by one or more ordered assemblies of molecular functions, such as cell death. Cellular component describes where the gene product is located at the levels of subcellular structures and macromolecular complexes. An example of the cellular component can be nuclear inner member, or inner envelope (Harris et al., 2004).

III Materials and Methods

3.1 Data Source

Nowadays, most of the experimental microarray data can be obtained from public websites and/or given by the authors upon request. As for the research, we will focus my microarray data source from the Stanford Microarray Database (SMD) and the Gene Expression Omnibus (GEO).

3.2 Classification of Cancer-Related Genes from OMIM Database

OMIM has included a variety of articles, or records, on genetic diseases and inherited genes that have been read and briefly summarized into few sentences by scientists. In other word, OMIM acts as a miniature reading environment for readers to view a variety of the article sources at once. In addition, OMIM is also a high-quality information source and considered a key referencing database by the genetics community. As a result, we have chosen OMIM to be our major resource to derive the cancer-related gene list.

We have limited our search to the key word “cancer” on the OMIM search engine to extract all cancer-related gene records. In order to narrow down the search field to only the text portion, we would only focus on the information contained in the “title”, “text” and “clinical synopsis” (Table 1).

Table 1: Explanation of the name, content and search tips in different search field. (*http://www.ncbi.nlm.nih.gov/Omim/omimhelp.html#SearchFields*)

Search Field	**Description**	**Qualifier**
All Fields	Contains all terms from all searchable database fields in the database.	[ALL]
Allelic Variant	Describes a subset of disease-producing mutations.	[AV] or [VAR]
Chromosome	The chromosome onto which a gene or disorder has been mapped, as reported in the OMIM Gene or Morbid Map.	[CH]or [CHR]
Clinical Synopsis	Clinical features of a disorder and the mode of inheritance (e.g., autosomal dominant, autosomal recessive, x-linked), if known.	[CS] or [CLIN]
Contributor	Contributor to an OMIM record. Names are in the format of *lastname* followed by *one or more initials* (with no periods), e.g., Smith AB	[AU] or [CTRB]
Creation Date	The date on which an OMIM record was created, in the format YYYY/MM/DD.	[CD] or [CDAT]
EC/RN Number	Number assigned by the Enzyme Commission or Chemical Abstract Service (CAS) to designate a particular enzyme or chemical, respectively.	[EC] or [ECNO]
Editor	Editor of OMIM record. Names are in the format of *lastname* followed by *one or more initials* (with no periods), e.g., Smith AB	[ED] or [EDTR]
Filter	Primarily used to retrieve subsets of records that contain crosslinks to other Entrez databases, and LinkOuts to external (non-Entrez) resources. There is a separate LinkOut Overview document which provides more detail about that service.	[FI] or [FILT]
Gene Map	Cytogenetic map location represented in the OMIM Gene Map	[GM]or [MAP]
Gene Map Disorder	Text words appearing in the Disorder column of the OMIM Gene Map.	[DIS] or [DI]
Gene Name	The official gene symbol, and alternate gene symbols, associated with a record. Currently limited to gene symbols present on the OMIM Gene Map. All gene symbols represented in OMIM (mapped or unmapped) can be searched in the Title Word field, described below.	[GN] or [GENE]

MIM Number	For information on the numbering system, see the OMIM FAQs.	[ID] or [MIM]
Modification date	Date on which the record was last modified, in the format YYYY/MM/DD.	[MD] or [MDAT]
Modification History	All dates on which an OMIM record was updated, in the format YYYY/MM/DD.	[MDH] or [HIST]
Properties	An index containing various properties of OMIM records, identifying those which have attributes such as Allelic Variants, Clinical Synopsis, or Gene Map locus. The most commonly used attributes are presented as check boxes on the Limits page. To see a complete list of attributes, you can browse the index of the Properties field by use the Index option.	[PR] or [PROP]
Reference	Contains author names and title words from the articles cited in an OMIM entry. Names are in the format of *lastname* followed by *one or more initials* (with no periods), e.g., Smith AB	[RE] or [REF]
Text Word	Contains terms from the main text-containing section of a record, which begins under the title of a record and ends above the Allelic Variants section (if present), or above the References section (if no Allelic Variants are described).	[TXT] or [WORD]
Title Word	Words in title of an OMIM record. Includes words in the primary title, alternative titles, and included titles.	[TI] or [TITL]

In other word, if the key word “cancer” is nowhere to be found in any of the three sections, we would assume that the record does not consist of any cancer relevant information. Next step would be to review each gene’s OMIM record to confirm its role in different cancer types, which are defined based on the ten leading mortality rate in cancer among Taiwanese population by Department of Health for year 2004. The ten cancer types are lung cancer, hepatocellular carcinoma (HCC), colorectal carcinoma, female breast cancer, gastric carcinoma, oral cancer, cervical cancer, prostate cancer, esophageal cancer and pancreatic cancer.

Further to the key word “cancer” search in the OMIM database, we have also used the ten cancer types for individual search so that a more comprehensive cancer-related gene list would be obtained. Since many different terms can be used to refer to one cancer type, all the possibilities therefore have to be taken into the searching consideration. For example, breast cancer can be described as a breast carcinoma, mammary gland neoplasm etc. As a result, we have used both the synonyms for each cancer type based on the classification by the International Classification of Disease for Oncology (ICD-O) plus the synonyms, near-synonyms and closely related concepts for cancers defined by the Medical Subject Headings (MeSH). ICD-O is used mainly for the cancer and/or tumour registries for coding the histology and site of the neoplasms (ICD-O Website, 2005). Table 2 has shown a summary of the synonyms for ten cancer types defined by ICD-O while Table 3 has illustrated all related terms for the listed cancer types in MeSH.

Table 2: Summary of the Synonyms for Ten Different Cancer Types Defined by ICD-O

Words	**Synonyms by ICD-O**
Cancer	cancer//carcinoma//leukaemia//leukemia//lymphoma//malignancy//melanoma//myeloma//neoplasm//tumor//tumour//
Lung	bronchiole//bronchogenic//bronchus//carina//hilus//lingula//lung//pulmonary//
Liver	liver//hepatocellular//hepatoma//
Colorectal	bowel//cecum//colon//colorectal//ileocecal//intestine//pelvirectal//rectal//rectosigmoid//rectum//sigmoid//
Female Breast	areola//breast//mammary//nipple//
Stomach	antrum//cardia//cardioesophageal//esophagogastric//fundus//gastric//"nos"//prepylorus//pyloric//pylorus//stomach//
Oral	alveolar//alveolus//buccal//cheek//frenulum//gingiva//"gum"//labial//linguae//molar//mouth//oral//palate//periodontal//retromolar//salivary//tongue//tonsil//tooth//uvula//
Cervical	cervical//cervix//endocervical//endocervix//exocervical//exocervix//internal os//nabothian//
Prostate	prostate//prostatic//
Esophageal	esophageal//esophagus//
Pancreatic	langerhans//pancreas//pancreatic//santorini//wirsung//

Table 3: Summary of the Synonyms for Ten Different Cancer Types Defined by MeSH

Cancer Type	**MeSH Headings**	**MeSH Synonyms**
Lung Cancer	Lung Neoplasms	Lung Neoplasms//Cancer of Lung//Lung Cancer//Pulmonary Cancer//Pulmonary Neoplasms//Cancer of the Lung//Neoplasms, Lung//Neoplasms, Pulmonary//Non-Small-Cell Lung Carcinoma//Carcinoma, Non-Small Cell Lung//
Liver Cancer	Liver Neoplasms	Liver Neoplasms//Cancer of Liver//Hepatic Cancer//Liver Cancer//Cancer of the Liver//Hepatic Neoplasms//Neoplasms, Hepatic//Neoplasms, Liver//Carcinoma, Hepatocellular//Hepatocellular Carcinoma//Hepatoma//
Colorectal Cancer	Colorectal Neoplasms	Colonic Neoplasms//Cancer of Colon//Colon Cancer//Cancer of the Colon//Colon Neoplasms//Colonic Cancer//Neoplasms, Colonic//Colorectal Neoplasms, Hereditary Nonpolyposis//Hereditary Nonpolyposis Colorectal Cancer//Hereditary Nonpolyposis Colorectal Neoplasms//Lynch Syndrome//Colon Cancer, Familial Nonpolyposis//Lynch Cancer Family Syndrome I//Lynch Syndrome I//Lynch Syndrome II//
Breast Cancer	Breast Neoplasms	Breast Neoplasms//Breast Cancer//Breast Tumors//Cancer of Breast// Cancer of the Breast//Human Mammary Carcinoma //Mammary Carcinoma, Human//Mammary Neoplasm, Human//Mammary Neoplasms, Human//Neoplasms, Breast//Tumors, Breast/
Stomach Cancer	Stomach Neoplasms	Stomach Neoplasms//Cancer of Stomach//Gastric Cancer//Gastric Neoplasms//Stomach Cancer//Cancer of the Stomach//Neoplasms, Gastric//Neoplasms, Stomach//
Oral Cancer	Mouth Neoplasms	Mouth Neoplasms//Cancer of Mouth//Mouth Cancer//Oral Cancer//Oral Neoplasms//Cancer of the Mouth//Neoplasms, Mouth//Neoplasms, Oral//Oral Cavity//Cavitas Oris//Cavitas oris propria//Mouth Cavity Proper//Oral Cavity Proper//Vestibule Oris//Vestibule of the Mouth//
Cervical Cancer	Cervix Neoplasms	Cervix Neoplasms//Cancer of Cervix//Cervical Cancer//Cancer of the Cervix//Cervical Neoplasms//Cervix Cancer//Neoplasms, Cervical//Neoplasms, Cervix//Cervical Intraepithelial Neoplasia//Neoplasia, Cervical Intraepithelial//Cervical Intraepithelial Neoplasia, Grade III//Cervical Intraepithelial Neoplasms//Intraepithelial Neoplasia, Cervical//
Prostate Cancer	Prostatic Neoplasms	Prostatic Neoplasms//Cancer of Prostate//Prostate Cancer//Cancer of the Prostate//Neoplasms, Prostate//Neoplasms, Prostatic//Prostate Neoplasms//Prostatic Cancer//Prostatic Hyperplasia//Adenoma, Prostatic//Benign Prostatic Hyperplasia//Prostatic Adenoma//Prostatic Hyperplasia, Benign//Prostatic Hypertrophy//Prostatic Hypertrophy, Benign//Prostatism//
Esophageal Cancer	Esophageal Neoplasms	Esophageal Neoplasms//Cancer of Esophagus//Esophageal Cancer//Cancer of the Esophagus//Esophagus Cancer//Esophagus Neoplasm//Neoplasms, Esophageal//
Pancreatic Cancer	Pancreatic Neoplasms	Pancreatic Ductal Carcinoma//Duct-Cell Carcinoma of the Pancreas//Duct-Cell Carcinoma, Pancreas//Ductal Carcinoma of the Pancreas//Pancreatic Duct Cell Carcinoma//Pancreatic Neoplasms//Cancer of Pancreas//Pancreatic Cancer//Cancer of the Pancreas//Neoplasms, Pancreatic//Pancreas Cancer//Pancreas Neoplasms//Carcinoma, Pancreatic Ductal//

Ten gene lists for ten different cancer types would be derived as a result of the reviewing and categorizing process of each OMIM record. We would define each individual gene within each cancer-specific gene list as one cancer-related gene since the gene has been confirmed by the OMIM to be related to this particular cancer type. In other word, we have had a total of ten specific cancer-related gene lists from the OMIM database.

As ten cancer-related gene lists have been identified, we would perform a ten cancer-related gene lists interaction to look for any common genes that are present across ten cancer types. First step would be to use the Microsoft Access software to create tables individually for ten cancer types. Each database table was created by using SQL language. For example, the script for creating the breast cancer table is as follows:

```
Create table breast (genesymbol varchar(15));
```

After creating ten tables, the next step would be to do the cancer-gene list interactions. SQL language is once again used to complete various cancer-gene list interactions. For example, the script can be seen below for the cancer-gene list interactions between the breast, cervical and prostate tissue:

```
SELECT Breast.GeneSymbol
FROM (Breast INNER JOIN Cervical ON
Breast.GeneSymbol = Cervical.GeneSymbol) INNER JOIN
Prostate ON Cervical.GeneSymbol = Prostate.GeneSymbol;
```

The process flow leading to the completion of obtaining the ten cancer-related gene lists as well as the common cancer-related genes are summarized in Figure 9 below.

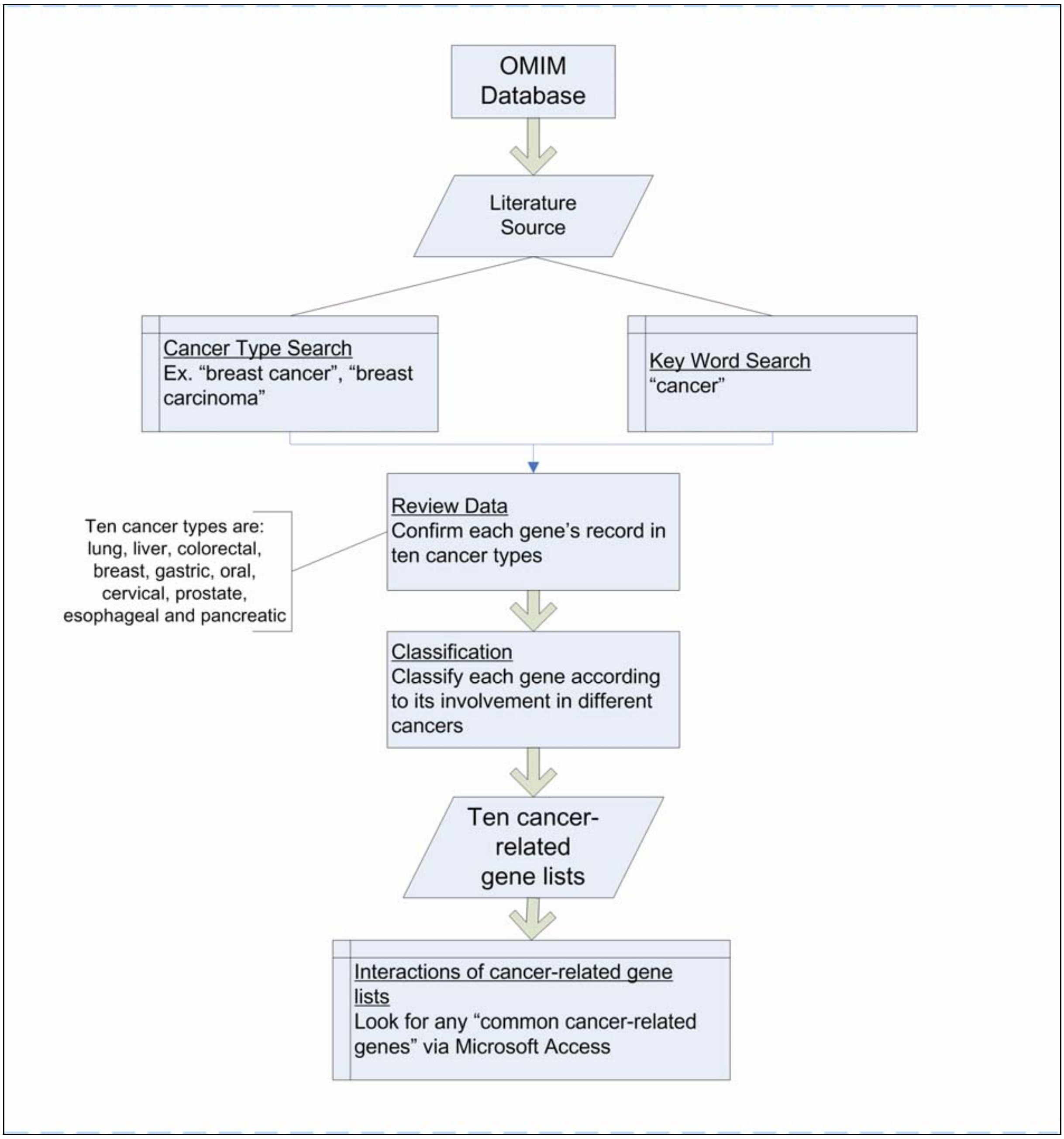

OMIM Database
Literature Source
Cancer Type Search
Ex. "breast cancer", "breast carcinoma"
Key Word Search
"cancer"
Review Data
Confirm each gene's record in ten cancer types
Ten cancer types are: lung, liver, colorectal, breast, gastric, oral, cervical, prostate, esophageal and pancreatic
Classification
Classify each gene according to its involvement in different cancers
Ten cancer-related gene lists
Interactions of cancer-related gene lists
Look for any "common cancer-related genes" via Microsoft Access

Figure 9: Flowchart for extracting cancer-related gene lists from the OMIM database.

3.3 Microarray Data Extraction from GEO Database

Many researchers have deposited their precious microarray datasets onto GEO; thus, we can download the relevant cancer datasets from its website. We used the GEO incorporated function “Browse->DataSets” to list out all available datasets within GEO (Figure 10).

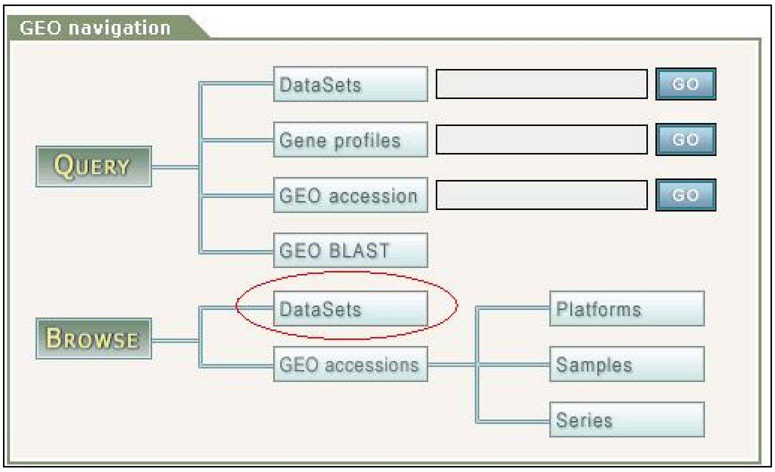
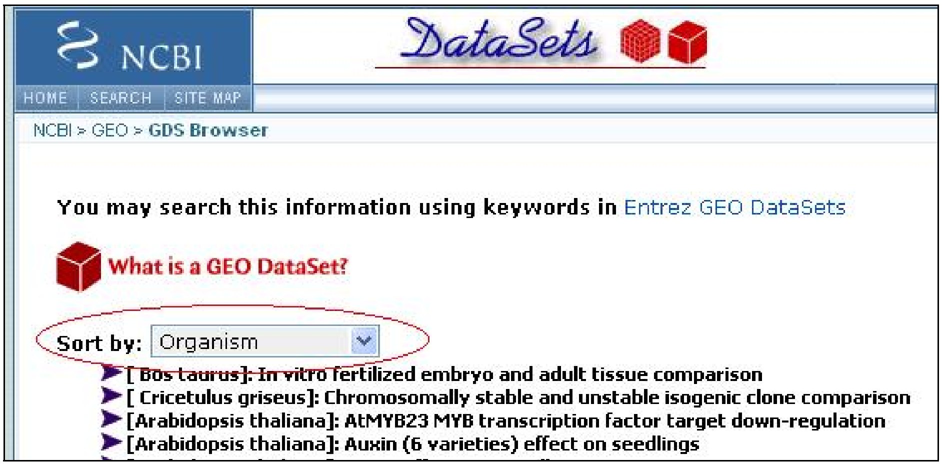


Figure 10: Browse all the available datasets on GEO (GEO Website, 2005).

Since we are only interested in human datasets, we would first use the function “Sort by Organisms” to have all human datasets listed together (Figure 11).



NCBI > GEO > **GDS Browser**

You may search this information using keywords in Entrez GEO DataSets

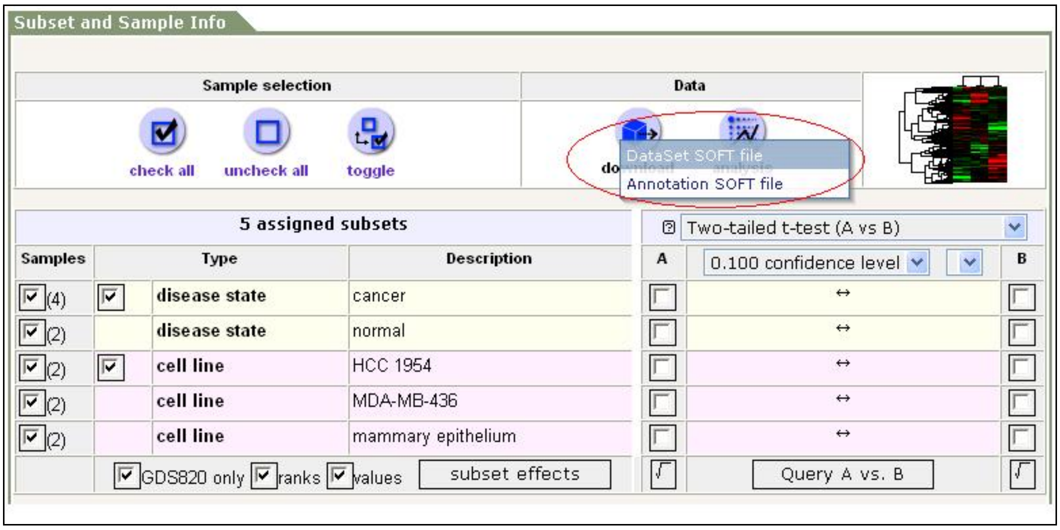
What is a GEO DataSet?

Sort by: Organism

- [Bos taurus]: In vitro fertilized embryo and adult tissue comparison
- [Cricetulus griseus]: Chromosomally stable and unstable isogenic clone comparison
- [Arabidopsis thaliana]: AtMYB23 MYB transcription factor target down-regulation
- [Arabidopsis thaliana]: Auxin (6 varieties) effect on seedlings

Figure 11: Sort out all GEO datasets according to organisms (GEO Website, 2005).

Next step would be to focus on the datasets using oligonucleotide chips and remove those that were uploaded by authors who have published their data on SMD database already. In addition, we were most interested in dataset record that consisted of either normal or cancer tissue samples, or both at the same time but no cell line samples (Figure 12).



5 assigned subsets				Two-tailed t-test (A vs B)		
Samples		Type	Description	A	0.100 confidence level	B
☑(4)	☑	disease state	cancer	□	↔	□
☑(2)		disease state	normal	□	↔	□
☑(2)	☑	cell line	HCC 1954	□	↔	□
☑(2)		cell line	MDA-MB-436	□	↔	□
☑(2)		cell line	mammary epithelium	□	↔	□
	☑GDS820 only ☑ranks ☑values		subset effects	√	Query A vs. B	√

Figure 12: GEO dataset download and information page (GEO Website, 2005).

As a result of the pre-filtering process, three published datasets for each breast cancer and gastric cancer, one for each prostate cancer, colorectal cancer, cervical cancer and hepatocellular carcinoma, and two for lung cancer were left for thorough reviewing. Moreover, the Su *et. al*'s two published datasets in which one consists of various types of human normal tissues, including lung, prostate, liver, etc., and the other includes a variety of tumour samples would also include into our reviewing process (Su et al., 2001; Su et al., 2002). The final datasets for analysis use were the ones from the breast, prostate, liver, lung and pancreas tissues. The details for the datasets we used for analysis are summarized in Table 4 below.

Table 4: Summary of cancer-related literatures used for analysis from the GEO database

Cancer Type	Author	Paper Title	Array Type	Experiment Description	Used for Our Analysis
Breast	Mecham BH, et al.	Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements	HG-U95A	2 normal and 4 cancer state breast tissues	Yes
			HG-U133A	2 normal and 4 cancer state breast tissues	Yes
			HG-U133B	2 normal and 4 cancer state breast tissues	Yes
			cDNA (G4100A)	2 cell line sets	No
N/A	Su et al.	Large-scale analysis of the human and mouse transcriptomes	HG-U95A	2 breast normal, 2 prostate normal, 2 liver normal, 2 lung normal and 2 pancreatic normal	Yes
N/A	Su et al.	Molecular classification of human carcinomas by use of gene expression signature	HG-U95A	31 prostate cancer samples, 7 liver cancer samples, 28 lungcancer samples, 6 pancreatic cancer samples, 23 breast cancer samples	Yes

Upon obtaining the five specific cancer-related gene lists, we would again perform cancer-related gene lists intersection to look for any common genes that are present across five cancer types via Microsoft Access. The detailed steps on how we achieved the gene lists intersections have been mentioned in Section 3.2 previously. The process flow from collecting the microarray datasets from GEO database to receive the common cancer-related gene lists are summarized in Figure 13 below.

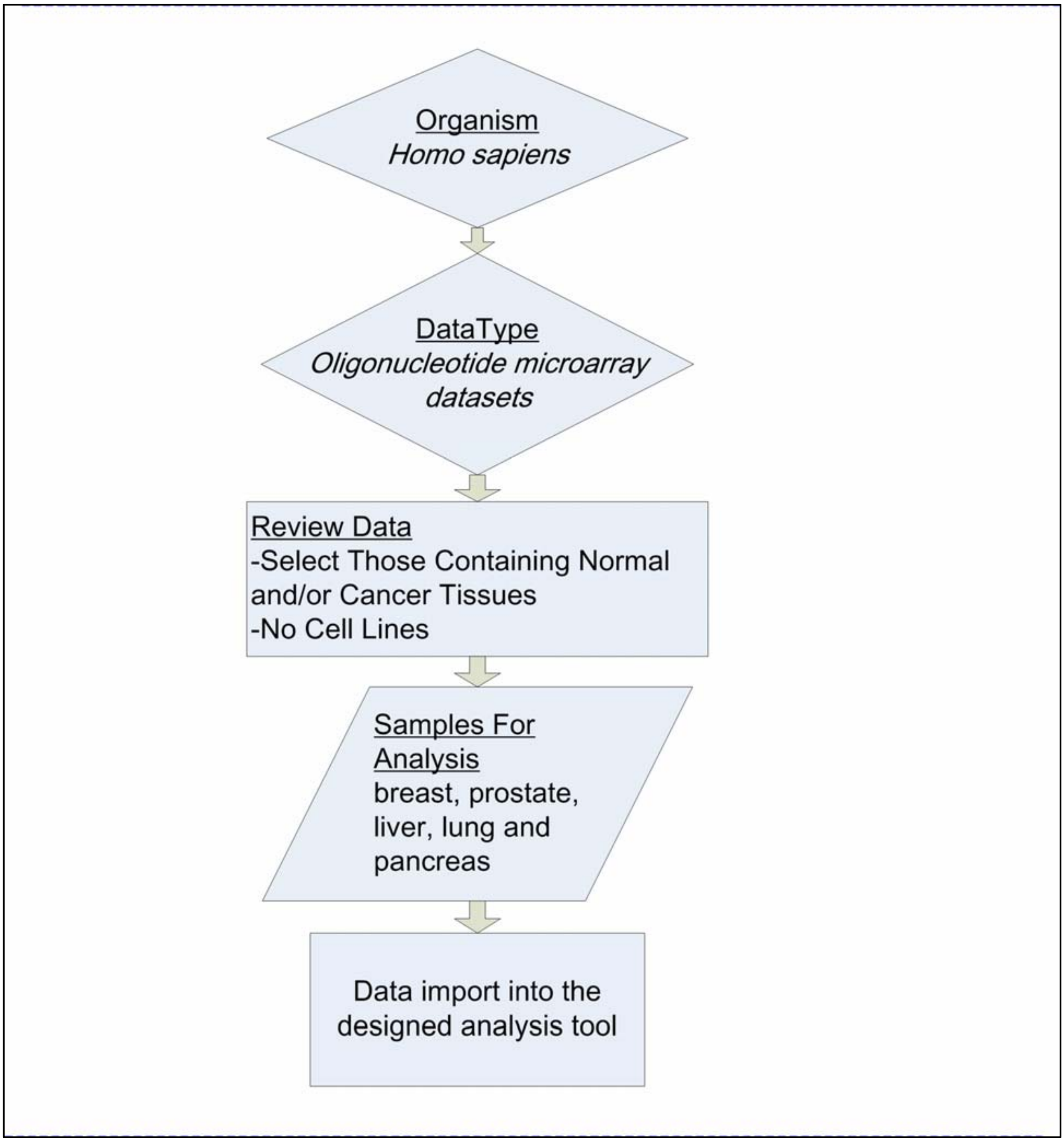
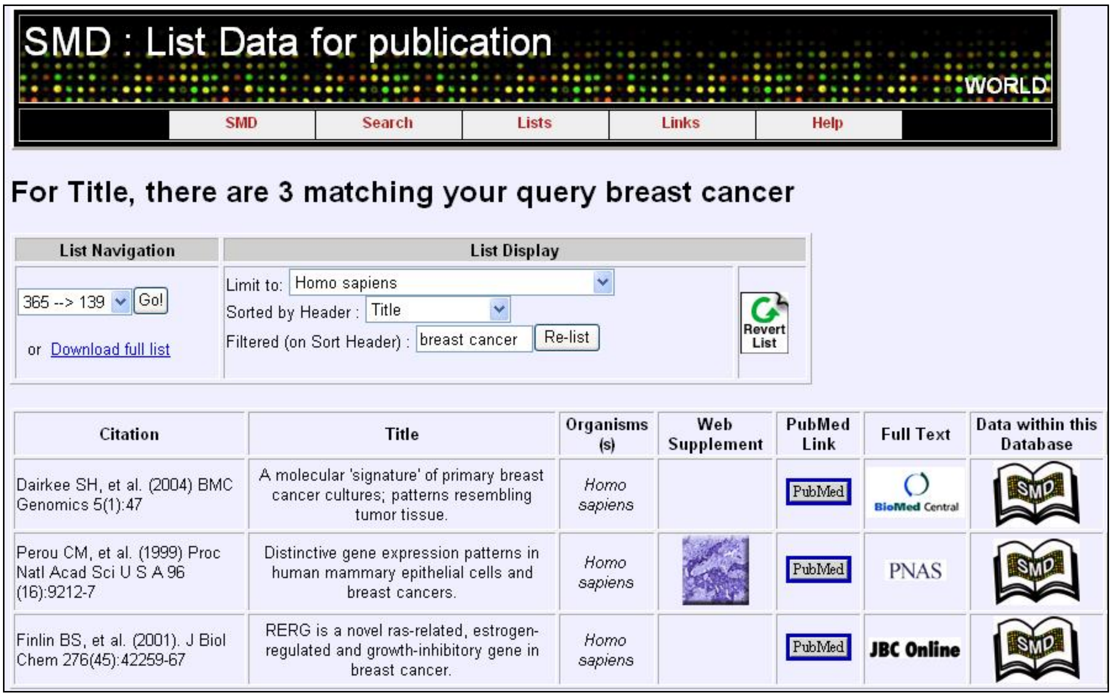

Organism
Homo sapiens
DataType
Oligonucleotide microarray datasets
Review Data
-Select Those Containing Normal and/or Cancer Tissues
-No Cell Lines
Samples For Analysis
breast, prostate, liver, lung and pancreas
Data import into the designed analysis tool

Figure 13: Process flowchart for extracting the datasets from GEO database.

3.4 Microarray Data Extraction from SMD Database

SMD is another famous microarray database that consists mainly of cDNA microarray datasets. Each SMD publication includes considerable amount of sample numbers which in terms increase the reliability of the experiment. We have used the search engine developed by Stanford University to query all the cancer-related publications (Figure 14).



For Title, there are 3 matching your query breast cancer

List Navigation

365 --> 139 Go!

or Download full list

List Display

Limit to: Homo sapiens

Sorted by Header : Title

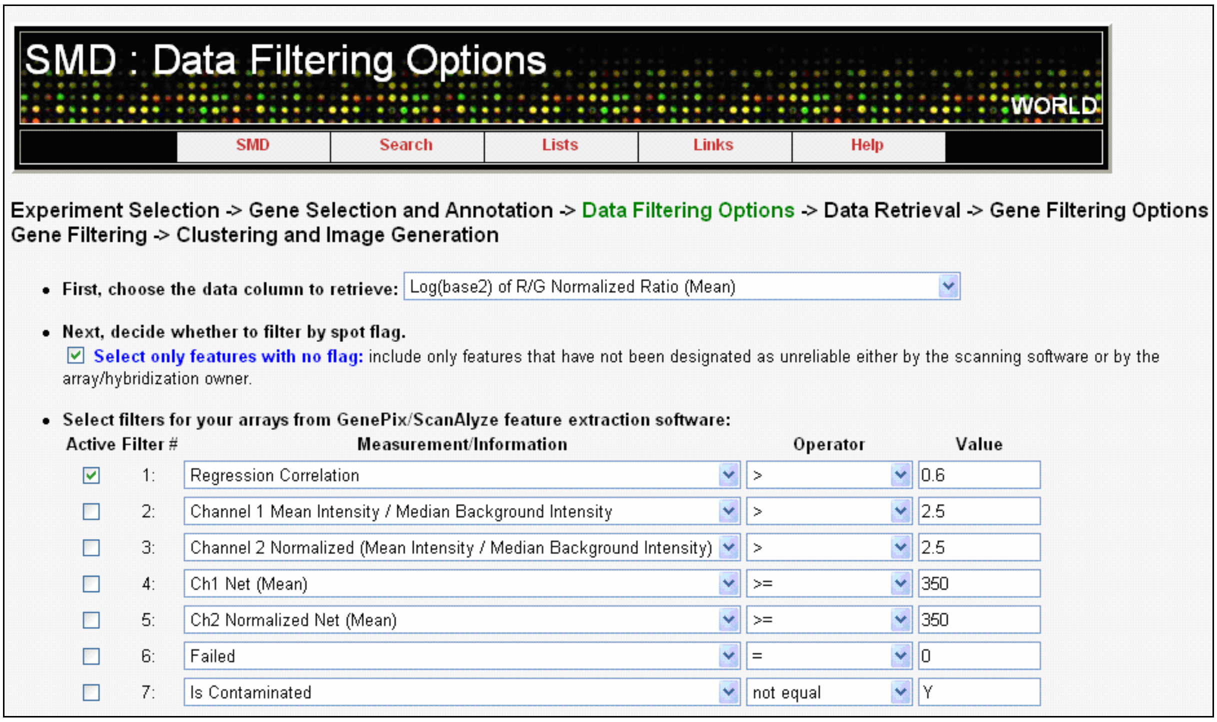
Filtered (on Sort Header) : breast cancer Re-list

Revert List

Citation	Title	Organisms (s)	Web Supplement	PubMed Link	Full Text	Data within this Database
Dairkee SH, et al. (2004) BMC Genomics 5(1):47	A molecular 'signature' of primary breast cancer cultures; patterns resembling tumor tissue.	*Homo sapiens*		PubMed	BioMed Central	SMD
Perou CM, et al. (1999) Proc Natl Acad Sci U S A 96 (16):9212-7	Distinctive gene expression patterns in human mammary epithelial cells and breast cancers.	*Homo sapiens*		PubMed	PNAS	SMD
Finlin BS, et al. (2001). J Biol Chem 276(45):42259-67	RERG is a novel ras-related, estrogen-regulated and growth-inhibitory gene in breast cancer.	*Homo sapiens*		PubMed	JBC Online	SMD

Figure 14: Query publications related to breast cancer using the search engine (SMD Database Website, 2004).

Next step was to filter out only those publications that matched to our ten cancer types and to download those microarray data from the SMD website. In total, we have got twelve experimental datasets for breast cancer, three for gastric cancer, one for hepatocellular carcinoma, six for prostate cancer and two for each lung cancer and pancreatic cancer. Since we want to filer out datasets containing cell line samples and include only those with both normal and cancer tissue samples at the same time, only one dataset for each of the breast cancer, lung cancer, gastric cancer, prostate cancer, pancreatic cancer and hepatocellular carcinoma would be analyzed (Table 6). We then used the SMD incorporated data analysis and retrieval system to pull out required information, such as the gene symbol, log base 2 of R/G normalized ratio of the mean, etc. (Figure 15).



Experiment Selection -> Gene Selection and Annotation -> Data Filtering Options -> Data Retrieval -> Gene Filtering Options Gene Filtering -> Clustering and Image Generation

- First, choose the data column to retrieve: Log(base2) of R/G Normalized Ratio (Mean)
- Next, decide whether to filter by spot flag.
  ☑ Select only features with no flag: include only features that have not been designated as unreliable either by the scanning software or by the array/hybridization owner.
- Select filters for your arrays from GenePix/ScanAlyze feature extraction software:

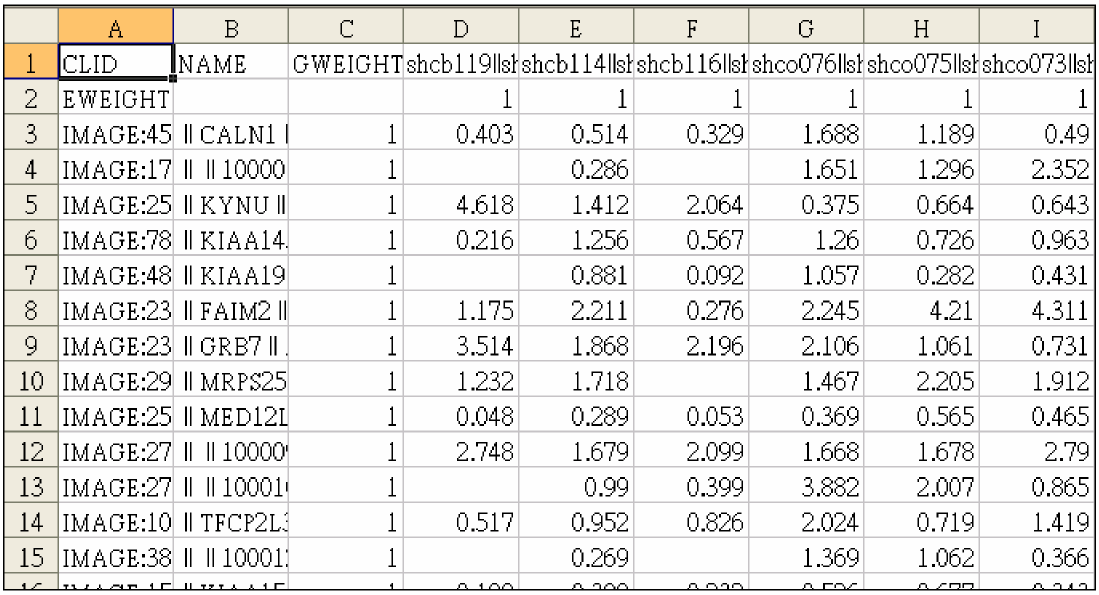
Active Filter #		Measurement/Information	Operator	Value
☑	1:	Regression Correlation	>	0.6
□	2:	Channel 1 Mean Intensity / Median Background Intensity	>	2.5
□	3:	Channel 2 Normalized (Mean Intensity / Median Background Intensity)	>	2.5
□	4:	Ch1 Net (Mean)	>=	350
□	5:	Ch2 Normalized Net (Mean)	>=	350
□	6:	Failed	=	0
□	7:	Is Contaminated	not equal	Y

Figure 15: Using SMD-incorporated data analysis and retrieval system to extract required information, such as gene symbol, log base 2 of R/G normalized ratio of the mean, etc. (SMD Database Website, 2004)..

Table 6: Summary of cancer-related literatures used for the analysis from SMD database

Cancer Type	Author	Paper Title	Array Type	Experiment Description	Used for Our Analysis
Breast	Sorlie T, et al.	Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.	cDNA	4 normal breast tissues and 81 primary tumour	Yes
Lung	Garber ME, et al.	Diversity of gene expression in adenocarcinoma of the lung	cDNA	6 normal tissue, 61 primary lung tumour	Yes
Gastric	Chen X, et al.	Variation in gene expression patterns in human gastric cancers	cDNA	103 gastric cancer tissues and 29 non-neoplastic gastric tissues	Yes
Gastric	Leung SY, et al.	Expression profiling identifies chemokine (C-C motif) ligand 18 as an independent prognostic indicator in gastric cancer	cDNA	23 non-tumour tissue and 103 primary tumour (Part of Chen X Data)	Yes
Gastric	Leung SY, et al.	Phospholipase A2 group IIA expression in gastric adenocarcinoma is associated with prolonged survival and less frequent metastasis.	cDNA	23 non-tumour tissue and 103 primary tumour (Part of Chen X Data)	Yes
Prostate	Lapointe J, et al.	Gene expression profiling identifies clinically relevant subtypes of prostate cancer	cDNA	62 primary prostate tumors, 41 normal prostate specimens and nine lymph node metastases	Yes
Pancreatic	Iacobuzio-Donahue CA, et al.	Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays	cDNA	17 infiltrating pancreatic cancer tissues, and 5 samples of normal pancreas	Yes
Liver	Chen X, et al.	Gene expression patterns in human liver cancers	cDNA	102 primary HCC tumour tissues, 74 non-tumour liver tissues, 10 metastatic cancers, 3 adenoma tumour samples and 4 FNH tumour samples	Yes

The downloaded file is in text format containing much information about this publication's datasets. An example of the download dataset format is shown in Figure 16. Column A contains the information for clone ID, column B indicates each gene's name; column D and onwards include each experimental slide's log base 2 of R/G normalized ratio of the mean. Before doing any further analysis, we would have to categorize each experimental slide manually into cancer and normal groups.



<table>
<tr><th></th><th>A</th><th>B</th><th>C</th><th>D</th><th>E</th><th>F</th><th>G</th><th>H</th><th>I</th></tr>
<tr><td>1</td><td>CLID</td><td>NAME</td><td>GWEIGHT</td><td>shcb119||sh</td><td>shcb114||sh</td><td>shcb116||sh</td><td>shco076||sh</td><td>shco075||sh</td><td>shco073||sh</td></tr>
<tr><td>2</td><td>EWEIGHT</td><td></td><td></td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td></tr>
<tr><td>3</td><td>IMAGE:45</td><td>|| CALN1 |</td><td>1</td><td>0.403</td><td>0.514</td><td>0.329</td><td>1.688</td><td>1.189</td><td>0.49</td></tr>
<tr><td>4</td><td>IMAGE:17</td><td>|| || 10000</td><td>1</td><td></td><td>0.286</td><td></td><td>1.651</td><td>1.296</td><td>2.352</td></tr>
<tr><td>5</td><td>IMAGE:25</td><td>|| KYNU ||</td><td>1</td><td>4.618</td><td>1.412</td><td>2.064</td><td>0.375</td><td>0.664</td><td>0.643</td></tr>
<tr><td>6</td><td>IMAGE:78</td><td>|| KIAA14</td><td>1</td><td>0.216</td><td>1.256</td><td>0.567</td><td>1.26</td><td>0.726</td><td>0.963</td></tr>
<tr><td>7</td><td>IMAGE:48</td><td>|| KIAA19</td><td>1</td><td></td><td>0.881</td><td>0.092</td><td>1.057</td><td>0.282</td><td>0.431</td></tr>
<tr><td>8</td><td>IMAGE:23</td><td>|| FAIM2 ||</td><td>1</td><td>1.175</td><td>2.211</td><td>0.276</td><td>2.245</td><td>4.21</td><td>4.311</td></tr>
<tr><td>9</td><td>IMAGE:23</td><td>|| GRB7 ||</td><td>1</td><td>3.514</td><td>1.868</td><td>2.196</td><td>2.106</td><td>1.061</td><td>0.731</td></tr>
<tr><td>10</td><td>IMAGE:29</td><td>|| MRPS25</td><td>1</td><td>1.232</td><td>1.718</td><td></td><td>1.467</td><td>2.205</td><td>1.912</td></tr>
<tr><td>11</td><td>IMAGE:25</td><td>|| MED12L</td><td>1</td><td>0.048</td><td>0.289</td><td>0.053</td><td>0.369</td><td>0.565</td><td>0.465</td></tr>
<tr><td>12</td><td>IMAGE:27</td><td>|| || 10000</td><td>1</td><td>2.748</td><td>1.679</td><td>2.099</td><td>1.668</td><td>1.678</td><td>2.79</td></tr>
<tr><td>13</td><td>IMAGE:27</td><td>|| || 10001</td><td>1</td><td></td><td>0.99</td><td>0.399</td><td>3.882</td><td>2.007</td><td>0.865</td></tr>
<tr><td>14</td><td>IMAGE:10</td><td>|| TFCP2L</td><td>1</td><td>0.517</td><td>0.952</td><td>0.826</td><td>2.024</td><td>0.719</td><td>1.419</td></tr>
<tr><td>15</td><td>IMAGE:38</td><td>|| || 10001</td><td>1</td><td></td><td>0.269</td><td></td><td>1.369</td><td>1.062</td><td>0.366</td></tr>
</table>

Figure 16: Snapshot of the downloaded dataset format.

As we had the six specific cancer-related gene lists on hand, we again performed cancer-related gene lists intersection to look for any common genes that are present across six cancer types via Microsoft Access. The detailed steps on how we achieved the gene lists interactions have been mentioned in Section 3.2 previously. An overall process flow is shown in Figure 17 to illustrate what we did step by step towards extracting the required data from the SMD database.

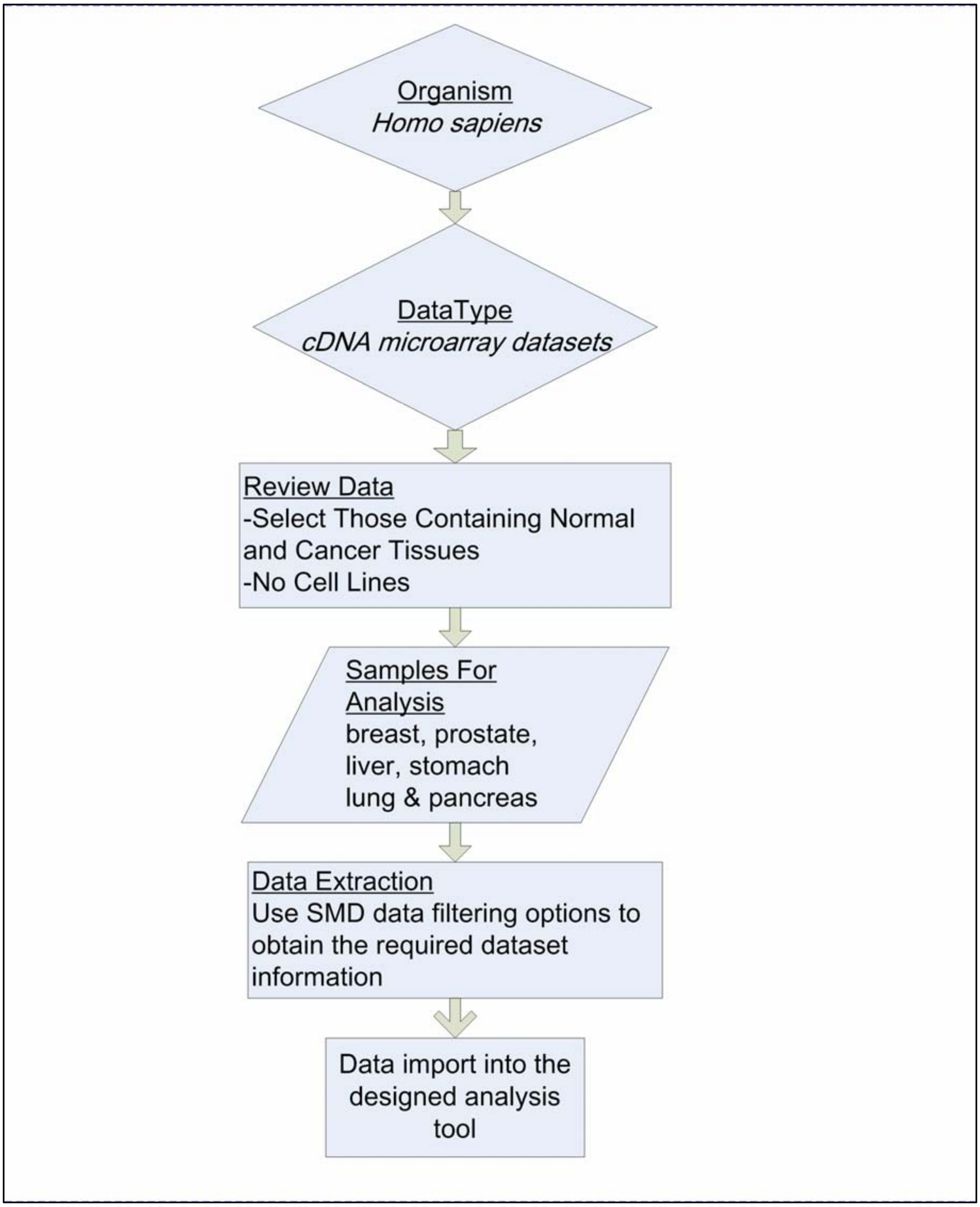
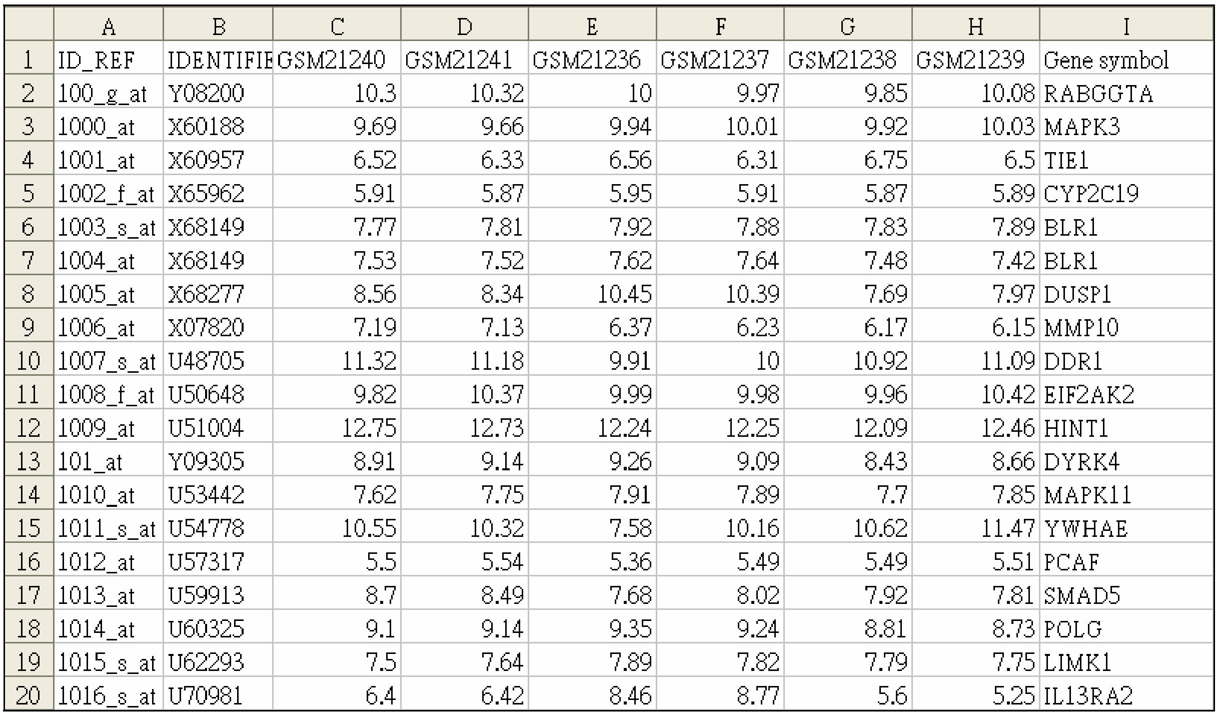

Organism
Homo sapiens
DataType
cDNA microarray datasets
Review Data
-Select Those Containing Normal and Cancer Tissues
-No Cell Lines
Samples For Analysis
breast, prostate, liver, stomach lung & pancreas
Data Extraction
Use SMD data filtering options to obtain the required dataset information
Data import into the designed analysis tool

Figure 17: Process flowchart to extract datasets from SMD database.

3.5 Datasets Analysis

3.5.1 Downloaded Datasets Processing

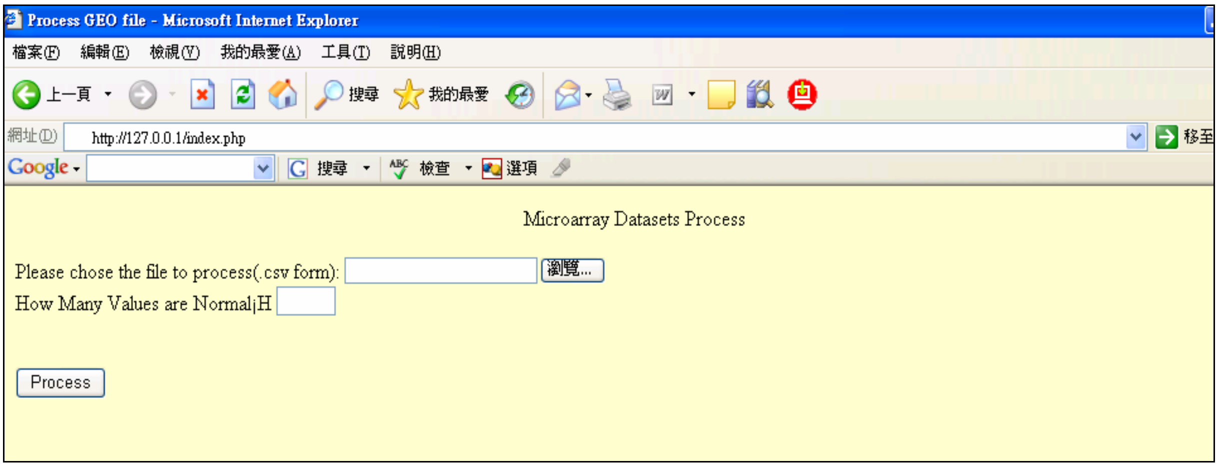
The downloaded datasets would be analyzed via a designed tool to extract the cancer-related gene lists. The dataset format has to be organized into the format as shown in Figure 18 before having imported into the designed tool. For example, column A has to be the microarray unique ID, in which the oligonucleotide array data from Affymetrix would use the probe set ID while the cDNA data from SMD would use the clone image ID. Column B has to be the GenBank accession numbers then followed by a series of experimental slides' data points. Lastly, the final column has to insert the gene symbol.



	A	B	C	D	E	F	G	H	I
1	ID_REF	IDENTIFIE	GSM21240	GSM21241	GSM21236	GSM21237	GSM21238	GSM21239	Gene symbol
2	100_g_at	Y08200	10.3	10.32	10	9.97	9.85	10.08	RABGGTA
3	1000_at	X60188	9.69	9.66	9.94	10.01	9.92	10.03	MAPK3
4	1001_at	X60957	6.52	6.33	6.56	6.31	6.75	6.5	TIE1
5	1002_f_at	X65962	5.91	5.87	5.95	5.91	5.87	5.89	CYP2C19
6	1003_s_at	X68149	7.77	7.81	7.92	7.88	7.83	7.89	BLR1
7	1004_at	X68149	7.53	7.52	7.62	7.64	7.48	7.42	BLR1
8	1005_at	X68277	8.56	8.34	10.45	10.39	7.69	7.97	DUSP1
9	1006_at	X07820	7.19	7.13	6.37	6.23	6.17	6.15	MMP10
10	1007_s_at	U48705	11.32	11.18	9.91	10	10.92	11.09	DDR1
11	1008_f_at	U50648	9.82	10.37	9.99	9.98	9.96	10.42	EIF2AK2
12	1009_at	U51004	12.75	12.73	12.24	12.25	12.09	12.46	HINT1
13	101_at	Y09305	8.91	9.14	9.26	9.09	8.43	8.66	DYRK4
14	1010_at	U53442	7.62	7.75	7.91	7.89	7.7	7.85	MAPK11
15	1011_s_at	U54778	10.55	10.32	7.58	10.16	10.62	11.47	YWHAE
16	1012_at	U57317	5.5	5.54	5.36	5.49	5.49	5.51	PCAF
17	1013_at	U59913	8.7	8.49	7.68	8.02	7.92	7.81	SMAD5
18	1014_at	U60325	9.1	9.14	9.35	9.24	8.81	8.73	POLG
19	1015_s_at	U62293	7.5	7.64	7.89	7.82	7.79	7.75	LIMK1
20	1016_s_at	U70981	6.4	6.42	8.46	8.77	5.6	5.25	IL13RA2

Figure 18: Dataset format for the analysis

Next step would be to import the data into our tool for further analysis. The tool interface is shown below in Figure 19. Before having submitted the data for analysis, we had to indicate how many data columns contained normal values. In other word, we had to insert the normal experimental slides' data right after Column B.


Process GEO file - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
網址(D)
移至
Google
搜尋
檢查
選項

http://127.0.0.1/index.php

Microarray Datasets Process

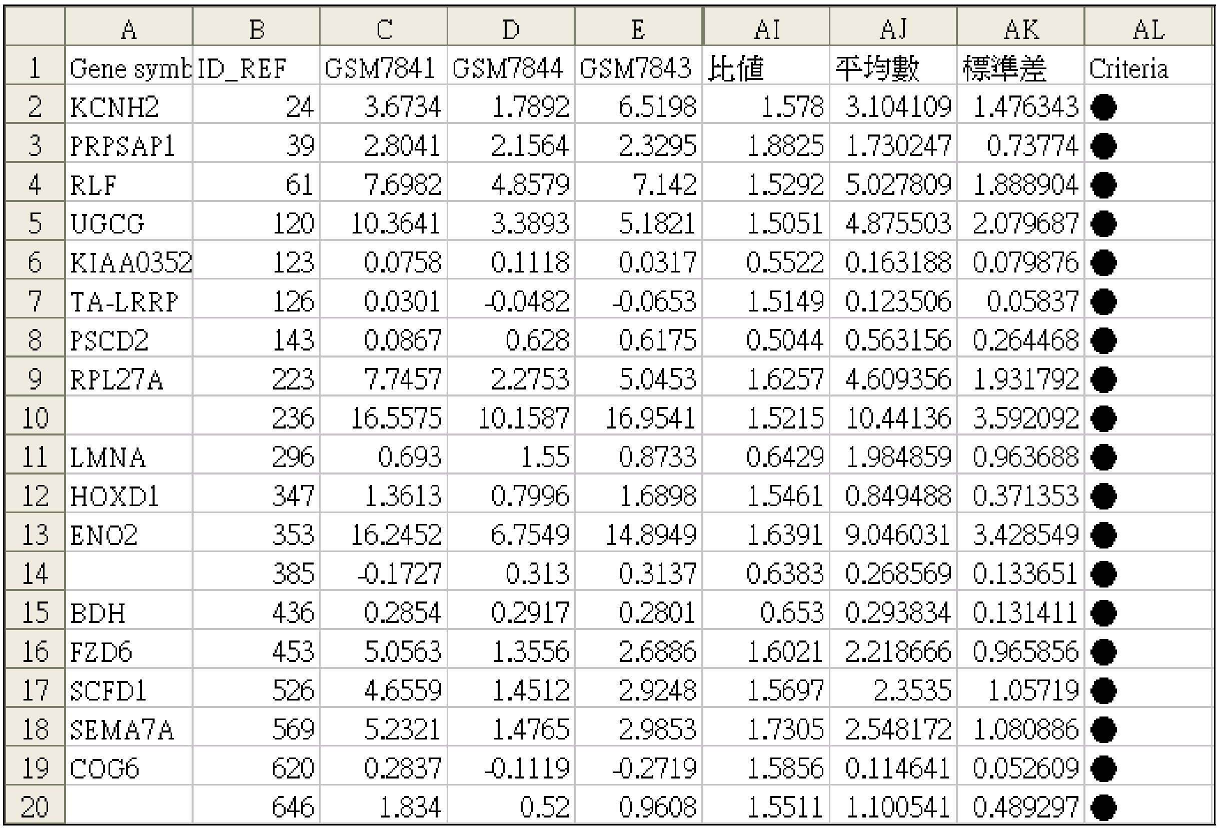
Please chose the file to process(.csv form): 瀏覽...

How Many Values are Normal¡H

Process

Figure 19: Interface of the analytical tool.

Throughout the analytical process, the first step would be to eliminate those genes that have more than half of the expression values are missing. Then the tool would help us to calculate each gene's expression ratio between normal and abnormal tissues, the mean expression level as well as the standard deviation of the gene. Since we would only want to extract genes that have the ratio greater than 1.5 or less than 2/3 fold, we would use a solid circle to indicate genes that did meet the set criteria. Figure 20 has shown an example of the result format after the analysis is complete. The last four columns have specified the result of our analysis. One column would contain the calculated ratio of the cancer and normal samples, followed by a column indicating the average expression levels. One more column would have the standard deviation values and the last column would point out which gene has fulfilled our criteria to be included into the cancer-related gene list.



	A	B	C	D	E	AI	AJ	AK	AL
1	Gene symb	ID_REF	GSM7841	GSM7844	GSM7843	比值	平均數	標準差	Criteria
2	KCNH2	24	3.6734	1.7892	6.5198	1.578	3.104109	1.476343	●
3	PRPSAP1	39	2.8041	2.1564	2.3295	1.8825	1.730247	0.73774	●
4	RLF	61	7.6982	4.8579	7.142	1.5292	5.027809	1.888904	●
5	UGCG	120	10.3641	3.3893	5.1821	1.5051	4.875503	2.079687	●
6	KIAA0352	123	0.0758	0.1118	0.0317	0.5522	0.163188	0.079876	●
7	TA-LRRP	126	0.0301	-0.0482	-0.0653	1.5149	0.123506	0.05837	●
8	PSCD2	143	0.0867	0.628	0.6175	0.5044	0.563156	0.264468	●
9	RPL27A	223	7.7457	2.2753	5.0453	1.6257	4.609356	1.931792	●
10		236	16.5575	10.1587	16.9541	1.5215	10.44136	3.592092	●
11	LMNA	296	0.693	1.55	0.8733	0.6429	1.984859	0.963688	●
12	HOXD1	347	1.3613	0.7996	1.6898	1.5461	0.849488	0.371353	●
13	ENO2	353	16.2452	6.7549	14.8949	1.6391	9.046031	3.428549	●
14		385	-0.1727	0.313	0.3137	0.6383	0.268569	0.133651	●
15	BDH	436	0.2854	0.2917	0.2801	0.653	0.293834	0.131411	●
16	FZD6	453	5.0563	1.3556	2.6886	1.6021	2.218666	0.965856	●
17	SCFD1	526	4.6559	1.4512	2.9248	1.5697	2.3535	1.05719	●
18	SEMA7A	569	5.2321	1.4765	2.9853	1.7305	2.548172	1.080886	●
19	COG6	620	0.2837	-0.1119	-0.2719	1.5856	0.114641	0.052609	●
20		646	1.834	0.52	0.9608	1.5511	1.100541	0.489297	●

Figure 20: Microarray expression datasets analysis result.

3.5.2 Search for Cancer-Related Gene Involvement in Biological Pathway

Upon obtaining the common cancer-related genes after the gene list interactions, we would determine those genes' functions and their locations in each biological pathway using KEGG database (http://www.genome.ad.jp/kegg-bin/mk_point_html). KEGG is short for Kyoto Encyclopedia of Genes and Genomes, a publicly available pathway database containing updated knowledge on molecular interaction networks, which includes metabolic pathways, regulatory pathways and molecular complexes (KEGG Website, 2005). The KEGG has provided a huge collection of biological pathways diagrams that can clearly view gene-to-pathway relationships. In other words, each gene's specific location in each biological pathway can easily be seen on the diagram.

IV. Result

4.1 Cancer-Related Genes from OMIM

Upon reviewing the OMIM queried gene list data for different cancer types, we have obtained ten individual specific cancer-related gene lists for our designated ten cancer types. A summary of number of genes that have associated with each cancer type is briefed in Table 7.

Table 7: Number of cancer-related genes in relation to each specific cancer type

Cancer Type	Breast	Cervical	Colon	Esophageal	Liver	Lung	Pancreas	Oral	Prostate	Stomach
Number of cancer-related genes	388	106	287	136	610	491	193	47	236	133

Ten specific cancer-related gene lists then were imported into KEGG pathway database to obtain pathway lists. Cancer-related genes in relation to biological pathways have been summarized in Table 8. Only twelve biological pathways in which cancer-related genes present across all ten cancer types are extracted. Moreover, although there are quite some numbers of specific cancer-related genes mentioned in OMIM literatures, only APC, CDKN2A and PTEN genes are found to be present across ten different cancers. Gene APC is mainly involved in the environmental information processing – wnt signaling pathway and cellular processes regulation of actin cytoskeleton. Gene CDKN2A can be found to have a role in cellular process cell cycle. Gene PTEN takes parts in both the environmental information processing phosphatidylinositol signaling system and inositol phosphate metabolism.

Table 8: Summary of OMIM defined cancer-related genes in relation to the biological pathways. Each number represents how many genes have associated with each pathway individually.

Pathway Name	**Breast**	**Cervical**	**Colon**	**Esophageal**	**Liver**	**Lung**	**Oral**	**Pancreas**	**Prostate**	**Stomach**	**Total**
Environmental Information Processing MAPK signaling pathway	22	8	16	14	26	18	3	13	8	17	145
Environmental Information Processing Wnt signaling pathway	14	5	18	4	20	23	2	10	11	9	116
Cellular Processes Regulation of actin cytoskeleton	24	8	15	7	16	15	3	6	4	11	109
Environmental Information Processing Cytokine-cytokine receptor interaction	8	2	9	8	27	25	1	4	5	10	99
Cellular Process Cell Cycle	7	4	13	4	10	14	1	8	7	3	71
Environmental Information Processing Neuroactive ligand-receptor interaction	7	3	5	2	13	17	2	6	10	3	68
Cellular Processes Adherens junction	6	3	12	6	10	9	2	3	2	6	59
Cellular Process Apoptosis	4	2	8	6	12	6	1	4	3	8	54
Environmental Information Processing TGF-beta signaling pathway	7	1	4	3	8	10	1	4	3	3	44
Cellular Processes Focal adhesion	5	3	5	2	5	4	2	4	3	5	38
Environmental Information Processing Phosphatidylinositol signaling system	11	2	4	1	5	4	1	3	2	3	36
Metabolism Inositol phosphate metabolism	9	2	3	1	5	3	1	3	1	2	30
Total Genes	124/388	43/106	112/287	58/136	157/610	148/491	20/47	68/193	59/236	80/133	

Furthermore, when we classified those cancer tissue types into different groups, we would get some quite intriguing results. When having intersected the cancer-related gene lists obtained from stomach, pancreatic, liver and colon tissues altogether, a total of twenty-three genes are considered as present in all four tissues (Figure 21; Appendix I). As having queried the associated pathways for those twenty-three genes on the KEGG system, we have found that fourteen of the twenty-three genes do not have any involvements in any of the pathways. The remaining nine genes have shown there are no common pathways present among them.

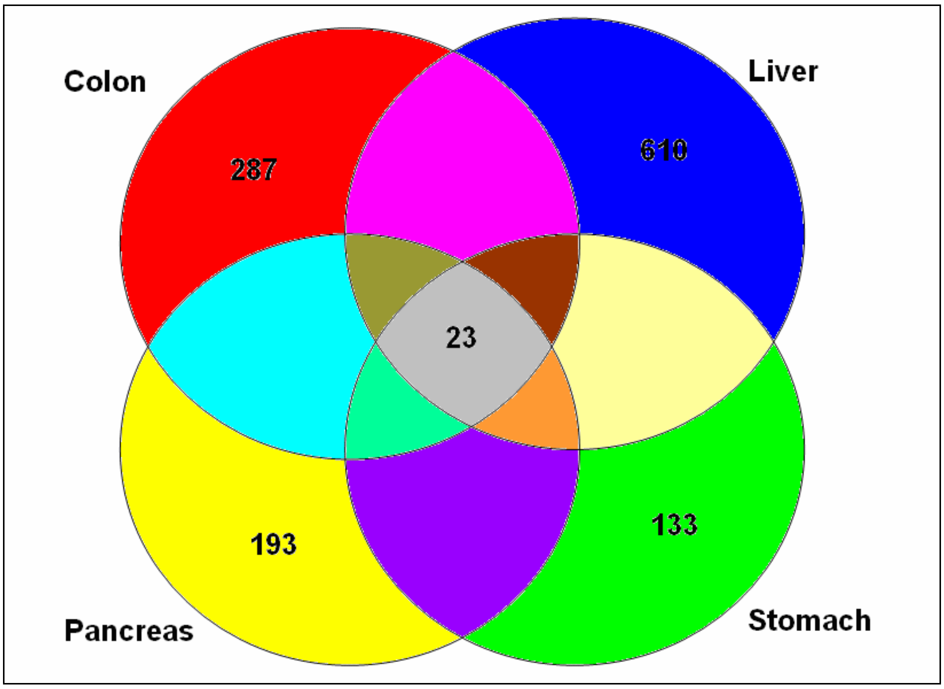


Figure 21: Intersection of OMIM cancer-related gene lists for the colon, liver, pancreas and stomach tissues. The grey coloured overlapping part in the middle represents the number of common cancer-related genes among those four tissues.

On the other hand, when putting together the cancer-related gene lists from breast, cervical and prostate tissues, nineteen genes are found present in those three tissues (Figure 22; Appendix II). Again, when having checked each gene's association with the pathways via KEGG system, only eight out of the nineteen genes are considered to have a role in one or more pathways. Those eight genes, still, do not share any roles in one or more common pathways.

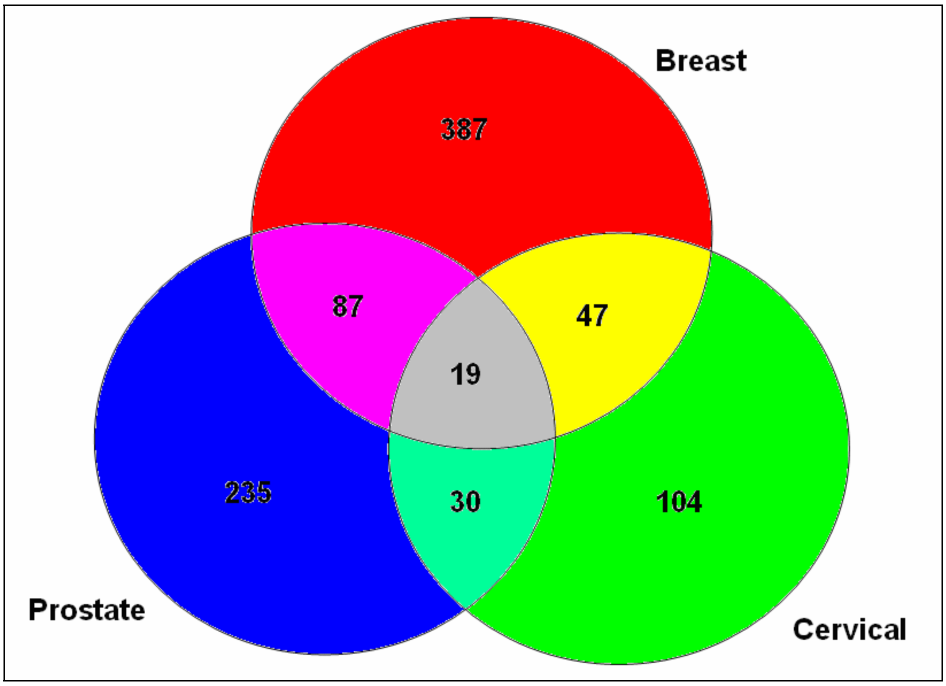


Figure 22: Intersection of OMIM cancer-related gene lists for the breast, prostate and cervical tissues. The grey coloured overlapping part in the middle represents the number of common cancer-related genes among those three tissues.

The combined summary for the intersection of gene lists from liver, stomach, pancreatic and colon along with the of gene lists integration from breast, cervical and prostate can be seen below in Table 9. We have also completed one other intersection of cancer-related gene lists for breast, cervical, prostate, liver, lung, pancreatic and stomach tissues. Genes APC, BRAF, CDKN2A, PTEN and TP53 are resulted from this intersection.

Table 9: Combined summary for the intersection of gene lists from liver, stomach, colon and pancreas plus the gene lists intersection from breast, cervical and prostate in relation to the biological pathways.

Pathway Names	**Gene Symbol (from Liver, Stomach, Colon and Pancreas Interactions)**	**Gene Symbol (from Breast, Cervical and Prostate Interactions)**
Cellular Processes Adherens junction	CDH1, IQGAP1	
Cellular Processes Apoptosis	TP53	TP53
Cellular Processes Axon guidance	DCC	
Cellular Processes Cell cycle	CDKN2A, MADH4, TP53	CDKN2A, CHEK2, TP53
Cellular Processes Focal adhesion	BRAF, HRAS	BRAF
Cellular Processes Regulation of actin cytoskeleton	APC, BRAF, HRAS, IQGAP1	AFC, APC, BRAF
Environmental Information Processing Hedgehog signaling pathway	IHH	
Environmental Information Processing MAPK signaling pathway	BRAF, HRAS, TP53	NF1, BRAF, TP53
Environmental information processing phosphatidylinositol signaling system and metabolism inositol phosphate metabolism	PTEN	PTEN
Environmental Information Processing TGF-beta signaling pathway	MADH4	
Environmental Information Processing Wnt signaling pathway	APC, FZD4, MADH4, TP53	APC
Human Diseases Amyotrophic lateral sclerosis (ALS)	TP53	TP53
Human Diseases Huntington's disease	TP53	TP53
Metabolism Fatty acid biosynthesis (path 1)		FASN
Metabolism Fatty acid biosynthesis (path 2)		BASE, FASN
Prostaglandin and leukotriene metabolism		PTGS2
Reactome Event:Cell Cycle Checkpoints 69620		CHEK2
Reactome Event:DNA Repair 73894	XPA , XPC	BRCA1, BRCA2

4.2 Microarray Gene Expression

4.2.1 Cancer-Related Genes from GEO Database

After thoroughly reviewing and analyzing the downloaded datasets from each publication, we have done individual analysis for each of the dataset. A list of 1,471 breast cancer-related genes has been obtained after the analysis of the breast carcinoma datasets. A total of 1,928 cancer-related genes for hepatocellular carcinoma have been found to express differently between normal and cancer tissues of liver tissues. A total of 1,531 cancer-related genes for lung carcinoma were resulted after the analysis. 2,613 cancer-related genes for pancreatic cancer have been confirmed to have quite different expression level between normal and cancer tissues while 829 cancer-related genes are resulted from the analysis of prostate cancer datasets.

As we derived the cancer-related genes from the analysis of microarray gene expression datasets for breast, liver, lung, pancreas and prostate tissue types, we performed an intersection among those five lists to examine if there are any common cancer-related genes. We have received a total of nine genes that are present across all five cancer types (Figure 23). Those nine common cancer-related genes are the FOXM1, HNRPDL, BIN1, BUB3, CCNI, PMS1, PRKCBP1, PURA and RPA3. Unfortunately, when we look up the biological pathway involvements via KEGG database for those nine genes, we could not locate any of those nine genes in any of the biological pathways.

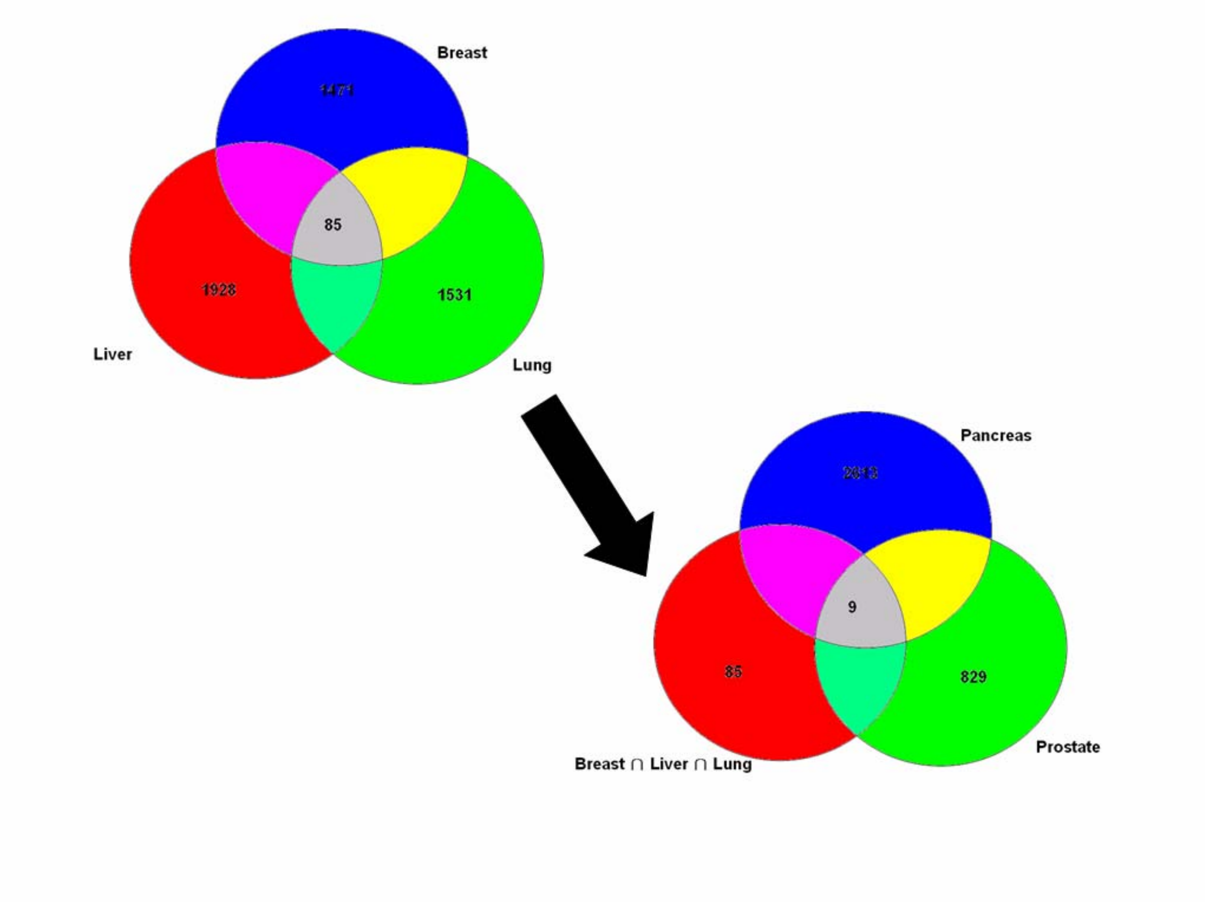


Figure 23: Intersection of GEO cancer-related gene lists for the breast, liver, lung, prostate and pancreatic tissues. The grey coloured overlapping part in the middle of the right bottom diagram represents the number of common cancer-related genes among those five tissues

4.2.2 Cancer-Related Genes from SMD Database

After comprehensively reviewing and analyzing downloaded datasets from each publication, we have done individual analysis based on the nature of the datasets. Genes that have expression level varies greatly between normal and cancer tissues have been filtered out for each cancer type. A list of 1,849 prostate cancer-related genes has been extracted out from Lapointe *et al.*'s prostate cancer datasets. A total of 1,077 cancer-related genes for breast carcinoma have been found to express differently between normal and cancer tissues using Sorlie *et al.*'s datasets for analysis. Garber *et al.*'s datasets have got a list of 3,653 cancer-related genes for lung carcinoma after the analysis while Chen *et al.*'s datasets have obtained 2,888 cancer-related genes for gastric cancer. Moreover, a list of 2,357 cancer-related genes from Chen *et al.*'s hepatocarcinoma data have been confirmed to have rather different expression level between normal and cancer tissues while 3,818 cancer-related genes are resulted from the analysis of Iacobuzio-Donahue *et al.*'s pancreatic cancer data.

Upon having derived the cancer-related genes from the analysis of microarray gene expression datasets for breast, liver, lung, pancreas, prostate and stomach tissue types, we intersected those six lists to see if there are any common cancer-related genes. We have determined a total of six genes that are present across all six cancer types (Figure 24). Those six common cancer-related genes are ARGBP2, CD53, FCGBP, JUN, MME and VBP1. Among those six genes, gene JUN is found to be actively involved in the environmental information processing MAPK signaling pathway, Wnt signaling pathway, Toll-like receptor signaling pathway, T cell receptor signaling pathway, focal adhesion and B cell receptor signaling pathway. In addition, gene MME is also being recognized to have a role in both the Alzheimer's disease and the Hematopoietic cell lineage. Gene ARGBP2, CD53, FCGBP, and VBP1, however, are not yet found to have participated in any of the biological pathways.

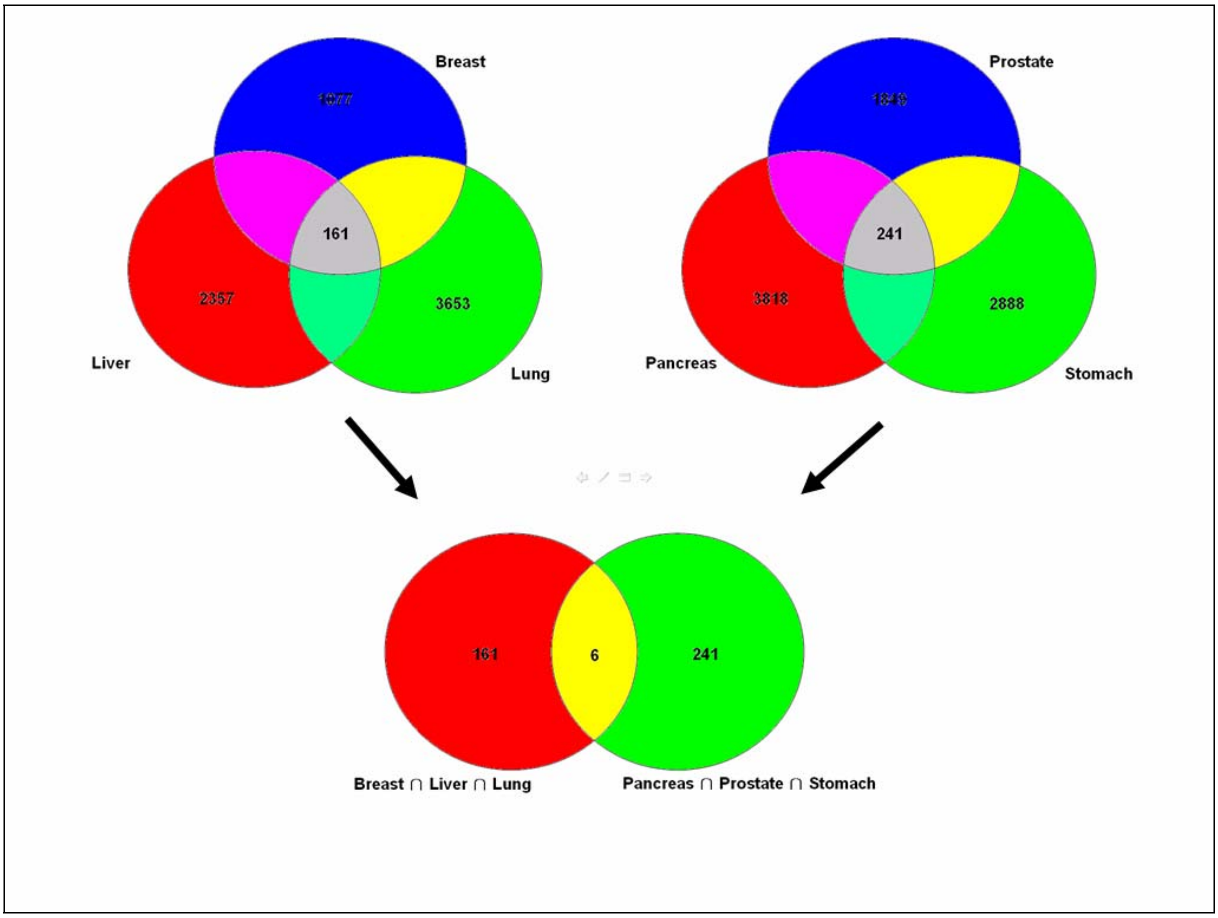


Figure 24: Intersection of SMD cancer-related gene lists for the breast, liver, lung, prostate, pancreatic and stomach tissues. The yellow coloured overlapping part in the middle of the bottom diagram represents the number of common cancer-related genes among those six tissues.

4.3 Expression Patterns for Cancer-Related Genes

4.3.1 GEO-Downloaded Expressions for OMIM Cancer-Related Genes

Figure 25 has demonstrated the average expression level of the five OMIM cancer-related genes among the Affymetrix datasets obtained from the GEO database. Apparently none of these five cancer-related genes has shown significant difference in expression levels between normal and cancer tissues; therefore, those five genes were not in the common cancer-related gene list for the GEO database. Across those five cancer-related genes, the BRAF gene seems to have fairly close expression levels while the expression levels for other four cancer-related genes vary greatly across five tissues. BRAF is a kind of protein that plays a central role in both the growth and survival of cancerous cells. The mutation of BRAF gene often leads to malignant melanoma, and sometimes also causes lung, colon or breast cancers (BRAF Mutation Website, 2006). By looking at Figure 25 for BRAF expression patterns, we found that the mean expression level for lung cancer tissues was marginally higher than that of the normal lung tissues, in which the ratio of difference equals 1.97. For BRAF expression levels in breast tissues, however, the ratio between the cancer and normal tissues approximately equals to 1.23, which is lower than the filtering criteria of 1.5 fold difference. As a result, the BRAF gene was not able to pass the filtering criteria to be included in the GEO cancer-related gene list.

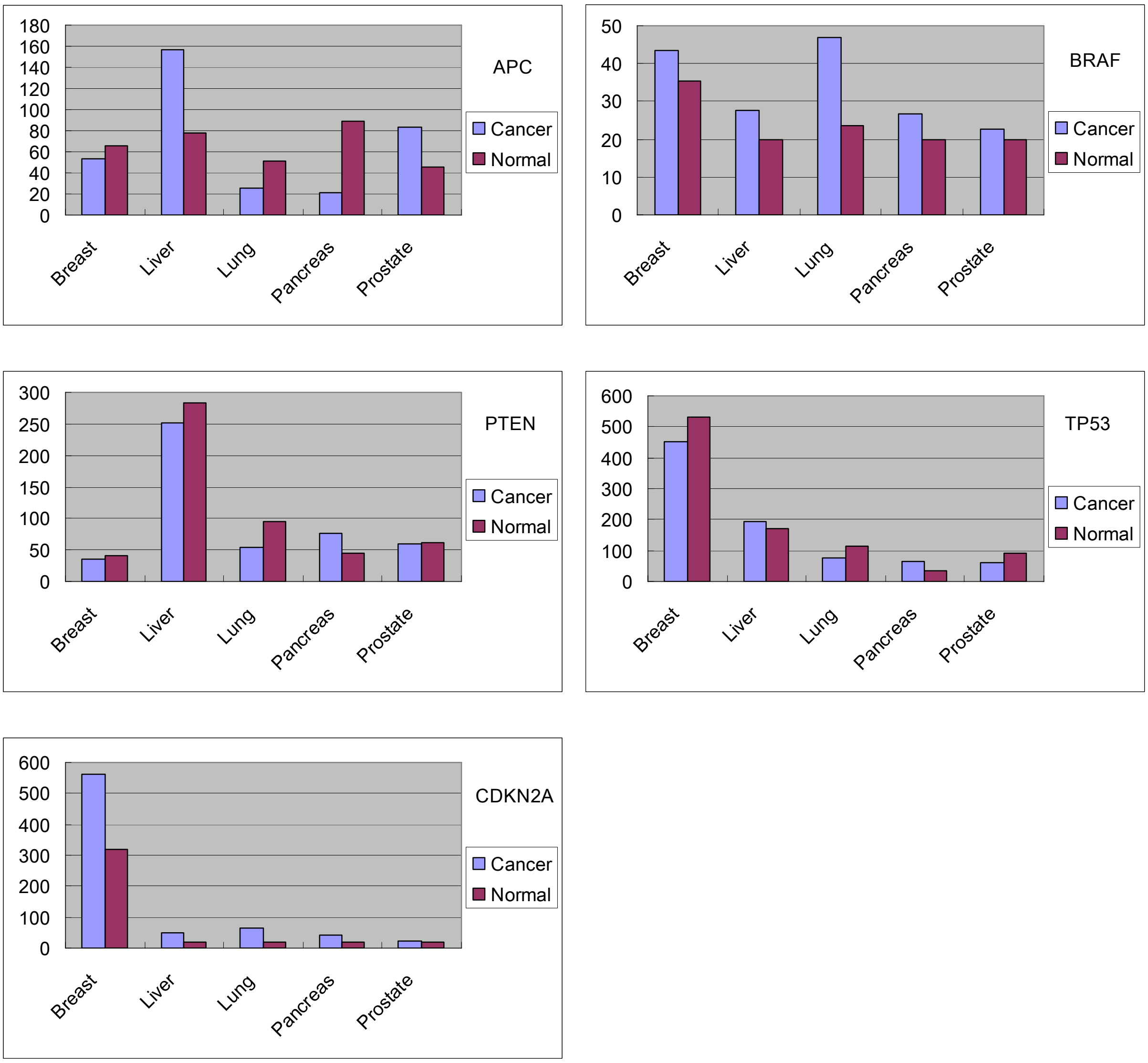


Figure 25: Gene expression levels from GEO database for OMIM cancer-related genes.

4.3.2 SMD-Downloaded Expressions for OMIM Cancer-Related Genes

Figure 26 has shown the average expression level of the five OMIM cancer-related genes among the datasets obtained from SMD database. None of these five cancer-related genes has demonstrated significant difference in expression levels between normal and cancer tissues. For example, only lung (ratio=0.59), prostate (ratio=0.45) and stomach tissues (ratio=1.87) for PTEN gene passed our expression filtering criteria to be included in the cancer-related gene lists for the above three tissue types, but not for breast (ratio=0.72), liver (ratio=0.88), pancreas (ratio=0.86). As a result, the PTEN was not on the common cancer-related gene list for the microarray datasets although it is well-known for its function as a tumour suppressor gene, which means it helps to ensure that the cell grows, divides and dies in a controlled manner (PTEN: Genetics Home Reference Website, 2005). Moreover, the mutations of PTEN genes usually are found in a large number of human tumours, which include cancers of the breast, prostate, colon and lymphoma (PTEN Mutation Website, 2005). Based on the microarray expressions, the normal and cancer expression for breast tissues are quite at the same level; therefore, the PTEN gene was not able to pass the filtering criteria in this case.

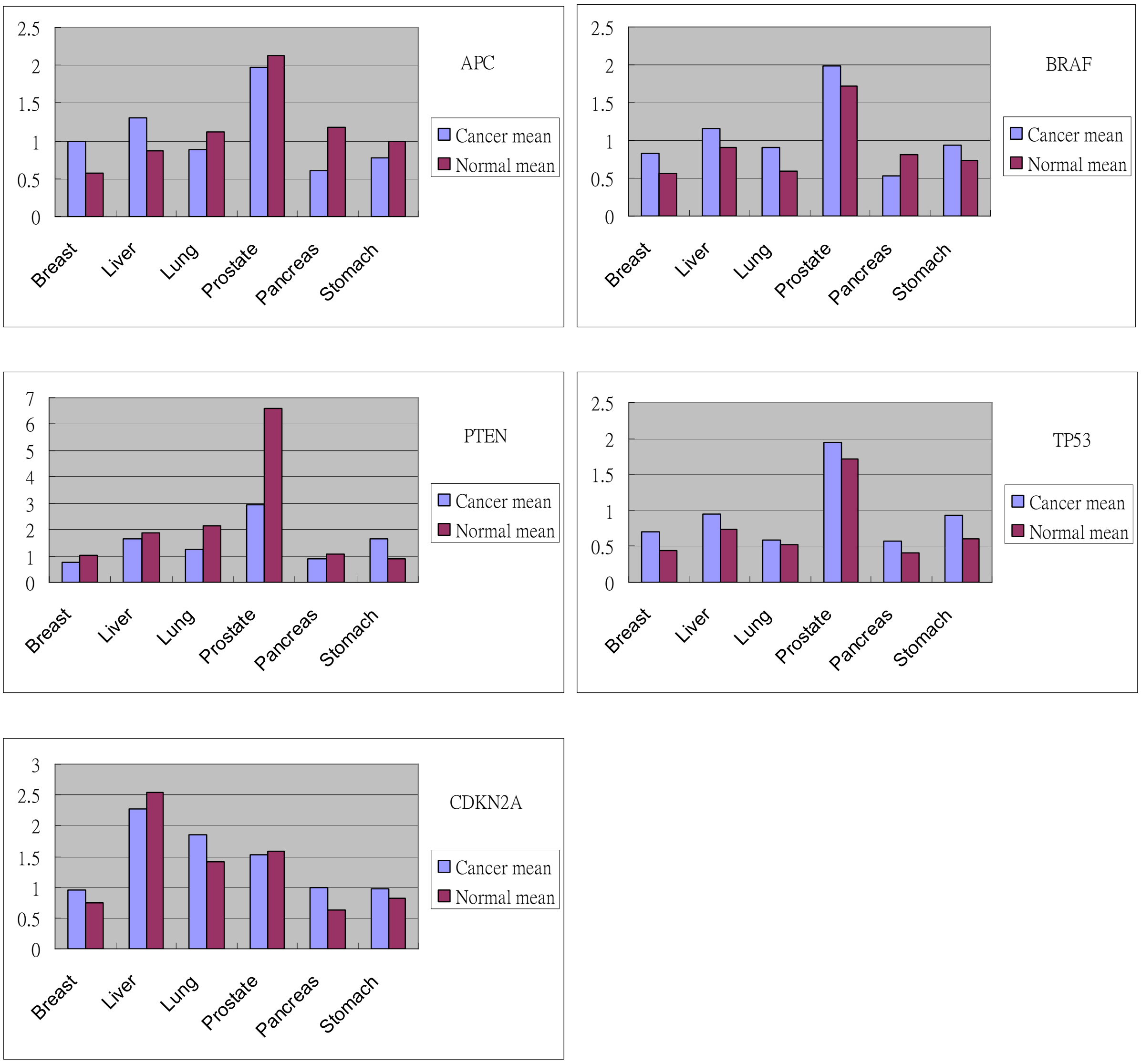


Figure 26: Gene expression levels from SMD database for OMIM cancer-related genes

4.4 Gene Ontology for the Cancer-Related Genes

The cancer-related genes obtained from OMIM, GEO and SMD databases were then imported into the Fatigo.Org – Data Mining for Gene Ontology (http://www.fatigo.org/) website to look for any gene ontology relationships among those genes. We would focus on the molecular function section of the gene ontology. Figure 27 has shown the molecular function of the gene ontology for five OMIM cancer-related genes. As shown in the figure, 40% of the OMIM cancer-related genes are involved in hydrolase activity, protein binding, nucleotide binding and transferase activity.

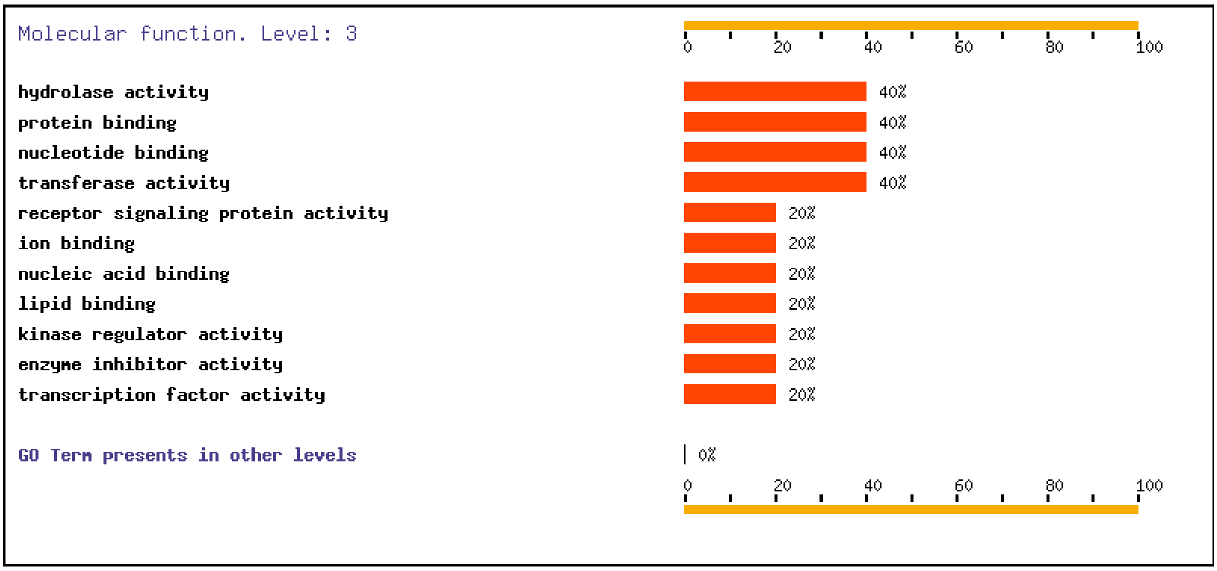


Figure 27: Gene ontology for OMIM cancer-related gene in terms of molecular function. (Fatigo.Org Website, 2005)

The gene ontology for nine GEO cancer-related genes in relation to their molecular functions is illustrated in Figure 28. As shown in the figure, about 85.71% of the GEO cancer-related genes are confirmed to participate in the nucleic acid binding.

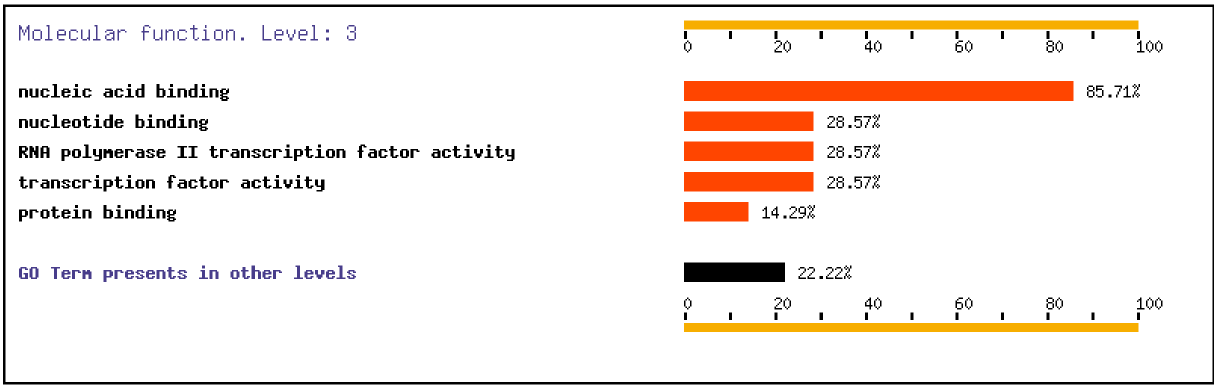


Figure 28: Gene ontology for GEO cancer-related genes in terms of molecular function. (Fatigo.Org Website, 2005)

Lastly, the molecular function of the gene ontology for six SMD cancer-related genes is demonstrated in Figure 29. From the figure, we could see that 66.67% of the SMD cancer-related genes are related to protein binding, ion binding and nucleic acid binding.

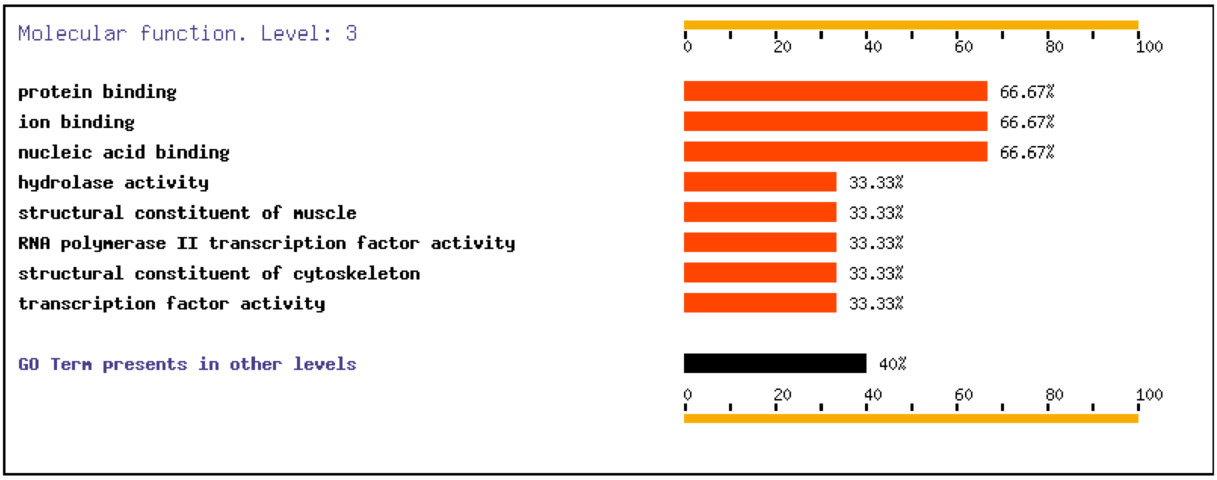
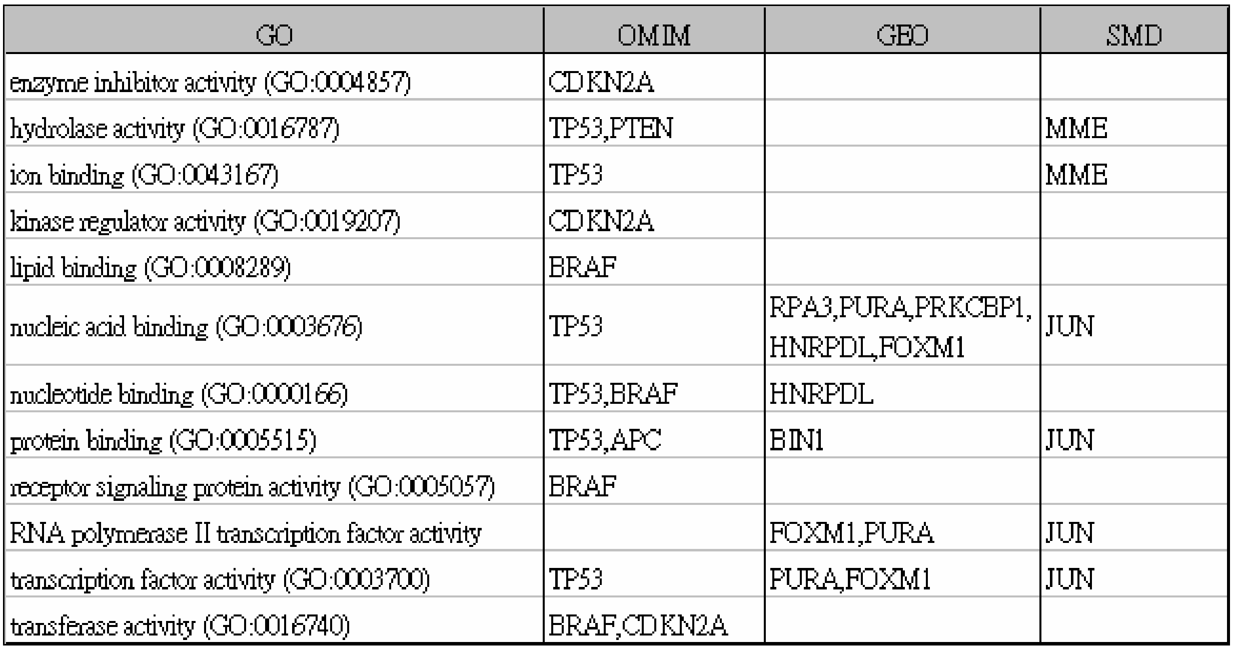


Figure 29: Gene ontology for SMD cancer-related genes in terms of molecular function (Fatigo.Org Website, 2005).

By classifying the cancer-related genes from OMIM, GEO and SMD based on their gene ontology in terms of molecular functions, we have found that OMIM cancer-related gene, TP53, is linked to hydrolase activity, ion binding, nucleic acid binding, nucleotide binding, protein binding and transcription factor activity. The TP53 gene has been recognized in six out of twelve GO terms. Among GEO cancer-related genes, PURA and FOXM1 genes are shown to be associated with nucleic acid binding, RNA polymerase II transcription factor activity and transcription factor activity. For SMD cancer-related genes, however, only MME and JUN are related to the GO terms in the molecular function category.

Table 10: Gene ontology for cancer-related genes from OMIM, GEO and SMD in the molecular function category



GO	OMIM	GEO	SMD
enzyme inhibitor activity (GO:0004857)	CDKN2A		
hydrolase activity (GO:0016787)	TP53,PTEN		MME
ion binding (GO:0043167)	TP53		MME
kinase regulator activity (GO:0019207)	CDKN2A		
lipid binding (GO:0008289)	BRAF		
nucleic acid binding (GO:0003676)	TP53	RPA3,PURA,PRKCBP1, HNRPDL,FOXM1	JUN
nucleotide binding (GO:0000166)	TP53,BRAF	HNRPDL	
protein binding (GO:0005515)	TP53,APC	BIN1	JUN
receptor signaling protein activity (GO:0005057)	BRAF		
RNA polymerase II transcription factor activity		FOXM1,PURA	JUN
transcription factor activity (GO:0003700)	TP53	PURA,FOXM1	JUN
transferase activity (GO:0016740)	BRAF,CDKN2A		

V. Discussion and Conclusion

5.1 Biological Pathway for OMIM and Microarray Cancer-Related Genes

From our analysis on both the OMIM and the microarray cancer-related genes, we have obtained a list of pathways which have one or more defined cancer-related genes present. While there are quite a few biological pathways involved in cancer development process, we would focus our attention on the *Wnt signaling pathway*.

Wnt signaling pathway is essential in various biological processes throughout the daily life. The OMIM defined cancer-related genes APC and TP53, as well as the microarray-defined cancer-related gene JUN, can be found in this particular pathway. According to some reports, chronic activation of the *Wnt signaling pathway* can result in the development of human malignancies, including hepatocellular carcinoma, colorectal carcinoma, ovarian cancer, etc. (WNT Signaling Pathway Website, 2005). Mutations in various regular genes, such as APC, as well as in other pathway components have been confirmed to have major impacts in causing tumorigenesis in human. Specifically for gene APC, more than 90% of the mutation currently reported for APC gene are related to colorectal carcinoma, while there are still a few reports indicating that the mutation cause results in breast cancer, hepatocellular carcinoma, pancreatic carcinoma and some other major cancer types (APC Introduction Website, 2006). Figure 30 has shown a graphic representation of the pathway with various gene involvements starting from extracellular of the cell to the cell nucleus.

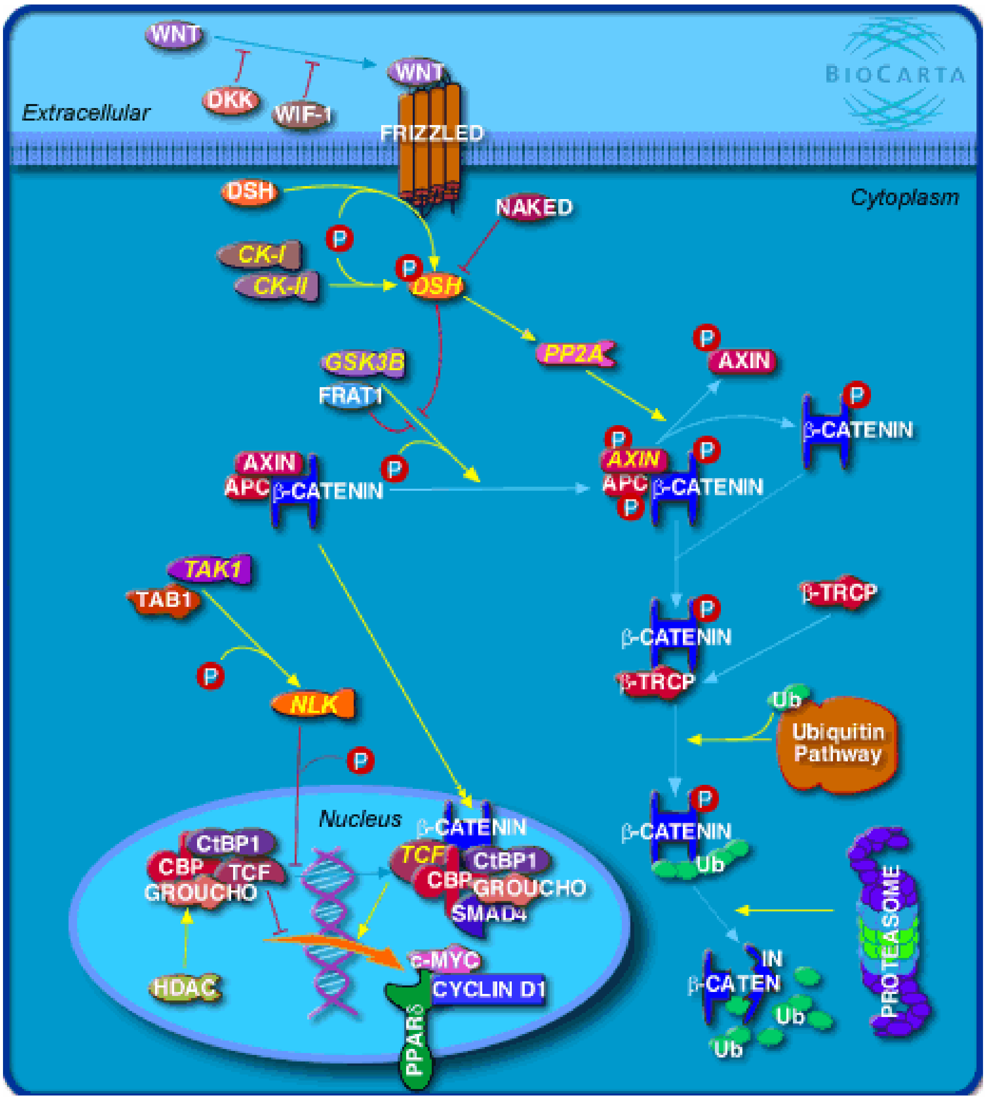


Figure 30: Schematic presentation of the Wnt signaling pathway.
(*http://www.biocarta.com/pathfiles/h_wntPathway.asp*)

5.2 Comparison between OMIM and Microarray Databases

Each OMIM record has consisted of information on many literatures that are reviewed and summarized in few sentences by scientific personals. The literatures that are being reviewed are published in either well-known journals or conferences. In other word, OMIM database acts like an evident and resourceful treasure box for us to dig and extract the cancer-related information, especially the genes that are involved in causing cancers. As for the microarray databases, both the cDNA microarray from the SMD database and the oligonucleotide arrays from the GEO database use the high throughput method to monitor gene expression levels. Although both cDNA and oligonucleotide arrays are able to view great numbers of gene expressions at the same time, experimental deviations and analytical variations usually influence the interpretation of the results.

5.3 Perspective

With the available common cancer-related gene lists, it is possible for researchers to do animal studies or use human samples to perform microarray experiments to find out whether mutations of those genes do lead to tumourgenesis. Moreover, researchers can further verify gene functions for this group of cancer-related genes to confirm if they do possess any not-yet-determined roles inside our body. By this means, the cancer-related genes can finally be identified and more effective cancer treatment targeted on those genes can be developed.

5.4 Conclusion

Cancers are certainly complex diseases with multiple genetic and environmental factors contributing to their development. Most cancers, however, are sporadic and appear in people who do not have a clear family history of the disease. As a result, this research is aimed to combine OMIM cancer-related gene list and microarray gene expressions to help discover unexpected trends and patterns from large sets of data. From our research, we have found that the five OMIM cancer-related genes do not match with either the nine oligonucleotide microarray cancer-related genes or the six cDNA microarray cancer-related genes. Moreover, OMIM mentioned cancer-related genes were not necessarily supported by microarray gene expression patterns

VI. References

Affymetrix. Statistical algorithms reference guide, Technical report, Affymetrix. 2001.

Aranda-Anzaldo A. Cancer development and progression: a non-adaptive process driven by genetic drift. Acta Biotheor 2001;49(2):89-108.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19(2):185-93.

Chen X, Leung SY, Yuen ST, et al. Variation in gene expression patterns in human gastric cancers. Mol Biol Cell 2003;14(8):3208-15.

Chen X, Cheung ST, So S, et al. Gene expression patterns in human liver cancers. Mol Biol Cell 2002;13: 1929-1939.

Claverie JM. Computational methods for the identification of differential and coordinated gene expression. Hum Mol Genet 1999;8(10):1821-32.

Garber ME, Troyanskaya OG, Schluens K, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci USA 2001;98(24):13784-13789.

Gibson GaM, S.V. A primer of genome science. Sinauer Associates, Inc: USA 2002.

Gudermann T, Grosse R, Schultz G. Contribution of receptor/G protein signaling to cell growth and transformation. Naunyn Schmiedebergs Arch Pharmacol 2000;361(4):345-62.

Harris MA, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32:D258-61.

Iacobuzio-Donahue CA, Maitra A, Olsen M, et al. Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. Am J Pathol 2003; 162(4):1151-1162.

Lapointe J, Chunde L, Higgins JP, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci USA 2004;101(3):811-6.

Liotta L, Petricoin E. Molecular profiling of human cancer. Nat Rev Genet 2000;1(1):48-56.

Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Genet 1999;21(1 Suppl):20-4.

Lu YaH, J. Cancer classification using gene expression data. . Information Systems 2003;28:243-68.

Mecham BH, Klus GR, Strovel J, et al. Expression genomics of cervical cancer: molecular classification and prediction of radiotherapy response by DNA microarray. Nucleic Acids Res 2004;32(9):e74.

Mecham BH, Klus GT, Strovel J, et al. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. Nucleic Acids Res 2004;32(9):e74.

Nelson DLaC, M.M. Lehninger Principles of Biochemistry, Third Edition. Worth Publishers: United States 2000.

Ogata H, Goto S, Fujibuchi W, Kanehisa M. Computation with the KEGG pathway database. Biosystems 1998;47(1-2):119-28.

Rhodes DR, Yu, J, Shanker, K., Deshpande, N., Varambally, R. et al. Large-scale mata-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci USA 2004;101:9309-14.

Rhodes DR, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 2004;6(1):1-6.

Schadt EE, Li C, Su C, Wong WH. Analyzing high-density oligonucleotide gene expression array data. J Cell Biochem 2000;80(2):192-202.

Su AI, Cooke MP, Ching KA, et al. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A 2002;99(7):4465-70.

Su AI, Welsh JB, Sapinoso LM, et al. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res 2001;61(20):7388-93.

Tsao JL, Tavare S, Salovaara R, Jass JR, Aaltonen LA, Shibata D. Colorectal adenoma and cancer divergence. Evidence of multilineage progression. Am J Pathol 1999;154(6):1815-24.

Welsh JB, Sapinoso LM, Su AI, et al. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Res 2001;61(16):5974-8.

Yang YH, Dudoit S, Luu P, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30(4):e15.

Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinoma distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001; 98(19):10869-74.

Wong YF, Selvanayagam ZE, Wei N, et al. Expression genomics of cervical cancer: molecular classification and prediction of radiotherapy response by DNA microarray. Clin Cancer Res 2003;9(15):5486-92.

Web Resource

APC Introduction – CancerQuest : The Genes of Cancer. Available at: http://www.cancerquest.org/index.cfm?page=303. Access Jan 02, 2006.

BRAF Mutation – Mutations in the BRAF gene predict sensitivity to a novel class of cancer drugs. Available at: http://www.biologynews.net/archives/2005/11/06/mutations_in_the_braf_gene_predict_sensitivity_to_a_novel_class_of_cancer_drugs.html. Access Jan 02, 2006.

CGAP – Cancer Genome Anatomy Project. Available at: http://cgapncinihgov/. Access Jun 07, 2005.

DOH – Department of Health, Executive Yuan, Taiwan, R.O.C. Available at:

http://wwwdohgovtw/cht/indexaspx. Access Dec. 11, 2004.

Fatigo.Org – Data Mining for Gene Ontology. Available at: http://www.fatigo.org/. Access Dec. 20, 2005.

GEO – Gene Expression Omnibus. Available at: http://wwwncbinlmnihgov/geo. Access Jan 07, 2005.

KEGG: Kyoto Encyclopedia of Genes and Genomes. Available at: http://www.genome.ad.jp/kegg/. Access Jan 03, 2005

ICD-O – World Health Organization – International Classification of Diseases. Available at: http://www.who.int/classifications/icd/en/. Access Dec. 29, 2005

PTEN Mutation – Mutation in PTEN gene can cause cancer and autoimmune disease. Memorial Sloan-Kettering Cancer Center. Available at: http://www.scienceblog.com/community/older/1999/B/199901898.html. Accessed Dec. 27, 2005.

PTEN: Genetics Home Reference. Available at: http://ghr.nlm.nih.gov/gene=pten. Accessed Dec. 27, 2005.

SEER's Training Web Site - Cancer as a Disease. Available at: http://trainingseercancergov/module_cancer_disease/cancer_disease_homehtml. Access Jan 20, 2005.

SMD Database – Stanford Microarray Database. Available at: http://genome-wwwstanfordedu/microarray. Access Dec. 11, 2004.

WNT Signaling Pathway and Its Role in Human Solid Tumors. Available at: http://www.infobiogen.fr/services/chromcancer/Deep/WNTSignPathID20042.html. Access Dec. 20, 2005.

WHO Report – World Health Organization. Available at: http://wwwwhoint/cancer/en/2004. Access Dec. 11, 2004.

Appendix

Appendix I – Twenty-three cancer-related genes and their associated pathways from the intersection of OMIM cancer-related gene lists for the colon, liver, pancreas and stomach tissues

Gene Name	**Biological Pathways**
APC	Environmental Information Processing Wnt signaling pathway Cellular Processes Regulation of actin cytoskeleton
BRAF	Environmental Information Processing MAPK signaling pathway Cellular Processes Focal adhesion Cellular Processes Regulation of actin cytoskeleton
CDKN2A	cellular process cell cycle
PTEN	environmental information processing phosphatidylinositol signaling system metabolism inositol phosphate metabolism
HRAS	Environmental Information Processing MAPK signaling pathway Cellular Processes Focal adhesion Cellular Processes Regulation of actin cytoskeleton
FZD4	Environmental Information Processing Wnt signaling pathway
IQGAP1	Cellular Processes Adherens junction Cellular Processes Regulation of actin cytoskeleton
CDH1	Cellular Processes Adherens junction
TP53	Environmental Information Processing MAPK signaling pathway Cellular Processes Cell cycle Cellular Processes Apoptosis Environmental Information Processing Wnt signaling pathway Human Diseases Amyotrophic lateral sclerosis (ALS) Human Diseases Huntington's disease
PHB, DCC MADH4 CEACAM5 KLK9, PTPRH MEN1, AFP IHH, XPC HFE, SHH XPA, RCV1	X

Appendix II – Nineteen cancer-related genes and their associated pathways from the intersection of OMIM cancer-related gene lists for the breast, prostate and cervical tissues.

APC	Environmental Information Processing Wnt signaling pathway Cellular Processes Regulation of actin cytoskeleton
BRAF	Environmental Information Processing MAPK signaling pathway Cellular Processes Focal adhesion Cellular Processes Regulation of actin cytoskeleton
CDKN2A	cellular process cell cycle
PTEN	environmental information processing phosphatidylinositol signaling system and metabolism inositol phosphate metabolism
TP53	Environmental Information Processing MAPK signaling pathway Cellular Processes Cell cycle Cellular Processes Apoptosis Environmental Information Processing Wnt signaling pathway Human Diseases Amyotrophic lateral sclerosis (ALS) Human Diseases Huntington's disease
NF1	Environmental Information Processing MAPK signaling pathway
FASN	Metabolism Fatty acid biosynthesis (path 1) Metabolism Fatty acid biosynthesis (path 2)
BASE	Metabolism Fatty acid biosynthesis (path 2)
CHEK2, DDX26 CTAG1B, BRCA2 DLC1, ESR1 TBC1D3, BRCA1 PTGS2, AXUD1 KAI1	X