Identification of IDH1 Mutation-Related Gene Signature of Glioblastoma Multiforme

Master's thesis

Zhaoran Zhou
Institute of Biomedical Technology
University of Tampere

May, 2014

Acknowledgements

The studying experience in Finland bestows me much benefits. Thus I would like thank to Finland for giving me the opportunity to study.

I would like to give the most gratitude to my supervisor Matti Nykter, who has helped me throughout the working on this thesis. I deeply appreciated his instructive advices and patient guidance, which are indispensable for completing this thesis.

I took 25 courses in total during the master degree studying, and I thank to all the instructors for what they have taught me.

I am very grateful to the staff of IT service department, for their helping in providing the necessary support and equipment to write and accomplish my thesis.

Finally, I would like to express my thanks to my parents for their continuous supporting.

Master's thesis

Place:	University of Tampere Institute of Biomedical Technology Computational Biology Group
Author:	Zhaoran Zhou
Title:	Identification of IDH1 Mutation-Related Gene Signature of Glioblastoma Multiforme
Pages:	74 and 5 appendix pages
Supervisors:	Prof. Matti Nykter
Reviewers:	Prof. Matti Nykter, Dr. Juha Kesseli
Date:	May, 2014

Abstract

Background

Glioblastoma multiforme (GBM) is a type of commonly occurred malignant astrocytoma with an extremely poor prognosis. GBMs display a remarkable genetic variability, and it is essential to study the genomic alterations and pathway dysregulations based on the different tumor entities.

The gene IDH1 encodes cytosolic isocitrate dehydrogenase 1, which catalyzes the reactions of oxidative decarboxylation of isocitrate to α–ketoglutarate. Different types of mutation of IDH1 has been found in gliomas and GBMs, especially in secondary GBMs. Among the IDH1 mutations, R132H mutation is the most prominent one. IDH1 mutation in GBMs is correlated with a longer survival time, and no IDH1 mutations are reported in many other tumor types. Thus IDH1 is hypothesized as crucial in the pathogenesis of GBMs, and it is regarded as a potential drug target.

The fundamental goal of this study is to identify a gene signature correlated with IDH1 mutation in GBMs. And related genes and biological pathways are also studied.

Methods

Most of the work of data collection and analysis are achieved with R pacakges. The step-down maxT method is adopted to perform the multiple testing procedure in order to find differently expressed genes. The p-values of statistical tests are corrected by controlling FWER. The clustering result is explicated as heatmap, and clinical data is elucidated with boxplot and Kaplan Meier-plot. Analysis of GO and KEGG pathways are used to extract more information from the genes. And the results are visualized as graphs in Cytoscape.

Results

A framework is created for identifying gene signatures as well as studying biological pathways. The expression data from 548 samples are collected, and 58 genes out of 12042 genes are identified as differently expressed. Finally a gene signature with 50 genes are proposed.

Conclusion

Microarray technology and statistics methods are effective for the studying of alterations in gene expression and biological pathways. The gene signature proposed by this study can distinguish samples harboring IDH1 mutation from GBMs. And future researches are necessary to corroborate and extend the results.

Contents

1. Introduction

Glioblastoma multiforme (GBM) is a common brain malignancy[1], and it is one of the most lethal and treatment-refractory cancer[2]. Numerous studies has been conducted in order to understand the biology of GBM and develop novel treatments. However, there is no significant breakthrough in this field. As the name indicated, GBM is exemplified by the cytologic and histologic variation and contains extensive genetic and biological variability[3]. The tumorigenesis of GBM is complicated by the diverse dysregulation of genome. Based on distinct genomic features, GBMs are divided into 4 subtypes: Proneural, Neural, Classical and Mesenchymal. Due to the heterogeneity of GBMs, it is necessary to study the aberrant pathways and phenotype in terms of different molecular characteristics.

IDH1 (Isocitrate Dehydrogenase 1) is enzyme function as a catalyst to oxidatively decarboxylate isocitrate producing α–ketoglutarate (αKG or 2-OG). During this process, NADP+ is reduced to NADPH[4]. IDH1 has been found as frequently mutated in GBM and be associated with increase in overall survival[5]. Among all types of IDH1 mutations, R132H mutation is the most commonly found one. However, mutant IDH1 has not been found in a wide ranges of cancers so far[6]. In addition, more mutant IDH1 is detected in secondary GBMs and younger GBM patients. Therefore, IDH1 is speculated to play a unique role in tumorigenesis of GBM, and studying on IDH1 can facilitate the development of novel therapy.

Microarray technology is widely used to quantitatively monitor the expression level of thousands of genes simultaneously. Using the microarray expression data, this thesis is concentrated on identifying gene signature which is correlated with IDH1 mutation state in GBM samples. The GBM samples are divided into 2 groups (IDH1+ and IDH1-) according to whether they harbor IDH1 mutation. The main method is to find the differently expressed (DE) genes by statistical tests. And the clinical data is collected to get insights of the difference of clinical features between GBM samples with and without IDH1 mutation. In the interest of extracting more information from gene signature, GO and KEGG pathways are analysis. Further, the gene signature is validated by hierarchical clustering. Hopefully, the results of this study can provide information for future studies on IDH1 mutation and aberrant pathways in GBM.

2. Literature Review

2.1 Cancer and Glioblastoma

Cancer is a class of diseases characterized by out-of-control cell growth. To be specific, the cells involved in cancer usually grow and reproduce in an uncontrollable way, and those cells breach and even destroy healthy tissue, including organs[7]. Without treatments, cancer is inclined to grow into worse progressively and can potentially lead to death[8].

The impaired cells may form abnormal mass of tissue or lumps, which are referred to as tumors. Tumors can grow and become harmful by interfering with the nervous, circulatory and digestive systems, or simply by pressing against nerves or blood vessels. Some tumors even result in releasing hormones that disrupt body functions.
Note that tumor is not equivalent of cancer, and it does not necessarily cause a health threat. Actually, tumors can be categorized basically as 3 groups: benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). And only the malignant tumors can be called cancer. Once the malignant tumors grow fast, they tend to become aggressive and spread to distant parts of the body by invasion and metastasis. However, it is difficult to draw clear lines of demarcation between benign, pre-malignant and malignant tumors. Some benign tumors may eventually become premalignant, and then malignant.

Cancer cells can migrate and penetrate into neighboring tissues directly by invasion. And metastasis is the process by which cancer cells penetrate into lymphatic and blood vessels, circulate through the bloodstream, and then invade normal tissues elsewhere in the body[9]. “Primary” means the original site of the cancer, while “secondary” implies any additional sites where the cancer has spread[10].

There are many different types of tumors, and scientists employ a variety of technical names to distinguish them. Tumors’ names usually indicate the locations they appear in and their shapes. For example, “blastoma” refers to as those tumors derived from embryonic tissue or immature “precursor” cells. And “blastoma” is often used as a suffix to describe tumors such as “glioblastoma”.

Tumor grading is usually based on the microscopic examination of a tumor and its abnormality[11]. Tumor grading assign most cancers a numerical grade, which indicating the likely behavior of a tumor and its responsiveness to treatments. Generally, the grading system gives a low number grade (grade I or II) to cells or tissues with fewer abnormalities, and scores higher numbers (grade III, IV) to those with more abnormalities. Higher-graded cancers tend to grow and spread faster with worse prognosis. The factors taken into consideration in tumor grading can vary between different types of cancer.

Glioma is a type of tumor that arise from glia, which is a tissue helps to keep the neurons in place and functioning well. Glioma can occur in brain and spine, and the former is more common. Like many other tumors, the degree of severity of a glioma depends on its grade. Grade I is the least serious and grade IV is the most serious.

There are 3 kinds of glia cells can produce tumors: astrocytes, oligodendrocytes, and ependymal cells. If the tumor is made up from more than one type of glia cells, it is called "mixed glioma". According to the histological details, gliomas can be divided as many subtypes, such as astrocytoma, dendroglioma, pilocytic astrocytoma, and ependymoma.

Astrocytic original gliomas, also called astrocytomas, are the most common type of glioma. Astrocytomas consist of various types, and they are thus graded on a scale from I to IV[12], ranging from the slowly growing juvenile pilocytic astrocytoma (grade I) to the highly malignant glioblastoma multiforme (Grade IV). Astrocytomas have been found in many part of the brain and nervous system, including the central areas of the brain, the brainstem, the cerebellum, the cerebrum and the spinal cord[13].

Glioblastoma multiforme (also called GBM, Astrocytoma Grade IV, and Glioblastoma Grade IV) is the most common astrocytomas. And it is also the most aggressive form of malignant primary brain cancer in adults. The GBM has one of the worst prognosis among human tumor types[14]. The word "multiforme" indicates the significant intratumoral heterogeneity on the cytopathological, transcriptional, and genomic levels[15]. The GBM usually occurs in frontotemporal region and parietal lobes. But it is rarely found in the cerebellum and spinal cord. Primary glioblastoma, as the name implies, arises de novo without antecedent history of low-grade disease, while secondary glioblastoma evolves progressively from previously diagnosed low-grade gliomas[16].

Glioblastomas are composed predominantly of poorly differentiated, fusiform, round, or pleomorphic cells[17]. There are few biomarkers of favorable prognosis and, accordingly, few therapies strongly influencing disease outcome. According to some sources, the GBM are defined as many different subtypes, which are caused by different genomic aberrations and require different therapeutic approaches[18]. The genomic-based classification of the subtypes of GBM facilitates scientists to get insight into the molecular mechanism leading to this disease. And the sequence-based mutation detection is a kind of effective method to study the GBM.

2.1.1 Hallmarks of Cancer

The proliferation, differentiation and death of normal cells are under the control of many factors, and the molecular machinery relating to that has been studied for a long time. Usually, normal cells will grow only when stimulated by the growth factors, and they need a blood supply. And once the cells are damaged, anti-growth signals will prevent them from dividing until they are repaired. If the cells cannot be repaired, they will die through the apoptosis. Normal cells can only divide a limited number of times and they always remain where they belong. Each mechanism is controlled by several proteins. In one word, normal cells can balance cellular proliferation and death. When all the mechanisms are "conquered", the normal cells transform into malignant tumor cells. Such disruptions of those mechanisms are caused by the damages of the related proteins: the corresponding genes are damaged by acquired or somatic mutations.

An influential and highly cited paper asserts that the development of human cancer is a multistep process, and the cells gain six biological capabilities during this progressive conversion. Such six biological capabilities are shared by all cancers as common traits, and they are called "hallmarks". Those hallmarks are: self-sufficiency in growth

signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis[19]. And the authors proposed four new hallmarks of human cancers as an updated, which are genome instability and mutation, tumor-promoting inflammation, reprogramming energy metabolism, evading immune destruction[20].

These cancer hallmarks provide a solid foundation for understanding the biology of cancer.

2.1.2 Gene expression-based molecular classification of GBM

An influential and highly cited paper classify GBM into 4 distinct subtypes based on the gene expression data. The subtypes of GBM comprises: Proneural, Neural, Classical, and Mesenchymal[21]. Although the subtypes share similar genomic aberrations, they have distinctiveness in terms of gene expression patterns, mutation and copy numbers. Proneural GBM is characterized as mutations of TP53 and IDH1as well as amplification of PDGFRA. Patients with Proneural GBM are significantly younger and they tend to survive longer. However, TP53 mutation is not found in Classical subtype, which is described as a high level expression of EGFR and EGFR vIII mutant. The main features of Mesenchymal subtype are the frequent mutation of NF1 and TP53 tumor suppressor gene and an epithelial-to-mesenchymal transition (EMT). Neural is a normal-like subtype, which do not contain significantly higher or lower rates of mutations. Neural subtype is the expression of several gene types of the brain's noncancerous nerve cells, or neurons. Patients of Neural subtype is oldest among the 4 subtypes. In response to aggressive treatment, Classical subtype presents longest survival time, while Proneural subtype seems cannot get any benefits from the treatment.

2.2 Genome Methylation and GBMs

DNA methylation is the first discovered epigenetic marks and remains the most studied. DNA methylation occurs almost exclusively in the context of CpG dinucleotides. CpG dinucleotides do not distribute randomly in genome. Most CpG dinucleotides are in CpG islands, which are regions which contain more than 500 bases with a CG content of more than 55%[22]. The ratio of CpG dinucleotides form in a CpG island and statistically expected CpG should be 0.65 at least. CpG islands are important, because there are about 60% human gene promoters having relation with them[23], and the methylation state of these CpGs are widely regarded as critical indicators of gene regulation. Most of human gene promoters are usually unmethylated in normal cells. Generally, CpG island methylation is associated with gene silencing[24]. Some methylated DNA can promote the recruitment of methyl-CpG-binding domain (MBD) proteins[25], which can recruit histone-modifying and chromatin-remodeling complexes to methylated sites. However, unmethylated CpG islands generate a chromatin structure preferring to recruit Cfp1, which is better for gene expression. Except CpG islands, CpG island shores are another kind of regions which are tend to occur DNA methylation. CpG island shores are close to CpG islands (the distant is not more than 2kb), and contain less CpG dinucleotides[26]. There are two features of methylation of CpG island shores: (1) They are highly conserved between human and mouse; (2) Most of them are conjectured to be tissue-specific. Therefore methylated CpG island shores can be used to distinguish different tissues. Genome methylation can not only inhibit gene expression, but also can promote it[27]. Besides, methylated CpGs in repetitive

elements are found to protect chromosomal integrity[28]. Non-CG methylation has been found in CHG and CHH sites[29], and 5-hydroxymethyl-2′-deoxycytidine has also been observed[30].

It has been conjectured that DNA methylation has an intimate connection with many human diseases. For example, cancer cells can be characterized by a massive global loss of DNA methylation and acquisition of specific patterns of hypermethylation[31]. And rise to hyper- and hypomethylated sites of DNA sequences may lead to a great deal of neurological diseases[32]. In fact, statistically significant associations are found between the DNA methylation states and histological subtypes and grades of gliomas. Notably, mutation of genes encoding the isocitrate dehydrogenase (IDH) is considered to be associated with distinct DNA methylation phenotype in gliomas. In addition, a pervasive and highly conserved DNA repair enzyme, O^6-methylguanine-DNA methyltransferase (MGMT), is associated with resistance to alkylating agent cancer therapy, which is applied to many patients with glioblastoma[33]. To be specific, the promoter methylation status of MGMT influences glioblastoma sensitivity to alkylating agent. Some researchers suggest that MGMT promoter methylation assessment could provide a prognostic or predictive biomarker for benefit from alkylator-based chemotherapy.

2.3 The predominant genes related to gliomas and GBMs

Following the brief introduction of genetic characters of GBM, a relatively detailed description of some important genes and corresponding enzymes are exhibited. As was discussed before, transformation from normal cells to glioma or GBM cells is a multistep process, whereby each genetic change confers a proliferative advantage.
Some cancer cells of GBM show stem cell-like features, and that is one of the reasons why GBM is resistant to many treatments and has a high recurs frequently. A deep understanding of those genes and enzymes will provide insights into the tumorigenesis of glioma and GBM.

2.3.1 Genomic and epigenetic characterization of GBM

GBM is characterized by distinctive histopathologic features such as cellular heterogeneity, necrosis, and endothelial proliferation. Numerous studies focusing on the genomic and epigenetic characters have been performed to comprehend the pathological mechanism of glioma and devise targeted therapeutics.

According to a widely cited paper[34], some important genetic events in GBM has been detected: (1) dysregulation of growth factor signaling via amplification and mutational activation of receptor tyrosine kinase (RTK) genes; (2) activation of the phosphatidylinositol-3-OH kinase (PI3K) pathway; and (3) inactivation of the p53 and retinoblastoma tumor suppressor pathways. Based on the previous studies and the powerful The Cancer Genome Atlas (TCGA) pilot project, many novel genomic characters of GBM or glioblastoma have been identified.

Many amplification and deletion events are found in GBM samples with the analysis of genomic copy number alterations (CNAs). Those CNAs affect the expression of many genes, including some known cancer-related genes. For example, the inactivating mutations of NF1, activations of EGFR family and mutations in PI3K complex are detected in most glioblastoma samples, and the aberration of TP53, PTEN, PDGFRA, ERBB2, MET, ARF, MDM2, MDM4, CDKN2B, CDK4 and RB1 genes occur

frequently. Those genes may play a key role in the development of golioblastoma. Besides, MGMT methylation, together with the mismatch repair (MMR) genes mutations, is found to be associated with the alteration of mutation spectrum of samples which are exposed to alkylating agent chemotherapy. Notably, although primary and secondary GBM have similar pathology, they carry distinct patterns of genetic abnormalities.

Glioblastomas also harbor more hypermethylated CpG loci relative to other types of gliomas. In fact, a study exhibits that the ratios of hyper- to hypomethylated CpG loci are statistically significantly different across glioma histological subtypes[35]. In addition, the hypermethylation of histone residue H3K9 occurs frequently in gliomas[36]. And a group of metabolic pathways are commonly hypomethylated in gliomas.

The genetic and epigenetic pattern of different GBM samples may be valuable and informative for the clinical decision−making. In other words, patients with different pattern of mutations should receive distinct treatments. And that is one of the reason for the analyses of genomic data.

2.3.2 FGF and FGFR

FGF (Fibroblast growth factor) family is an enzyme family with varied functions in regulating of cell mitogenesis, chemotaxis, angiogenesis, proliferation, migration and differentiation[37]. FGFs have been found involved in the development of many systems, including skeletal, nervous, and vascular systems. FGFs are highly conserved in both gene structure and amino-acid sequence. All of the FGFs share certain structural characteristics, and most of them can bind heparin strongly.

FGFs can signal cross epithelial-mesenchymal boundaries directionally and reciprocally. Some members of this family are crucial for the neuronal signal transduction in the central and peripheral nervous systems. And FGFs are homeostatic factors and play a role in angiogenesis, tissue repair and response to injury in adult organism. In addition to the functions in normal development, FGFs and FGF signaling pathway play significant roles in tumor development and progression. Some of FGFs are observed improperly expressed and contribute to tumors[38].

Fibroblast growth factor receptors (FGFR) are the receptors that bind to FGF proteins, and they are receptor tyrosine kinase (RTK). FGFRs transmit extracellular signals to various cytoplasmic signal transduction pathways through tyrosine phosphorylation FGFRs aid FGF signaling system to achieve diverse effects on diverse target cells[39]. Presently, 4 FGFRs (FGFR1-4) have been widely studied, and they share the common extracellular region containing 3 immunoglobulin (Ig)-like domains (designated IgI, IgII and IgIII), and thus belong to the immunoglobulin (Ig) superfamily[40]. Other important member of Ig superfamily include platelet-derived growth factor receptor (PDGFR), cluster of differentiation 7 (CD7) and Interleukin 1 receptor (IL-1R).

Binding with FGFRs, FGFs can activate many genetic programs and stimulate cell growth by promoting cell cycle progression and inhibiting pathways of cell death. If any step in the process becomes unregulated, the cells will grow beyond control and might lead to tumor. Therefore, components in such pathways are potential oncoproteins.

FGF signaling pathway is reported to stimulate GBM growth. FGF1 (acidic FGF or aFGF) shows an elevated expression in most tumor samples compared with the control brain tissues[41]. In an early research[42], FGF2, a pro-angiogenic molecule, is found expressing in most human gliomas, while FGF2 cannot be detected in mormal brain.

The expression of FGF2 also increases with the degree of malignancy and vascularity in human gliomas. In a recent study, FGF2, which can activate FGFR1-4, displays a growth-promoting effect in several GBM cell lines. Inhibition of FGF signaling pathway produces a small but significant growth inhibition of GBM cells in vitro[43]. In addition, FGF4 is also detected in tumor cells, and it is conjectured to participate in glioma angiogenesis. Both FGF2 and FGF4 are demonstrated to regulate vascular endothelial growth factor (VEGF) and the formation of new blood vessels[44]. However, the limited ability of FGF signaling pathway to promote GBM cell proliferation suggests that it is not the only pathway in driving GBM cell growth.

Not surprisingly, alterations in FGFRs can disrupt the FGF signaling pathways, and FGFRs are thus considered to be associated with tumors. Previous studies reveal that FGFR1 expression is significantly increased in malignant tumors relative to normal white matter[45]. By contrast, FGFR2 expression is absent in malignant astrocytomas, while abundant in normal white matter as well as in all low-grade astrocytomas[46]. Scientists hypothesize that FGFR1 signals through Mitogen-activated protein kinase (MAPK) pathway.

2.3.3 PDGF and PDGFR

PDGF (The family of platelet-derived growth factor) includes 4 members: PDGF-A, -B, -C, and -D. PDGFs participate in normal embryonic development, central nervous system (CNS) development, cellular differentiation, tissue homeostasis, and response to tissue damage. All members of PDGF family share the PDGF/VEGF homology domain, which is a highly conserved growth factor domain[47]. Due to the diverse transcriptional regulatory mechanisms and structures, PDGFs have numerous functions in the developments of normal systems and tumors.

Platelet-derived growth factor receptor (PDGFR), as its name implicates, is the receptor of PDGF. PDGFRα and PDGFRβ are the two isoforms of PDGFR, and they are encoded by genes PDGFRA and PDGFRB respectively. As mentioned above, PDGFRs are classified as receptor tyrosine kinases (RTK), and they share Ig-like domains and a split intracellular tyrosine kinase domain.

PDGFs activate the signal transduction pathways and downstream gene transcription events by signaling via PDGFRs. Different PDGFs have different affinity to PDGFRα and PDGFRβ[48].

Aberrant activity of the PDGFs and their receptors are associated numerous pathological conditions, and it has been reported that PDGFs and PDGFRs play important roles in the pathogenesis of gliomas[49]. Actually, overexpression and hyperactivity of PDGFs and PDGFRs are very common in human gliomas of all grades. The establishment of PDGF autocrine loop is not only an initial event of GBM progression, but also crucial in the late stages of GBM. PDGFs are also related to glioma angiogenesis, but the angiogenic effects of PDGFs are weaker than FGFs and VEGFs.

According to an early study, PDGFA and PDGFB are detected to be expressed in low-grade and anaplastic astrocytomas as well as in glioblastomas, and the expression correlates positively with tumor grade[50]. Particularly, PDGFA expresses at higher levels than PDGFB. In addition, PDGFC and PDGFD are also present in glioma and primary glioblastoma[51]. Notably, PDGFC is undetectable in normal fetal and adult brain tissues and PDGFD express as a lower level in normal brain tissues.

Amplification and overexpression of PDGFRs will lead to the activation of some important pathways in gliomas, including RTK signaling pathways, RAS/MAPK and PI3K pathways. PDGFRα is found at a high level in all grades of gliomas, while PDGFRβ is absent in glioma cells. And PDGFRα expresses in glioblastomas at highest level[52]. Nonetheless, another study shows that the expression of PDGFRβ as well as PDGFB is detected in hyperplastic tumor endothelial cells in glioblastoma[53]. It is possible that the expression of PDGFRβ is confined in tumor endothelial cells. And PDGFRβ also displays a positive correlation with glioma grades. To date, activating rearrangements of PDGFRA gene are rarely detected in gliomas.

2.3.4 EGF, TGF and EGFR

EGF (Epidermal growth factor) stimulates cell growth, proliferation, and differentiation. Overactive signaling of EGF system are detected in many aggressive cancers including GBM[54], and numerous evidences have demonstrated its importance in glioma transformation and angiogenesis.

TGF (Transforming growth factor) refers to 2 classes of polypeptide growth factors: TGFα and TGFβ. However, they are not similar in the structural or genetical aspects. TGFα shares 42% homology with EGF and regulates normal growth and development of many tissues[55]. TGFα is demonstrated as a mediator of the proliferation and transformation of human glioma cells, and it has been reported to involve in the angiogenesis of gliomas[56]. TGFα expresses in gliomas, and its expression is correlated with tumor grade as well as the expression of EGFR and Ki-67[57].

TGFβ serve as an inhibitor of proliferation in various systems. However, TGFβ is detected to be mitogenic for many glioma cell lines[58]. One of the possible explanation for the switched function of TGFβ is the dysregulation of the TGFβ signaling pathway. Besides, TGFβ protein induces expression of PDGFA, PDGFB, and PDGFRβ in glioma cells, and that is also a factor for TGFβ to convert from inhibitor to mitogen.

TGFβ plays a role in glioma angiogenesis. TGFβ expresses in glioblastoma but is almost absent in low-grade glioma or normal brain. And TGFβ is also inversely correlated with the survival of patients with malignant gliomas[59].

Epidermal growth factor receptor (EGFR, ERBB or ERBB1) is the receptor of EGF and TGFα, and it is a RTK. EGF and TGFα exert their effects on many pathways through binding EGFR. High level of expression in many types of cancer suggests that EGFR is strongly associated with the pathogenesis and tumor aggressiveness of multiple cancers. In fact, amplification and overexpression of EGFR is a striking feature of GBM[60]. The most common mutant form of EGFR is EGFRvIII, which is more oncogenic than wild type of EGFR (wtEGFR)[61]. EGFRvIII usually coexpresses

wtEGFR, and it is correlated with HB-EGF expression in GBM. However, understanding about the oncogenic potential of EGFRvIII is incompleted.

EGFR is rarely detected in normal glial cells but is widely expressed in human gliomas[62]. The gene encoding EGFR, is the most frequently amplified RTK gene in glioblastoma. Amplification of EGFR often occurs in many primary GBM samples and is associated with EGFR overexpression, whereas it is very rare in secondary GBMs. Particularly, EGFR amplification has a correlation with the histologic subtypes of GBM. Interestingly, EGFR gene rearrangements are observed in most samples with EGFR overexpression. Up to now many EGFR genetic alterations have been found, such as gene rearrangements, deletions, alternative splicing, and translational alterations. Those alterations could result in the expression of aberrant EGFR and contribute to an increased tumorigenicity. Evidence from studies about radiation and human head and neck carcinoma supports that the expression of EGFR is directly correlated with poor prognosis and radiation resistance[63]. Since inappropriate expression of EGFR contribute to the highly resistance to radiation treatments of GBM, EGFR signaling system is an attractive target for therapeutics designing.

2.3.5 NF1

NF1 (Neurofibromin-1) is a tumor suppressor and a negative regulator of the RAS signal transduction pathway[64]. Loss of NF1 expression will result in elevated activity of RAS, which is an important intracellular protonocoprotein in promoting cell growth and survival. Consequently, hyperactivation of RAS will activate a series of downstream intermediates, including AKT, and the mammalian target of rapamycin (mTOR).

Mutations in NF1 have been linked to the hereditary condition neurofibromatosis type-1, where patients are predisposed to glioma development[65]. Mutation or homozygous deletion of NF1 is also observed in some glioblastoma samples.

2.3.6 PTEN

PTEN (Phosphatase and tensin homolog deleted on chromosome TEN) is a protein containing a lipid-binding domain that allows anchorage to the plasma membrane. Since PTEN is a direct antagonist of the activity of PI3K, inactivation of PTEN will cause AKT hyperactivation and thus prompt the growth and proliferation of cells[66].

PTEN is originally discovery as tumor suppressor, which is mutated and lost in many types of cancer. A variety of mutations of PTEN are shown correlated with the development and progression of cancer. Mutations and deletions of PTEN are frequent and late events in high-grade gliomas, but rarely found in low-grade gliomas. Loss of PTEN is also significantly associated with a poor survival[67].

2.4 The predominant pathways related to glioma and GBM

Like many other types of cancer, a series of pathways malfunctions exist in GBMs. And those aberrance of pathways are essential for the normal cells to transform into tumor cells progressively and to become malignant. The deregulations of 3 pathways: RTK/RAS/PI3K signaling, the p53 and RB tumor suppressor pathways, are detected in most glioblastoma samples, implying the disruption of these pathways is a core requirement for glioblastoma pathogenesis. Moreover, with the recent advances in

technology and approaches, novel pathways contributing to gliomas and GBMs are reported[68].

2.4.1 The RB pathway

Members of RB (retinoblastoma protein) family are tumor suppressor proteins which are found as dysfunctional in some cancers. RBs function primarily as regulators of the mammalian cell cycle progression, and suppressors of cellular growth and proliferation. Generally, each RB bind and sequester distinct members of E2F family of transcription factors, and thus inhibit proliferation through repressing the transactivation of relevant genes[69]. Of note, E2Fs target genes encode proteins involved in DNA metabolism and synthesis and chromosomal replication. E2F DNA binding sites help to repress transcription in quiescent cells. However, activated CDK complexes by MAPK will phosphorylate RBs, enabling the expression of E2F-dependent genes that facilitate the G1/S transition and S-phase. Also, the $p16^{INK4a}$ transcribed from gene CDKN2A inhibits both CDK4 and CDK6 and maintains RB activation. Therefore, the inactivation of $p16^{INK4a}$ (the inhibitor of) will also disrupt RB functions[70].

Some genetic alterations inactivating RB pathway have been detected in gliomas. In high-grade glioma, amplification of the CDK4 and CDK6 gene, and mutation of RB1 gene are found. Markedly, inactivation of $p16^{INK4a}$ caused by allelic loss or hypermethylation prevails in cultured glioma cell line and high-grade gliomas[71], implying that $p16^{INK4a}$ is a very important suppressor of glioma tumor. In addition, loss of chromosome 13q is representative in the transition from low- to intermediate-grade gliomas[72].

2.4.2 The p53 pathway

Tumor suppressor p53 is a major regulator of multiple cellular responses encoded by the gene TP53. The p53 has been studied in a great depth and it is indispensable in cell division regulatory[73]. And p53 pathway contain hundreds of genes that response a wide range of stressing signals, involving cell cycle arrest, apoptosis or cellular senescence. Inducing apoptosis of neurons and neural progenitors, p53 plays an essential role during the development of central nervous system by controlling the cell number. Loss of p53 function facilitates the self-renewal of early neural progenitors.

Somatic mutations in the TP53 gene are the most common genetic changes found in human cancer, and alterations in genes which impact p53 functions also widely exist in most cancers. There are many factors that can cause the inactivation of p53, such as viral infection, loss of ARF, or overexpression of MDM2.

The p53 transcription factor can be activated in response to DNA damage, hypoxia, and oncogene activation. After post-translational modification by various genotoxic and cytotoxic stress-sensing agents, stabilized p53 function as a transcription factor regulating more than 2500 genes promoters. Among the genes stimulated by p53 are the p21 (cyclin-dependent kinase inhibitor1), MDM2, and many genes encoding proapoptotic proteins. Besides the target gene of p53, MDM2 also induce the p53 inactivation through inhibiting p53 transcription and catalyzing p53 ubiquitination. Actually, MDM2 is a key negative regulator of p53 during normal development and in tumorigenesis[74]. Another component in p53 pathway is MDM4, which inhibits p53 transcription and enhances the ubiquitin ligase activity of MDM2[75]. Importantly, CDKN2A (the gene encoding $p16^{INK4a}$) encode a second product: ARF protein ($p14^{ARF}$

in humans and p19[arf] in mice). ARF protein is a tumor suppressor that antagonize against MDM2 and stabilize p53[76]. And the expression of ARF is facilitated by CHD5[77]. Nevertheless, ARF function is not restricted to the p53 pathway. E2F1, the important component in RB pathway, can be inhibited by ARF. And MDM2 participated in the modulation of E2F1 activity by ARF[78]. In fact, evidence from many experiments supports that there are some connections between p53 pathway and RB pathway.

TP53 mutations are prevalent in glioblastomas, especially in the secondary glioblastoms. Loss of p53 caused by point mutations or chromosome 17p loss is a frequent and early event in the pathological progression of secondary GBM[79]. Amplification of chromosome harboring MDM2 gene is found in sporadic primary GBM samples[80]. Amplification of MDM4 is also found in GBM[81]. There are also some GBM samples containing loss of chromosome 1p, where the CHD5 gene is located. The inappropriate expression of p21 (the inhibitor of CDK2 encoded by gene CDNK1A) caused by p53 functional inactivity is also found in glioma, although there is no genomic alteration in CDNK1A gene. A study shows the evidence that p53 loss might cooperate in tumorigenesis by impairing neural stem cells differentiation potential[82].

2.4.3 The PI3K pathway

Phosphoinositide 3-kinases (PI3Ks) are a family of proteins regulating cell growth, metabolism, proliferation, glucose homeostasis and vesicle trafficking. There are 3 members in PI3K family: class-I, -II, and -III PI3K, and class-I PI3K is the most extensively studied member[83]. PI3Ks can phosphorylate the proteins with pleckstrin homology (PH) and PH-like domains, and those proteins are thus recruited to plasma membrane and transmit signals. Among those proteins, Protein Kinase B (AKT or PKB) is the best-characterized one. AKT phosphorylates about 100 substrates thereby modulating a wide range of cellular functions[84]. For example, AKT activates cell proliferation and exerts a strong anti-apoptotic effect by phosphorylation various proteins. AKT is also essential in forming genetically modified neural progenitors for GBM[85]. In addition, AKT also regulates a set of proteins implicated in growth, metabolism and angiogenesis. Importantly, AKT expedites the activation of mTORC1 pathway through the phosphorylation of TSC2 (tuberin) and PRAS40 (proline-rich AKT substrate of 40 kDa)[86]. Since mTORC1 is a complex regulating protein translation and ribosome biogenesis, activated AKT promotes the production of ribosomes and proteins. In fact, the PI3K/AKT/mTOR pathway plays an important role in apoptosis and hence cancer and longevity[87].

Downstream components of PI3K pathway confer strong feedback controls[88]. For example, feedback signaling can activate AKT and lead to a poor result in the treatments of cancers[89]. Hyperactivation of mTORC1 will result in the repression of PDGFRA and -B transcription, which impacts not only in PDGF signaling to AKT but also on the proper transmission of the signal from other growth factor receptors[90].
And many studies have shown that inhibiting mTORC1 will activate PI3K. Besides, transcriptional repression and inhibitory phosphorylation of IRS-1 by downstream elements also cause feedback inhibition of PI3K[91].

According to considerable works, PI3K pathway is a central integrator of metabolism and survival/growth signals, and aberrances of many members in PI3K pathway are

associated with development of malignancies. Firstly, mutations in PTEN, which is the upstream negative regulator of PI3K pathway, is commonly found in many cancers including GBM[92]. Secondly, PI3K class IA, which is on the top of this pathway, is mutated and amplified in a variety of cancers. And mutant PIK3R1 has been found in GBM[93]. Furthermore, amplification of AKT genes exists widely in human cancers, and a point mutation of AKT1 in several cancer patients is reported[94].

2.4.4 The RAS/MAPK pathway

Activated RAS proteins implicate in cellular signal transduction, which is crucial in many cellular processes, including proliferation, migration, differentiation and apoptosis. RAS is also one of the most common oncogene in human cancer[95].

RAS/MAPK pathway can be switched on by integrins. Integrins are transmembrane heterodimer receptors that mediate the interaction between the ECM (Extracellular matrix) and the cytoskeleton. Usually integrins create connections between ECM and cytoskeleton actin filaments by binding cytoplasmic anchor proteins and creating focal adhesion complex[96]. Those focal adhesion complexes will facilitate the cross-phosphorylation and activation of FAK (focal adhesion kinase). Upon activated, FAK prompts a signal transduction cascade recruiting Grb2 (the adaptor protein) and SOS (the RAS guanine nucleotide exchange factor) to phospho-FAK at the plasma membrane, and finally RAS is activated. Activated RAS (RAS-GTP) then activates serine/threonine kinase RAF, and RAF phosphorylates mitogen-activated protein kinase kinase (MEK), which in turn phosphorylates MAPK (mitogen-activated protein kinase)[97]. The activation of MAPK lead to the phosphorylation of many nuclear transcription factors that induce the expression of genes promoting cell cycle progression. Markedly, RAS can also activate PI3K.

Besides integrins, activated RTKs can also induce MAPK pathway. Particularly, activated RTK expedites receptor dimerization and cross-phosphorylation, which create binding sites for adaptor protein complexes such as Grb2/SOS. And RAS can be activated.

Some mutations of RAS gene family will generate permanently activated RAS proteins, which result in unintended and overactive signaling inside the cell and is associated with some cancers[98]. Although RAS mutation is very rare in GBMs, increased RAS pathway activity is detected in almost all GBMs. Therefore, it is also possible that upstream factors elevate RAS activity. For example, the integrin α3β1 which regulates glioma cell migration is consistently over-expressed in gliomas[99]. And RTKs like EGFR and PDGFR are highly activated in many GBMs. Moreover, in a mouse model, combined activation of RAS and AKT in neural progenitors induces GBM formation[100].

2.5 Isocitrate Dehydrogenase (IDH) and GBM

Isocitrate Dehydrogenase (IDH) is an enzyme that catalyzes the oxidative decarboxylation of isocitrate into alpha-ketoglutarate (αKG or 2-OG), which produces NADPH (or NADH) and is involved in tricarboxylic acid (TCA) cycle. Since the reaction catalyzed by IDH is reversible, IDH participates in the reaction forming isocritrate through reductive carboxylation of αKG. The isocritrate produced from this reaction will be further metabolized to facilitate lipid biosynthesis[101].

Isocitrate dehydrogenase 1 and 2 (IDH1 and IDH2) are two isoforms of isocitrate dehydrogenase. IDH1 localizes to the cytosol and peroxisomes, whereas IDH2 exists in mitochondria. IDH1 and IDH2 are $NADP^+$-dependent enzymes. And there is also an IDH enzyme called IDH3, which use NAD^+ as cofactor. The different IDH isoforms have overlapping function in cellular metabolism, but not redundant[102].

Heterozygous mutations in the gene encoding isocitrate dehydrogenases (IDH1 or IDH2) are frequently observed in gliomas and some other tumors[103], and such mutations are considered to be associated with the tumor formation in GBM. Mutations in IDH1 are detected in gliomas at a higher frequency than mutations in IDH2. The inverse correlation between mutations in IDH1and IDH2 suggests that they affect a similar pathway. According to previous studies, mutations in IDH1 are consistently found in codon 132 for arginine (R132), and mutations in IDH2 are largely confined to the analogous amino acid R172[104]. All the presently identified hotspot mutations are single-nucleotide substitutions in the respective arginine codons[105].

Mutations (usually the mutations are heterozygous) makes IDH reduce affinity for its substrates and thus lose activity to convert isocitrate into αKG. However, mutant IDH gains the ability to reduce αKG to D-2-hydroxyglutarate (D2HG or R-2-hydroxyglutarate) in an NADPH-dependent manner and results in accumulation of 2-hydroxyglutarate[106]. In fact, the level of αKG is slightly lower in IDH mutant gliomas, though this decrease was not statistically significant. And the tumor-derived mutant IDH dominantly inhibits the wild-type IDH[107]. Some in vitro experiments show that the mutant IDH promotes the proliferation and blocks differentiation of cells[108]. But some studies reveal that mutant IDH has very limited capability to promote the proliferation and inhibit differentiations in vivo independently.

Except an important TCA cycle intermediate, alpha-ketoglutarate (αKG) is also an essential cofactor for many enzymes, including Jumonji domain-containing histone demethylases, TET 5-methylcytosine hydroxylases, and EglN prolyl-4-hydroxylases. EglN prolyl-4-hydroxylases is the enzyme that tag HIF transcription factor for polyubiquitylation and proteasomal degradation.

D-2-hydroxyglutarate (D2HG) is an oncometabolite[109] having a connection with the increased risk for glioma. No physiologic functions of D2HG are found in normal metabolism. Usually, D2HG exists in normal cells at a very low level. The level of D2HG in IDH mutant tumors can be extremely increased. D2HG is structurally and chemically similar with αKG. Competing directly with αKG, D2HG inhibits αKG -dependent dioxygenases and thus hinders DNA demethylation, resulting in hypermethylation of CpG dinucleotides[110]. Besides, D2HG also inhibits αKG -dependent oxygenases and causes an increasing of histone methylation. And the histone demethylation is required for lineage-specific progenitor cells to differentiate into terminally differentiated cells. Such epigenetic alterations impact on differentiation and gene transcription, and contribute to the formation of tumors.

Furthermore, mutant IDH contributes to tumorigenesis by up-regulating HIF-1α target gene transcription. Hypoxia-inducible factor 1-alpha (HIF-1α) plays an important role in the transcriptional activation of genes involved in glucose metabolism, angiogenesis, and other crucial aspects of cancer biology. HIF-1α confers tumor cells a growth advantage by regulating the hypoxic response pathways. It has been found that HIF-1α

is associated with increased patient mortality in several cancer types[111]. However, a study on astrocytomas suggests that HIF-1α can be oncoprotein or tumor suppressor, depending on the extant microenvironment of the tumor[112]. And some researchers assert that HIF elevation in IDH mutant tumors is usually confined to areas of necrosis and presumed severe hypoxia. They also suggest that IDH mutation is not the only reason for the activation of HIF-1α pathway in gliomas[113].

As what mentioned before, the level of HIF-1α is increased in IDH mutant tumors compared with normal brains in an experiment of IDH1 R132H knock-in mice[114]. The accumulation of 2-hydroxyglutarte in IDH mutated tumors will inhibit the prolyl hydroxylase, which are αKG-dependent dioxygenases that hydroxylate HIF-1α for proteasomal degradation in the presence of oxygen. Therefore, the mutant IDH may stabilize HIF-1α and increase its steady-state levels through the decrease of enzyme activity and increased level of 2-hydroxyglutarte. However, several researches shows controversial results about the relationship between D2HG and HIF-1α. According to recent studies, it is the L-2-hydroxyglutarate (L2HG or S-2-hydroxyglutarate) enantiomer that inhibits HIF prolyl hydroxylases, while the D2HG enantiomer produced by mutated IDH stimulates their activity resulting in diminished HIF-1α levels[115]. In that experiment of knock-in mice, tissue hypoxia is also a reason for the elevated level of HIF-1α. Nonetheless, whether D2HG is sufficient to down-regulate HIF-1α remains obscure.

EglN1 is a member of αKG-dependent dioxygenases family, and it is the principal HIF prolyl-hydroxylase. As described above, HIF-1α will be degraded after tagged by EglN under normoxic conditions. When the activity of EglN1 is inhibited under hypoxic conditions, HIF-1α will have the chance to accumulate and activate the transcriptional response of cells to hypoxia. Consistent with the discovery about D2HG potentiating EglN1 activity, IDH mutant brain tumors show decreased HIF activation relative to their wild-type counterparts. Some researches deem that EglN plays a causal role in transformation of astrocytes by mutant IDH in cell culture models. Actually, some EglN inhibitors have been developed for the treatment of anemia and tissue ischemia[116], which may inspire the researches of novel treatments for IDH mutant tumors.

TET1, TET2 and TET3 are from a family of αKG-dependent DNA-modifying enzymes, and they catalyze reactions to convert 5-methylcytosine (5mC) to 5-hydroxymethycytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC)[117]. And the TET enzymes are crucial in epigenetic regulations of gene expression[118]. TET2 is considered to be a relevant target of D2HG in IDH mutant tumors. An in vitro experiment demonstrates that the catalytic activity of TET2 is potently inhibited by D2HG[110]. Although the connections between TET2 and IDH in glioma are still unclear, researchers believe that loss of TET2 activity is an important and frequent pathogenic event in brain tumors[119].

Another potential relevant target of D2HG is the Jumonji domain-containing (JmjC) family of histone lysine demethylases, which have crucial functions in regulating gene expression through mediating histone methylation and demethylation. This enzyme family has been connected with the pathogenesis of many cancers: some JmjC histone demethylases are considered to function as tumor suppressors but some promote tumor growth. KDM6B (JMJD3), a member of JmjC histone demethylases family, is reported to promote the terminal differentiation of glioblastoma cells[120]. And KDM6B is also

associated the regulation of p53. It has been observed that D2HG can inhibit many JmjC histone demethylases, and may thus contributes to transformation of mutant IDH expressing cells[121]. Some researchers speculate that the important alteration of histone methylation induced by D2HG that promote tumor formation are only at specific genetic loci. Although it is unknown whether D2HG affects histone methylation in primary IDH mutant tumors, different functions of the JmjC histone
demethylases in different tissues might explain the tissue specificity of IDH mutations in cancers.

Except what mentioned above, there are many enzymes are found to be inhibited by D2HG in IDH mutant tumors, including cytochrome c oxidase (complex IV), ATP synthase (complex V)[122] and PLOD family of lysyl-5-hydroxylases. Understanding the associations between those enzymes and tumors will help to understand how IDH mutant contribute to tumor formation and development. And it will also provide valuable information for the design of specific therapeutics.

IDH1 and IDH2 also have a function producing NADPH, which is the principal cellular and mitochondrial antioxidant preventing cells from oxidative stress and radiation damage[123]. IDH mutations will result in a lower concentration of NADPH, and thus disrupt the ability of cells for reductive processes in defense against reactive oxygen species. NADPH has also some essential regulatory functions in cells. The disruption of NADPH levels will have profoundly effects on the IDH mutant cells.

Most mutations of IDH occur frequently in grades II–III gliomas and secondary glioblastomas[124]. Spontaneous IDH mutations are thought as strong prognostic indicators in secondary glioblastomas[125]. And point mutations in IDH is one of the major feature of Proneural GBM (most known secondary GBMs were classified as Proneural). Nevertheless, such mutations are very rare in primary GBMs and pediatric GBMs[126], and none of the brain tumors of nonglial subtypes are found as harboring IDH mutations[127]. And the mutations of IDH1 and IDH2 seem to be mutually exclusive in brain tumors, because no cases are reported containing both IDH1 and IDH2 mutations[128].

IDH mutant tumors tend to be relatively indolent and are associated with increased overall survival and younger age, which is consistent with the fact that HIF-1α can work as both oncoprotein or tumor suppressor. However, it is obscure whether the difference is driven by IDH mutational status of tumors, or is just a reflection of other biological distinctions. IDH mutations are also found more commonly in tumors with TP53 mutations. Besides, IDH mutations have relation with a distinct DNA methylation phenotype and an altered metabolic profile in glioma[129]. IDH mutant brain tumors frequently displays a global DNA hypermethylation signatures. A study on glioma methylation patterns demonstrates that the ratio of hyper- to hypomethylated CpG loci in IDH mutant tumors is much higher than that of IDH wild-type tumors[130]. And IDH mutant tumor samples are more highly methylated than other samples without such mutations. Although IDH mutation is heterozygous, the methylation profile of IDH mutant tumors is generally homogenous. This study also concludes that IDH mutation is more robustly associated with methylation class compared with other classical glioma tumor genetic markers.

Mutations in IDH also lead to the hypermethylation of several cellular signaling pathways and the hypomethylation of some metabolism and biosynthesis pathways[131]. One possible explanation is that the relatively lower level of αKG and NADPH in IDH mutant tumor could drive the selection of cells with compensatory metabolic gene expression profiles. And the alterations in those cells are regulated by epigenetic factors such as methylation and chromatin configuration.

Taken together, IDH mutations play a crucial role in gliomagenesis and they are associated with a distinct phenotype. Mutant IDH is regarded as an oncogene. The genetic and epigenetic alterations in IDH mutant gliomas are not independent. And researching on the association between IDH mutations and phenotypes in glioma can provide profound implications for the development of diagnoses and therapies. However, there are still many facts about IDH and glioma remaining unknown. The relationship between IDH mutation and glioma tumor progression is much more complex than what the recent studies have shown.

2.6 Gene signatures

A gene signature is defined as a group of genes, whose combined gene expression alteration (or pattern) can be regarded as a unique characteristic of a medical or other condition[132]. In other words, the expression of the group genes are significantly associated with a certain condition. Ideally, a gene signature is predetermined and should have a specificity in terms of diagnosis, prognosis or prediction of therapeutic response. And such a specificity should be validated in independent groups of samples.

Identifying gene signatures can be applied to a wide range of biological and medical fields: from understanding tumor formation to initiating novel treatments. Microarray technology has become the most widely used method to find gene signatures. After decades of development, many gene signatures about tumors are identified. For example, a gene signature is found to be strongly associated with the clinical feature of breast cancer[133], and classification of human acute leukemias can be achieved based on gene signatures[134]. A variety of analysis methods for identifying gene signaures have been proposed, such as “bottom up”, “top-down” supervised approaches, and gene candidate approach[135]. In addition, gene signature models are developed based on known pathways or other information.

2.7 Introduction of statistics methods[1]

2.7.1 Statistics and parameters

A statistics represents a characteristic of samples, and it can be divided as descriptive statistics and inferential statistics. The main purpose of descriptive statistics is to demonstrate and summarize the data from samples. However, inferential statistics are used to make inference or prediction from data. Statistics are widely used in hypotheses testing.

One the other hand, a parameter refers to a characteristic of a population. And the inferential statistics are usually used to infer a parameter of the population from which samples are drawn.

[1] Most of the formulas in this section is from the paper “Resampling-based multiple testing for microarray data analysis”

Normally an appropriate test statistics should be determined before doing the hypotheses tests, and how to choose statistics depends on the details of the particular experiment. Only depending on samples, statistics collect and represent the information from a certain aspect of samples. Assuming that there are a series hypotheses to be tested, let T stand for the test statistic, and T_i means the statistic for each null hypothesis H_i. After choosing the critical value c_α, a null hypothesis can be rejected when $|T_i| > c_\alpha$, where α means confidence level indicating that the probability of rejecting a true null hypothesis is equal or less than α .

2.7.2 Hypothesis testing

Based on sample data, hypothesis testing is a procedure to identify whether there is enough evidence supporting a hypothesis with respect to a parameter. In most cases, hypothesis testing consists of 2 opposite hypotheses: null hypothesis (H_0) and alternative hypothesis (H_1). Null hypothesis is a statement about no effects or no difference. And the alternative hypothesis is in favor of that an effect or a difference does exist. Since most researches focus on the influence of drugs or conditions on samples, H_0 is generally expected to be rejected. And H_1 can be directional or non-directional (or two-tailed)[136]. Non-directional hypothesis does not make inference in a particular direction, while directional hypothesis indicates a directional relationship between groups. Although there are 3 possible alternative hypotheses, researchers can only select one as H_1.

For a typical gene expression data analysis, one null hypothesis is usually like: the gene does not express differently between the 2 conditions/phenotypes.

Upon collecting data from samples, hypotheses are evaluated with appropriate inferential statistical tests. A test statistic will be generated, which is a function of the sample. The sampling distribution of test statistics under the null hypothesis is calculable, so that the test statistic can be compared with critical values and the p-value can be gained. The critical values in a test are highly unlikely to occur if the null hypothesis is true. And the comparison between test statistic and critical values indicates the statistical significance, which means whether the observed difference occur by chance or is due to a genuine experimental effect.

A sampling distribution contains all the possible values the test statistic can be assumed, if infinite number of studies with the same size of sample as the study were conducted. Based on the sampling distribution, one can declare whether the observed difference between sample groups is due to chance. And p-value can be calculated, which refers to the probability of obtaining a test statistic ($T = t$) result at least as extreme as the one that was actually observed, under the assumption that H_0 is true.
The p-value is an alternative way to evaluate the hypotheses. The smaller the p-value is, the more convincing the evidence is in favor of the alternative hypothesis. One can set the criteria to decide whether to reject H_0, and the criteria is referred to as significance level for a test. The significance level is typically set at 5% or 1%: if the p-value is less than 5% or 1% assuming H_0 is true, this null hypothesis can be reject due to the unlikelihood of getting such a result by chance.

The confidence interval (CI) is a type of estimated range of values for an unknown population parameter. CI is used to measure the reliability of an estimate.

Correspondingly, confidence level indicates the frequency of the observed interval contains the parameter. In hypothesis testing, confidence level is complement of significant level, i.e. 95% confidence interval means to reject any value of H_0 that is outside the interval at a 5% significance level.

In the other hand, the set of values for the test statistic that leads to rejection of null hypotheses is called rejection region, while acceptance region is defined as the set of values for the test statistic which are consistent with the null hypotheses. Obviously, rejection region and acceptance region are complementary. In many analysis, rejection region is the only set needed to be determined. And rejection region at level α can be denoted as Γ_α . For a particular confidence level α, p-value is the minimum type I error rate over all possible rejection regions Γ_α containing the observed value $T = t$, i.e.,

$$p(t) = \min \Pr(T \in \Gamma_\alpha | H \; is \; true); t \in \Gamma_\alpha$$

When $p < \alpha$, rejecting null hypotheses provides control of the type I error rate at level α. Apparently, a smaller p gives stronger evidence to reject the null hypothesis. The p-values without adjustments are called raw p-value.

Usually, selections about the type of alternative hypothesis and the level statistical significance should be determined before conducting the experiments. And appropriate statistical testing procedures should be selected according to the characteristics of the data from studies.

2.7.3 Type I and Type II errors in hypothesis testing

In hypothesis testing, errors committed by researchers can be divided into Type I and Type II errors. A Type I error (or a false positive) occurs when a true null hypothesis is rejected. The likelihood of making a Type I error depends on the significance level. For instance, the possibility to commit a Type I error (usually represented by α) at a 5% significance level is 5%. A Type II error (or a false negative) is to retain a false null hypothesis.

Assume that among the m null hypotheses testing simultaneously, the number of true null hypotheses is m_0 . Let R denote the number of rejected hypotheses. Then a summary table of multiple testing problem can be built (table.2.1). Among the true null hypotheses, there are V hypotheses which are rejected (Type I errors) and U hypotheses which are not rejected. Similarly, for the $m - m_0$ non-true null hypotheses, S hypotheses are declared as significant, while T hypotheses are declared non-significant (Type II errors). The number R is observable random variables, while number V, U, S and T are unobservable random variables.

Table. 2.1 Summary table of multiple testing problem

	Not Rejected	Rejected	Total
True null hypotheses	U	V	m_0
Non-true null hypothese	T	S	$m - m_0$
Total	$m - R$	R	m

Therefore, two sets M_0 $\{i: H_i \; is \; true\}$ and M_1 $\{i: H_i \; is \; false\}$ can be defined, and obviously, $|M_0| = m_0$ and $|M_1| = m - m_0$. Therefore, the true hypotheses can be

written as $H_{M_0} = \bigcap_{i \in M_0}\{H_i \; is \; true\}$, and the total null hypotheses in the analysis is $H_M = \bigcap_{i=1}^{m}\{H_i \; is \; true\}$.

The likelihood of committing Type II errors (denoted as β) is inversely related to the likelihood of committing a Type I error. Namely, there is a trade-off between them. In hypothesis testing, the likelihood of making Type I error is the one to be controlled at a certain level. Because Type I errors are regarded more harmful, and Type II errors are not "really errors". When the testing result is not significant, it indicates that the evidence is not strong enough to support that H_0 is false. Therefore, lacking of significance does not mean H_0 is true.

2.7.4 Multiple testing problem

In large-scaled studies (for example, the microarray gene expression experiments), numerous hypotheses are tested simultaneously and incorrect rejections of H_0 are more likely to occur if no measures for control are taken[137]. To be specific, let m and α denote the number of hypotheses to be test and significant level respectively. And let a represent the false H_0 among all the n null hypotheses. Then there will be $\alpha * (m - a)$ hypotheses being incorrectly rejected by chance. Since the false H_0 usually forms a very small part of the overall hypotheses, the number of the false positive hypotheses approximately equals to $\alpha * m$, which is not a small amount. In other words, the probability of generating at least 1 false positive is $1 - (1 - \alpha)^m$. Obviously, such probability is too large to be accepted. And when the number of hypotheses (m) is very large, the results will be quite misleading. That is the multiple testing problem typically existing in many microarray experiments.

Several methods have been developed to address the multiple testing problem, and most of them try to control the FWER (family-wise error rate) and FDR (false discovery rate) through adjusting p-value and confidence level for each individual test.

FWER is the probability making at least one Type I error (or incorrectly rejecting the H_0) over the whole family of tests. And FDR is the expected proportion of falsely rejected H_0 among all rejected H_0.

FWER describes the likelihood of committing any error among all the hypotheses tests, while FDR reveals to what fraction of the rejected H_0 are, on average, really true. Bonferroni correction and Holm–Bonferroni method are the single-step procedures attempting to control FWER, whereas Benjamini–Hochberg procedure and Benjamini–Hochberg–Yekutieli procedure control FDR[138]. Westfall and Young permutation is an approach controlling FWER used when test statistics are strongly dependent. Methods controlling FWER are much more conservative. One need to select the controlling methods based on the research condition and goal.

2.7.5 Methods for controlling FWER in microarray data analysis

As discussed before, the Family-wise error rate (FWER) is defined as the probability of making one or more type I error, i.e.,

$$FWER = \Pr(V > 0)$$

Controlling the FWER under the true null hypotheses $H_{M_0} = \bigcap_{i \in M_0}\{H_i \; is \; true\}$ is called exact control.

$$FWER = \Pr\left(V > 0 \middle| H_{\mathrm{M}_0}\right)$$

However, in many cases the number of true null hypotheses is unknown, weak control is adopted to control the FWER under all the null hypotheses $H_{\mathrm{M}} = \bigcap_{i=1}^{m}\{H_i\ is\ true\}$.

$$FWER = \Pr(V > 0 | H_{\mathrm{M}})$$

Practically, there are some disadvantage of the weak control. Thus, Strong control is widely used in microarray experiments analysis by controlling every possible choice of the set of M_0. For FWER, strong control means to control of $max_{\mathrm{M}_{0\subseteq\{1,\ldots m\}}}\Pr\left(V > 0 \middle| H_{\mathrm{M}_0}\right)$.

In other words, if the strong control for FWER is at level α, $FWER \leq \alpha$ regardless of which or how many nulls in the family are true.

The adjusted p-value for an individual hypothesis ($\tilde{p}_i$) is defined as the smallest Type I error rate level α at which one would reject it, given the values of all test statistics involved. For instance, if the FWER is controlled, the adjust p-value for hypothesis H_i is:

$$\tilde{p}_i = \inf\{\alpha: H_i \text{ is rejected at FWER} = \alpha\}$$

In details, the way of adjusting p –values can be divided as 3 classes: single-step, step-down and step-up procedures.

For the single-step procedures, equivalent multiplicity adjustments are applied to all hypotheses, without taking consideration of the ordering of the test statistics or raw p-values. Bonferroni procedure is a widely used stingel-step method, and it set the confidence level of individual test as α/m. One null hypothesis will be rejected when the corresponding p-value is less than α/m. Thus the Bonferroni single-step adjusted p-values is defined as:

$$\tilde{p}_i = \min(mp_i, 1)$$

When the number of total hypotheses (m) is very large, it becomes extremely hard to reject one hypothesis. Therefore, Bonferroni procedure is regarded as somewhat conservative if there are a large number of tests. Furthermore, groups of genes may have highly correlated because of the co-regulation in many microarray experiments, and it makes the test statistics are positively correlated. Bonferroni procedure is not suitable in such cases because of its regardless of the correlations. To take into account the correlations between test statistics, Westfall & Young (1993) proposed two kind of single-step adjustments approaches: minP and maxT. Let P_l represent the random variable for the raw p-value of the lth hypothesis, and let H_{M} represent the set of all null hypotheses to be tested, the single-step minP adjusted p-values can be written as:

$$\tilde{p}_i = \Pr\left(\min_{1\leq l\leq m} P_l < p_i \middle| H_{\mathrm{M}}\right)$$

Employ T_l to denote the random variable for the test statistics of the lth hypothesis, and let t_i denote the realization of the random variable T_i, then the single-step maxT adjusted p-values can be written as:

$$\tilde{p}_i = \Pr\left(\max_{1\leq l\leq m} |T_l| \geq |t_i| \middle| H_{\mathrm{M}}\right)$$

Apparently, single-step minP and maxT procedures assume all the m null hypotheses are true, so the adjusted p-values control FWER in a weak way. To give strong control of FWER, data is assumed to have a property called "subset pivotality"[139]. For all the subsets K of $\{1, \dots m\}$, the joint distribution of a sub-vector of raw p-values $\{P_i: i \in \mathrm{K}\}$ are identical under the restrictions $H_{\mathrm{K}} = \bigcap_{i\in \mathrm{K}}\{ H_i \; is \; true\}$ and $H_{\mathrm{M}} = \bigcap_{i=1}^{m}\{H_i \; is \; true\}$. Subset pivotality is important, because it ensure the adjusted p-values calculated under all null hypotheses to provide strong control of FWER. Moreover, the following resampling can be done under all null hypotheses H_{M} conveniently (in most cases the set of true null hypotheses H_{M_0} is unknown).

For the gene expression values matrix, the data always have the subset pivotality property. Consider a subset $\mathrm{K} = \{i_1, i_2, \dots, i_\kappa\}$, and its complement $\{j_1, j_2, \dots, j_{m-\kappa}\}$. For a certain gene i, the test statistics T_i is computed solely from the data of this gene (the ith row of the matrix), so the joint distribution $(T_{i_1}, T_{i_2}, \dots T_{i_\kappa})$ is independent with the hypotheses $(H_{j_1}, H_{j_2}, \dots, H_{j_{m-\kappa}})$ if $(H_{i_1}, H_i, \dots, H_{i_\kappa})$ have same specification.

Compared with single-step methods, step-down procedures provide a less conservative but more powerful strong control of FWER or other error rates. Holm method is a step-down procedure improved from Bonferroni method. Firstly, it ranks all the m raw p-values in ascending order. Let p_{r_i} represent the ith ordered raw p-value, the ranking can be written as:

$$p_{r_1} \le p_{r_2} \le \cdots \le p_{r_m}$$

If the first hypothesis can be rejected at level α/m, set the level for rejecting the second hypothesis as $\alpha/(m-1)$. By analogy, if the first $(i-1)$ hypotheses can be rejected, set the significant level as $\alpha/(m-i+1)$for the ith hypotheses. The step is repeated until the ith hypothesis cannot be rejected. Therefore, The Holm step-down adjusted p-values can be given as:

$$\tilde{p}_{r_i} = \max_{k=1,2,\dots i}\left\{\min\left((m-k+1)p_{r_k}, 1\right)\right\}$$

If the data are assumed to have subset pivotality, the step-down minP and step-down maxT procedure can provide strong control of FWER under all null hypotheses. And they are less conservative than the Holm method.

The step-down minP adjusted p-values proposed by Westfall & Young (1993) is written as:

$$\tilde{p}_{r_i} = \max_{k=1,2,\dots i}\left\{\Pr\left(\min_{l=k,\dots,m} P_{r_l} \le p_{r_k} \middle| H_{\mathrm{M}}\right)\right\}$$

If the tests statistics is ranked as descending order $(|t_{s_1}| \ge |t_{s_2}| \ge \cdots \ge |t_{s_m}|)$, the step-down maxT adjusted p-values can be written as follow (larger test statistics suggest alternative hypothesis):

$$\tilde{p}_{s_i} = \max_{k=1,2,\dots i}\left\{\Pr\left(\max_{l=k,\dots,m} |T_{s_l}| \ge |t_{s_k}| \middle| H_{\mathrm{M}}\right)\right\}$$

2.7.6 Resampling methods for statistical testing

To get more reliable results, resampling method is used during the analysis. Usually the joint (and marginal) distribution of the test statistics is unknown. To avoid the parametric assumptions about the joint distribution of the test statistics, resampling

methods are adopted. One of the distinctive features of resampling methods is that the observed data itself constructs the relevant sampling distribution, and the scientists are restricted by any assumptions about the distribution of the underlying population. In each time of resampling, the observed variables' values are re-assigned randomly to different sample groups and the test statistics are re-computed. And a resampling distribution will be obtained after thousands of resampling. Then the original test statistics can be compared with the resampling distribution, which gives evidence about whether the corresponding hypotheses should be rejected.

The resampling p-value is calculated based on the test statistics distribution created by the resampling process. Especially, the resampling p-value is the proportion of resampled data sets yielding a test statistics as extreme as the original test statistics. Therefore, the results yielded from resampling methods are not based on any assumptions regarding an underlying population distribution[140].

Since the results depend exclusively on the observed data, resampling methods make the tests more robust by incorporating the distribution characteristics. According to Westfall & Young (1993), resampling methods "encompass many existing parametric multiple testing methods as special cases, and provide multiple testing solutions in situations where alternative methods are unavailable".

2.7.7 Chi-Square Test and Fisher Exact Test[2]

Usually, categorical data can be summarized by a form of an $r * c$ table, which consists of r rows and c columns. Such a table is referred to as contingency table, and it describes the frequency distribution of the variables. To be more specific, the data in each cell of contingency table is the number of observations that are categorized in the cell. Contingency table is very useful when identifying the dependence structure underlying the categorical variables with the help of hypothesis testing methods.

Chi-square test (χ^2-test) is used when the sampling distribution of the data in a contingency table is assumed as chi-square distribution. And the chi-squared distribution (χ^2-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. Generally, the problem chi-square test is employed to solve is that: in the underlying population(s) represented by the sample(s) in a contingency table, whether the observed cell frequencies are different from the expected frequencies or not.

The two main purposes of chi-square test is to test the homogeneity (whether the proportions of observations in a series populations are equal) and independence (the extend one variable influence another) of the variables. The computation processes for the two purposes are identical. The generic null and alternative hypothesis involve observed (o) and expected (ε) cell frequencies in the underlying population(s) represented by the sample(s).

The null hypothesis can be stated as "observed frequency of each of the $r * c$ cells is equal to the expected frequency of the cell":

$$H_0: o_{ij} = \varepsilon_{ij} \quad for\ all\ cells$$

[2] The table and formulas in this section are adapted from the book "Handbook of Parametric and Nonparametric Statistical Procedures, 2nd Edition"

where i and j represent the index of a cell in the contingency table.

The alternative hypothesis can thus be described as "observed frequency of at least one of the $r * c$ cells is not equal to the expected frequency of the cell":

$$H_0: o_{ij} \neq \varepsilon_{ij} \quad for \ \ at \ least \ one \ cell$$

For the contingency table, let the notation O_{ij} represents the number of observations in the cell that is in the ith row and the jth column. Similarly, $\mathrm{O}_{i.}$ denotes the number of observations in the ith row, and $\mathrm{O}_{.j}$ denotes the number of observations in the jth column. The notation n is used to represent the total number of samples.

The expected frequency or count of a cell (E_{ij}) can be calculated as:

$$\mathrm{E}_{ij} = \frac{(\mathrm{O}_{i.})(\mathrm{O}_{.j})}{n}$$

Therefore the test statistic (χ^2) for the chi-square test of $r * c$ tables is computed with equation:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left[\frac{(\mathrm{O}_{ij} - \mathrm{E}_{ij})^2}{\mathrm{E}_{ij}} \right]$$

The degree of freedom (df) for the relevant χ^2-distribution is:

$$df = (r-1)(c-1)$$

If the obtained test statistic is equal to or greater than the critical value at the particular level of significance, null hypothesis can be rejected.

However, the χ^2-distribution only provides an approximation of the exact sampling distribution for a contingency table. To be specific, the accuracy of the chi-square approximation increases as the size of the samples increases. And when the size of samples is small (less than 20) and the dimension of contingency table is $2 * 2$, Fisher exact test is recommended to be adopted instead of chi-square test.

For the $2 * 2$ contingency table, which is used in the pathway enrichment analysis, the χ^2-distribution is employed to approximate the hypergeometric distribution. And the genes mapped to a certain GO term might be small, which will lead to a small frequency or count in the corresponding cell of the contingency table. Consequently, Fisher exact test is selected to perform the analysis.

Fisher exact test is one of a class of exact tests, so called because the significance of the deviation from a null hypothesis can be calculated exactly, rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity, as chi-square test. The assumptions in Fisher exact test are identical as those mentioned in the chi-square test for $r * c$ tables. Besides, both the row and column sums of a $2 * 2$ contingency table are assumed to be predetermined in Fisher exact test. Nonetheless, this assumption is seldom met in practical.

In Fisher exact test, the p-values are computed based on hypergeometric distribution. The hypergeometric distribution is a discrete probability distribution. In such a model, there are two possible outcomes (to be designated Category 1 versus Category 2) in a

set of N trials. Sampling without replacement, the outcome of a trial will be dependent on the outcomes of previous trials. Assuming that there are N objects totally, and among them there are K objects in Category 1 (thus there are $(N-K)$ objects in Category 2). If n samples are drawn from the N objects without replacement, the probability $(P(X=k))$ of obtaining exactly k objects from Category 1 and $(n-k)$ objects from Category 2 is:

$$P(X=k) = f(k;N,K,n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Accordingly, if a one-tailed analysis is conducted, the probability (P) of obtaining a value equal to or more extreme than k in such experiments is:

$$P = 1 - \sum_{j=0}^{k-1} f(j;N,K,n)$$

Evidently, the smaller the probability is, the more impossible to observe such a result by chance. And if the probability is equal or less than the predetermined confidence level, the result can be regarded as statistically significant.

In illustrating the calculation of chi-square test and Fisher exact test, a $2*2$ contingency table is constructed by recording the data in the form of frequencies or counts (table.2.2).

Table.2.2 Contingency table for the calculation of p-values in fisher exact test

	Column1	Column2	Row Sums
Row1	a	b	a+b = n_1
Row2	c	d	c+d = n_2
Column Sums	a+c	b+d	a+b+c+d = n

The test statistic of chi-square test can be calculated quickly as:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

And the exact probability (P) of obtaining a specific set of observed frequencies for the 2*2 contingency table is:

$$P = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+c)!\,(b+d)!\,(a+b)!\,(c+d)!}{n!\,a!\,b!\,c!\,d!}$$

The null hypothesis of Fisher exact test can be stated as:“ In the underlying populations the samples represent, the proportion of observations in Row 1 that falls in cell a is equal to the proportion of observations in Row 2 that falls in cell c”.

The corresponding alternative hypothesis for one-sided test is:“In the underlying populations the samples represent, the proportion of observations in Row 1 that falls in cell a is greater or less than the proportion of observations in Row 2 that falls in cell c”. The alternative hypothesis (greater or less) should be consistent with the observed data. Note that the probabilities for any sets of observed frequencies that are even more extreme than the observed frequencies should also be taken into account.

2.7.8 Mann–Whitney U test[3]

Mann–Whitney U test (hereafter referred to as U-test), a nonparametric equivalent of Student's t-test, are selected to perform the analysis. For each hypothesis, expression data from all samples are arranged in ascending order, regardless which group the samples belong to.

For the gene expression analysis, consider that there are n samples in total. let x_{rj} denote the jth ranked gene expression value, so all the ranked data should be:

$$x_{r1} \leq x_{r2} \leq \cdots \leq x_{rn}$$

Then each expression value is assigned a rank. For example, the smallest expression value is assigned as 1, and the largest one will be assigned as rank n if there is no ties. (In some cases, the ranks can be reversed, which will not affect the final result of U-test.) Importantly, when there are two or more samples with equal expression values, the average of the ranks involved will be assigned to all the samples. However, the tie adjustment will not influence the sum and average of ranks for the two groups when the samples with ties are from the same group. After the adjustments of ranks, the sum of all ranks ($\sum R_i$) for each group can be calculated. Then the statistics, usually called "U", can be computed. There are two groups in this study, and U_1, U_2 denote the U values for each group.

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1$$
$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_2$$

where n_1 and n_2 are the numbers of samples in the two groups.

The smaller one of U_1 and U_2 will be the obtained statistics U. In order to reject the null hypothesis, the value of U must be equal or less than the critical value at the prespecified level of significance.

The normal distribution can be applied to approximate the Mann–Whitney U statistic if there are many samples. Especially, normal approximation should be considered when the sample size is larger than those documented in the exact table of the U distribution. The normal approximation of the Mann–Whitney U test statistic (represented as "z") is:

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

,where U is the smaller one of U_1 and U_2. And the absolute value of z must be equal to or greater than the critical value at a specific level of significance, so that the null hypothesis can be rejected. The result yielded by the normal approximation is usually consistent with the result obtained when the exact table for the Mann–Whitney U distribution is used.

[3] The formulas in this section are adapted from the book "Handbook of Parametric and Nonparametric Statistical Procedures, 2nd Edition"

2.7.9 The class imbalance problem

In many practical studies, imbalance data sets are generated and they weaken the power of most classifiers. An imbalance data set is highly skewed: most of the instances belong to one class (major class), whereas much fewer instances are labeled as the other class (minor class)[141]. And in most cases, the minor class is much more important. For example, in the study about rare diseases or genetic mutation in a population, samples with the diseases or mutation are very few. However, most classifiers tend to be biased towards the major classes and the minor classes are hence be ignored[142]. Several techniques have been proposed focusing on the class imbalance problem[143], which are beyond the scope of this thesis.

2.8 An outline of microarray

The accelerating availability of new technologies has transformed both the theory and practice of cancer research. Among those technologies microarray is an important and lost-cost one. Microarray technology is applied increasingly in biological and medical research to address a wide range of problems, such as quantification of gene expressions or the classification of tumors.

Usually, microarray is on a solid substrate (a glass slide) and contain an ordered series of sample. And the number of ordered samples can be hundreds of thousands on one slide. There are several alternatives of microarrays for differently researching purposes. Basically, the type of microarray depends on the sample (DNA, RNA, protein or tissue) placed on it. The most commonly used microarray is DNA microarray, which can be used for monitoring of expression levels in cells for thousands of genes simultaneously. In addition, DNA microarray is also capable of analyzing mutations, SNPs, or methylation states of genes in a sample.

According to the devices, microarrays can be classified as single-channeled and dual-channeled, where one or two samples might be hybridized simultaneously. In the single-channeled DNA microarray, absolute levels of gene expression is assayed, while in the dual-channeled counterpart relative levels of expression is evaluated.

Microarray technology presents an effective way to identify genes and pathways, which is valuable in many aspects such as finding potential drug targets, initiating novel therapy and genetic diagnosis.

2.8.1 Basic workflow of DNA microarray

First of all, a suitable microarray should be prepared or selected according to the research goals. Upon the acquisition of microarray, the RNA can be extracted and purified from samples. Then the RNA is converted to cDNA or cRNA, which are labeled with fluorescent reagents. The most commonly used dyes are Cy3 and Cy5. Fluorescently labeled cDNA or cRNA can be used in the hybridization process. The probes which are complementary to the molecules on the microarray hybridize with the strands on the microarray slide. Subsequently, microarray are washed and “read” by some commercially available scanners, so that the gene expression levels can be quantitated based on the amount of emitted fluorescence. A summary of the workflow is display in figure.2.1[144].

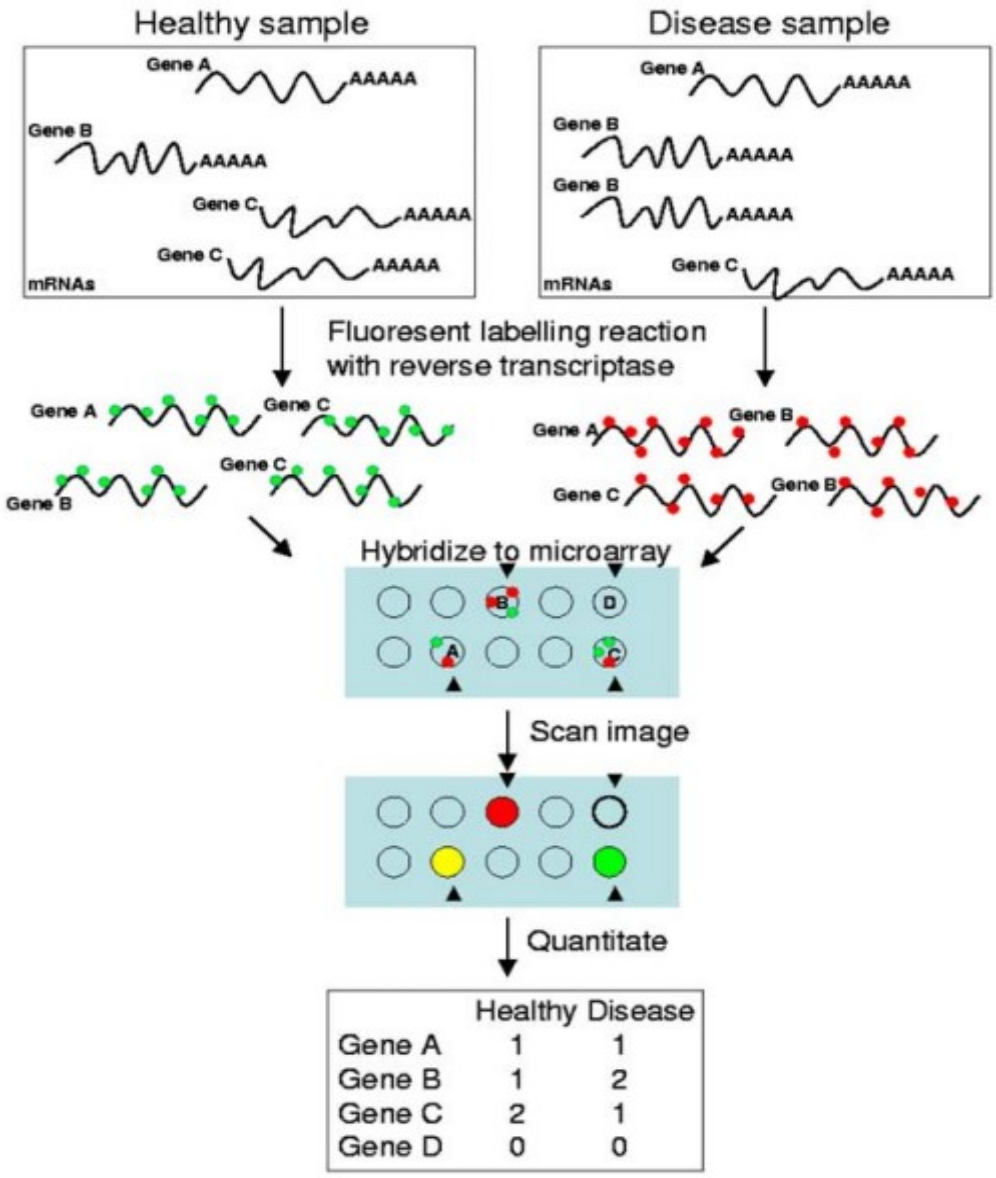

Figure 2.1 A typical workflow of expression microarray experiment

Importantly, making some replicates is strongly recommended for microarray experiments[145], and it can avoid many random errors. There are 2 types of replicates: biological replicates and technical replicates. Replicates can be designed based on the experiments, and samples can be also regarded as replicates in some cases. In addition, replicates facilitate to estimate the true expression of samples and potentially reduce the noise in data. However, bad replicates should be removed like other data with bad quality before any calculation.

2.8.2 Microarray data pre-processing and normalization

Data pre-processing is a necessary step in microarray experiments for the detailed analysis. Images produced by microarray experiments need to be parsed into numerical values to assay the intensities. The quality of images is significant for the following analysis. Good quality images should have high signal to noise ratio as well as a low background. There are various tools available for the quality control.

Missing values are common in experiment, which are defined as data whose intensity is below 0 or equals to 0, and they should be replaced by a certain estimated value (imputation) or just deleted.

Since the total brightness of a spot is composed by background brightness and labeled sample brightness, it is rational to make the spot intensities (or foreground intensities) be independent from background by subtracting it. Usually, background intensities should not vary multiplicatively with the spot intensities. And if such phenomenon occurs, there may be some problems in hybridization.

After removing the background, the relative expression ratio of genes can be calculated. The intensity ratio can be got easily using the formula:

$$E_i = \frac{R_i}{G_i}$$

where R_i and G_i is the median expression level of gene i (after background correction) in the red and green channel, respectively. However, the distribution of intensity ratio is highly skewed and asymmetric, because the ratios for up-regulated genes will range from 1 to infinity while the down-regulated genes will only have ratios between 0 and 1.

The log-transformation makes skewed distributions more symmetrical, so that the figure of variation becomes more realistic. In other words, it makes the variation of intensities more independent of absolute magnitude of intensity values. The most commonly used log-transformation is 2-based. However, log-transformation introduces systematic errors in the lower end of the expression value distribution.

Another data transformation method is fold change. If the intensity ratio is below 1, the value of fold change is inversed intensity ratio. But when the intensity ratio is higher than 1, fold change equals to intensity ratio. Obviously, fold change has some similar effect with log-transformation.

Data normalization is performed after transformation. The aim of normalization is to remove those systematic biases[146], but too strong normalization may make us miss the important biological variation from data. Hence, we need to find out the causes of systematic variations. Dye effect is the most common bias, and the scanner settings can also affect the measuring of intensity. As stated before, the replicate experiments may contain different sample variances due to differences in experimental conditions. And experimenter is one of the largest sources of systematic bias. Even though such systematic biases may be comparatively small, they may be confounding when searching for subtle biological differences. Importantly, bias introduced by the biological role of reporters or samples should not be normalized.

Normalization for microarray data also includes standardization and centralization. And the most widely used method is the log-transformation, which can make the data more normal-like and even out highly skewed distributions. There are a variety of methods for normalization, and the best choice of methods depends on the experimental design and results. For example, the linearity of data is a basis for choosing normalization methods. Linearity denotes that in the scatter plot of red channel versus green channel, the relationship between the channels is linear. When the data are linear, methods such

as median centering and scaling can be applied. For the non-linear data, lowess smoothing or other local method are preferred. Checking the linearity of the data also provides information about the reliability of the data, especially in the lower intensity range. Particularly, there are specific methods like RMA (robust multichip average) for the normalization of Affymetrix chips.

2.8.3 Microarray data analysis

After obtaining reliable data from microarray experiments, the biological information of interests will be derived from it.

One of the most common problem to be solved by microarray experiments is to find differently expressed (DE) genes. According to the expression levels from different types of samples, genes with statistically significantly different expression between the samples can be identified with some statistical methods. Apart from that, a ranked list about a genes based on "how distant is the expression level between different samples or conditions" can be obtained. In contrast, a set of genes can be pre-determined before experiments (for example, a group of genes involved in a certain pathway), and then they are checked whether to display significant differential expression as a whole in distinctive conditions.

There are many standard statistical tests available to find differently expressed genes, including t-test two-class comparisons, ANOVA (Analysis of variance) for multi-class comparisons, and Cox models for survival data. All of those methods are gene-wise, and the connections between genes might be omitted. To make use of the information of all the genes, hierarchical Bayes or empirical Bayes methods can be adopted. And the differential expression might be defined as a biologically meaningful way, so that "customized null hypotheses" can be used in statistical tests. The gene-by-gene approaches also generate multiple testing problem because thousands of hypotheses are tested simultaneously. If all the "rejected null" genes in the tests are regarded as differentially expressed, the study will end up with many false rejections. The most widely accepted method for the multiple testing adjustments is to control of the family wise error rate (FWER) or the false discovery rate (FDR). Statistical methods are also applied to answer questions like "whether members of a gene set are enriched in the differently expressed genes."

Furthermore, genes or samples can be classified as distinctive groups based on their expression patterns in the microarray experiments, which is usually accomplished by clustering. Besides, other machine learning methods like SOM (self-organizing map), SVM (support vector machines), PCA (principal component analysis) and DLDA (diagonal linear discriminant analysis) are used in the classification tasks of microarray data[147]. To represents the results in a lucid and concise fashion, visualization technology is required. Moreover, the biological interpretation from microarray data can be achieve through gene annotation. The relevant methods for data analysis are discussed in a separated sections.

2.9 Databases for genome annotation and biological pathways

2.9.1 Genome annotation

Usually, a list of genes or gene products that are supposed as being "interesting" is obtain after data analysis. And gene annotations provide more biological information about the experiment results, which is strongly desired.

The aim of genome annotation is to identify elements from genome and attach biological information for those elements. Basically, annotations can be either experimentally proven or computational predictions.

The hypothesis for the biological interpretation is generally like "If some cellular function is activated during the experiment, then several genes involved in this function should follow a similar pattern of activation." Hence it is expected to find annotations related to a certain function/process/pathway enriched in some cluster.

Fortunately, there are some databases about that available (for example, TOPSAN, ChemProt, DAnCER, TAIR, OryGenesDB, MGED and KEGG), which enable us to obtain annotation information more efficiently and easily. One of the difficulties of annotation work is that it is hard to collect all the kinds of phenotypes of samples while collecting their genomes. Presently, researchers are contriving to make a great amount of new annotation and correct some annotations, indicating that annotation work is relatively behindhand; and there are some mistakes in the existing annotations in databases. One reason is that those isoform transcriptons with low abundance did not get enough attention at the early stage of research. And earlier large-scale transcript sequencing projects emphasized on protein-coding genes, so there must be many unknown elements in other regions.

For microarray experiments, annotations facilitate researchers to fetch more information about the genes of interests. Often the first step is to locate an identifier for the genes (or probes), and identifiers often come from databases. Sometimes identifiers will change during the updating of databases, and genes may have different identifiers in different databases. Bioconductor provides many tools to solve such problem. For example, the package "org.Hs.eg.db" is for the genome wide annotation for human, primarily based on mapping using Entrez Gene identifiers. Some packages like "hgu95av2.db" support the mapping of identifiers from different kinds of microarray platform. There is also a list of annotation packages for different identifiers on the website of Bioconductor. Even if no identifiers are available, homology searching (BLAST) can assist us to identify those genes.

2.9.2 Gene Ontology

The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases[148]. Many gene products or biological elements are described and conceptualized diversely in different databases, which inhibits effective searching by both computers and people. Moreover, it has been found that there is a high level of sequence and functional conservation in many eukaryotes. It is possible to transfer the biological annotations from the experimentally tractable model organisms to the less tractable organisms based on gene and protein sequence similarity. So a computational system is required to compare and transfer the annotations among different species automatically or manually. Since much of the knowledge about those biological elements are deficient and updating rapidly, this computational system should allow the changing and updates constantly and be flexible enough. Gene Ontology (GO) Consortium was formed to solve such problems. The goal of the Consortium is to produce a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism.

Many databases have been included in the Consortium, and the full list of member organizations can be found in GO Consortium's webpage: http://www.geneontology.org/GO.consortiumlist.shtml#assoc

GO is structured hierarchically as a directed acyclic graph (DAG). As a node, each GO term is connected with other terms in the graph. The relationships between terms are represented as arcs in the graph, and they are categorized as "is a"; "part of"; and "regulates", "negatively regulates" and positively regulates. However, unlike the hierarchy, a GO term may have more than one parent term. There are 3 basic categories of GO: biological process (BP), molecular function (MF) and cellular component (CC). Biological process is defined as the biological objective to which the gene or gene product contributes. Molecular function is the biochemical activities of a gene product at the molecular level. And cellular component provides the information about the parts of a cell or its extracellular environment where a gene product is active. GO is being updated frequently with new terms being created and old ones rendered obsolete. If there are terms tagged "is_obsolete: true", there should be no new annotations attached to these terms.

Genes, gene products or other biological elements can be mapped to the corresponding GO terms according to their attributes. The information from GO consortium and the mapping relationships of genes and GO terms can be accessed by the R package GO.db.

2.9.3 KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a knowledge base for systematic analysis of gene functions, integrating genomic information with higher-level systemic functions of the cell, the organism and the ecosystem[149]. The GENES database is a collection of gene catalogs for all the completely sequenced genomes and some partial genomes with up-to-date annotation of gene functions. The PATHWAY database is one of the most well-known databases of KEGG, which graphically represents the cellular processes, including metabolism, membrane transport, signal transduction and cell cycle. Actually, KEGG pathway maps are widely used for biological interpretation of genome sequences and other high-throughput data. In detail, pathways in KEGG database are stored and represented as graphs, where nodes are molecules (proteins or compounds) and edges denote the relationship between molecules. And pathways can be downloaded for academic purposes from KEGG website as KGML format. It is sometimes necessary to parse and operate the pathways for biological interpretations. And visualization of those pathway graphs makes the analysis results more intuitive and readable. Fortunately, there are diverse tools available.

The KEGG knowledge base has expanded to contain 15 main databases including genomic, chemical, health and drugs information. KEGG base allows to extract meaningful information from large amounts of experimental data more effectively.

2.10 Tools for data analysis and visualization

2.10.1 R and Biocondcutor

R is a language and environment for statistical computing and graphics, with many sophisticated statistical functions implemented. R have many versions, allowing to be compiled and run on diverse platform, including Windows, Mac OS X, and Linux operating systems. Because of the free availability and the advantages in mathematical

and visualization, R has become the world-wide language for computational statistics, data science, visualization and bioinformatics. R is under constant development, and many new methods are added daily. Furthermore, R is supported by a diverse and active community of data scientists and programmers.

The Bioconductor project is an open source and open development software project. Based on the statistical computing environment R, Bioconductor project aims at providing tools for the analysis and comprehension of high-throughput genomic data. Presently, a large number of tools are available as R packages, and their functions cover the analysis and visualization of data from DNA microarray, sequence, flow, SNP, and other data[150].

2.10.2 Tools for microarray data analysis and visualization

Since microarray technology is widely applied in many fields, a series of tools are required to handle the massive amount of data. Currently, various software for data analysis and visualization are freely or commercially available[151].

SAM is a free software widely used for genomic expression data mining. "SAM" is the abbreviation of "significance analysis of microarrays", and the software adopts a modified t-test as well as permutation methods to identified differently expression genes[152]. The samr package for R language has been developed. Other software used for significance analysis of microarray data includes EDGE, Cyber-T and MeV.

Affymetrix gene chip platform is a very popular platform for the studying genes expressions. And many tools have been developed aiming at the data analysis of Affymetrix. The Affymetrix GeneChip Command Console Software (AGCC) is the latest generation of instrument control software for GeneChip systems, and Affymetrix Expression Console Software is used to conduct the probe set summarization, quantification, and normalization integrating the AGCC software. In addition, plentiful free software are available for the Affymetrix platform, such as DNA-Chip Analyzer (dChip), TM4 and RMAExpress.

DAVID Bioinformatics Resource provides online tools for annotation and functional analysis[153]. Containing a biological knowledge base, DAVID establishes a high throughput and integrated data mining environment. There are 4 distinct modules on DAVID website: functional annotation, gene functional classification, gene ID conversion and gene name batch viewer. Expression Analysis Systematic Explorer (EASE) is a downloadable version of DAVID with few added features.

Cytoscape is a general platform for complex network analysis and visualization, especially for biological network[154]. The main function of Cytoscape is to visualize biological network and integrate the network with expression profiles, phenotypes, and other molecular states. Cytoscape is capable of connecting the network with large DNA and protein databases, which enhances this software power. The functions of Cytoscape are extended by many Apps (or plugins), which are available in Cytoscape App Store. Cytoscape has become the standard network visualization tool in molecular biology.

It is deserved to mention that Bioconductor provides a large number of packages based on R language for the analysis and visualization of microarray data. "Affy" is a package designed for the quality assessment, prepossessing and analysis for Affymetrix gene

chip microarray probe level data. And “limma” is a package for differential expression analysis of microarray data. Limma is initiated for the analysis of complex experiments as well as simple ones. Although limma can be used for data input and normalization, the main function of this package is to fit a linear model for the expression data for each gene or probe. There are auxiliary functions for constructing design matrix and contrast matrix as well as estimating “average” variability with empirical Bayes statistics method. Furthermore, “multtest” package also conducts differential expression analysis, but it uses statistical tests. And the distinct advantage of multtest is that many plicable resampling-based single-step and stepwise multiple testing procedures (MTP) for controlling a broad class of Type I error rates are implemented. The multtest package supports a lot of methods for the adjustment of p-values controlling FWER and FDR, so that the results can be more reliable. In some cases, genes or probes are known not express in the tissues or pathways of interests, or they may have similar expression levels across all samples or conditions. Those genes (or probes) with low expression or low variability can be removed before the statistical analysis, so that the number of genes to be tested will reduce considerably and the reliability of result can be enhanced. Such filtering can be accomplished by the package “genefilter”. Using this package, Genes from microarray datasets can be discarded according to a variety of different filtering mechanisms. And users are allowed to create different criteria.

After identifying the differently expressed genes, biological interpretation is always achieved through data annotation and visualization. On bioconductor website, many relevant packages can be found. “GO.db” is an annotation package that combines the structure of the GO terms with the assignment of genes to terms. Different types of gene or probe identifiers can be converted with the help of a series of annotation packages, which are widely used in the enrichment analysis together with GO.db. The enrichment analysis of GO terms can be done by package “topGO”, which implements a number of test statistics and algorithms. And the enrichment results obtained from different methods can be compared easily. In addition, topGO provides visualization function for identifying how the significantly enriched GO terms distribute across GO graph. Users can choose how many significant GO terms should be involved in the graph. “KEGGgraph” package contains the unique function to parse KEGG pathways from KGML files into graphs. Other functions of this packages include graph operation and visualization. Collaborating with other graph package, KEGGgraph is able to address versatile biological problems.

“RCytoscape” package integrates Cytoscape with the statistically powerful programming environment of R[155]. This package remedies the limitation of Cytoscape, which lacks of a full-featured, bioinformatically capable scripting language. Inside of performing manually, all the details of a biological networks are allowed to be defined by the commands and functions of R. Therefore, the reproducibility and efficiency of data exploration is enhanced. When importing network data from R to Cytoscape, R package “XMLRPC” and the plugin “CytoscapeRPC” is required.

2.10.3 Graphs as analysis tools

Graphs have a long-standing history in the applications of numerous scientific fields. The structure of a graph usually is composed of nodes and edges, where nodes represent objects of interest and edges represent relationships between the nodes. Besides, edges in the graphs can have weights and directions if necessary. Because of the flexibility and simplicity, graphs are quite useful for bioinformatics analysis, especially for the

network and sequence analyses. In the analysis of biological network, nodes always stand for genes or proteins, and edges are used to elucidate the diverse pairwise relationships, like co-expression, interactions, inhibition, etc.

As aforementioned, GO ontology is structured as a DAG (directed acyclic graph), which is a graph with directed edges but without cycles. And genes annotated to a certain GO term is a subset of those annotated to its parent nodes.

Most pathways in KEGG are organized as graphs and can be viewed as networks. And the edges of graphs in KEGG can have different attributes representing varied relationships between the objects of networks.

With the help of graph theory, biological networks or pathways can be operated (e.g. subset or merge) and more information can be extracted. Presently, some functionality has been implemented in R/Bioconductor, so that the data from biological databases can be handled with graph algorithms. Parsing the biological data as graphs enable to visualize the networks in an insightful way. The "graph" package carries out basic graph handling capabilities. R package "RBGL" interface with the boost graph library, which contains algorithms for probing and analysis mathematical graphs. "Rgraphviz" connects with AT&T GraphViz software, and it is powerful to display the topology of connectedness between nodes. Rgraphviz is adept to show the graphs with particular attributes, by using different layout features[156]. For the large graphs with high-dimensional data, GGobi is used for visualization. The "rggobi" package complements GGobi's graphical user interface, and supports data transition between R and GGobi.

2.11 Introduction of clustering methods

Clustering is an unsupervised technique used to group together objects which are "close" to one another in a multidimensional feature space, usually for the purpose of uncovering some inherent structure which the unlabeled data possesses. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Obviously, a similarity criterion should be introduced. Distances and concepts are the two commonly used clustering criteria. Euclidean distance, correlation distances, Manhattan distances, Hamming distance, and Edit distance are widely used to measure the dissimilarity between each pairs of data points. And a distance matrix is usually constructed. Depending on the distance measure used, different pairs of data may be considered as "more similar". There are also some way to compute the distance between several probability distributions, such as Chi-Square and KL-Divergence.

There has been a plethora of clustering algorithms (they can be classified as exclusive, overlapping, hierarchical and probabilistic clustering); each of them has advantages and disadvantages. And one of the problems in clustering is to find out the most appropriate algorithm to the particular experiment data. Therefore the results of clustering should be validated. Criteria such as silhouette width, connectivity and dun index can be used to evaluate the clustering results. And for bioinformatics, clustering results should have not only statistical significance but also biological meanings.

Clustering methods can be also classified as grouping data methods and partitioning data methods. Grouping data approaches seek to probe how data are clustered by reconstruction data relation. One of the commonly used grouping data approaches is

hierarchical clustering. Partitioning data approaches attempt to detect and predict the hidden structure in the available data. K-means clustering and fuzzy C-Means clustering are partitioning data approaches.

2.11.1 Hierarchical clustering

Hierarchical clustering is an agglomerative clustering algorithm. It yields a dendrogram which can be cut at a chosen height to produce the desired number of clusters. Each observation is initially placed in its own cluster, and then the clusters are gathered successively according to their "closeness". The closeness between data points is determined by distance matrix. Distance between new formed clusters can be calculated in different ways: single-linkage from complete-linkage and average-linkage. In single linkage clustering, the distance between one cluster and another is considered to be equal to the shortest distance from any member of one cluster to any member of the other cluster. By contrast, in complete linkage clustering, the distance between two clusters is stipulated to be equal to the longest distance from any member of one cluster to any member of the other cluster. And in average linkage, the distance is the average distance from any member of one cluster to any member of the other cluster.

Single linkage may lead to remarkably skewed results, so it is not the best approach for hierarchical clustering. But it is good for picking outliers that are connected in the very last steps of the process. Complete linkage tends to produce very tightly packed clusters. The method is very sensitive for the quality of the data.

Hierarchical clustering has 2 distinct advantages: well visualization of relation between data points and interpretation data with merging distance. Hierarchical clustering is often applied in the analysis of patient samples to organize the data based on the cases, and it suffers from low noise tolerance[157].

2.11.2 K-means clustering

K-means clustering is one of the simplest and fastest clustering methods. It finds iteratively k clusters such that the within-cluster distances from the cluster centroid are minimized.

K-means clustering starts with an initial guess for the cluster centroids, which should be placed in a cunning way. Each point (or observation) is placed in the cluster to which it is nearest. Then the cluster centroids are updated and each point will be associated again to the nearest new centroid. This process (loop) will be repeated until the cluster centroids no longer change. Thus, K-means is an iterative method minimizing within-class sum of squares for a given number of clusters. K-means clustering is also an exclusive clustering algorithm.

The different initializations for number of clusters centroid may lead to different clustering results. There is no guarantee to find a globally optimal result. And a poor choice of k can give poor results. K-means algorithm can be run multiple times to reduce this effect.

Sometimes the number of clusters is given by biological knowledge. When the relevant biological knowledge is unknown, evaluation for the goodness of each clustering result is highly recommended, so that results from different k can be compared.

Furthermore, a parameter like "within-cluster diversity" can be introduced to evaluate the clustering results. However, within-cluster diversity cannot always work well, some statistical measures are thus used to select the models in k-means clustering. For example, Akaike information criterion (AIC) and Bayesian information criterion (BIC) are two criteria which help to select the best model if a probabilistic model for the data can be used. A good model will have a small value of AIC or BIC[158].

2.11.3 Fuzzy C-Means clustering

Fuzzy C-Means clustering allows one piece of data to belong to two or more clusters. This algorithm starts with an initial guess for cluster centers. And each data point will have estimated "memberships" for the k clusters. Then the k cluster centers will recomputed based on the membership values. The process is similar with K-means clustering. And this algorithm suffers from the similar problems with k-means clustering. Thus, Akaike information criterion (AIC) and Bayesian information criterion (BIC) are also used in this method to assist to find the optimal k clusters. Fuzzy C-Means clustering is frequently used in pattern recognition[159].

2.11.4 K-nearest neighbor (KNN)

KNN is the simplest classification method forming clusters by building a classifier. Optimally, two sets are made as training set and test set. The former is used for building the classifier while the latter is used for the validation of the classifier. KNN contains three phases: neighborhood analysis (to select genes for classifying), class prediction (to classify the selected genes into different groups), and validation (to verify results and to rule out the effect of sampling error on the construction of the classifier).

KNN is used for finding a set of genes that differentiates two or more groups of samples. And it performs better when the number of nearest neighbors (K) used for building the classifier is smaller (10-20% of the group).

KNN can only fit a linear discriminator to the dataset. However, data from microarray experiments are always multidimensional, KNN method may not work very well. When KNN classification produces many misclassifications, it is better to use other methods which can fit polynomial discriminators.

2.11.5 Model-based clustering

Model-based clustering uses certain models for clusters and attempt to optimize the fit between the data and the model. In model-based clustering each cluster can be mathematically represented by a parametric distribution. A dataset can thus be modeled by a mixture of these distributions. An individual distribution used to model a specific cluster is often referred to as a "component distribution". And the mixture components and group memberships are estimated using maximum likelihood (EM) algorithm. In simple cases such distributions may be multivariate Gaussians[160].

2.11.6 Visualization of clustering

With the help of visualization methods, clustering results can be understood and analyzed in an intuitive way.

Spot plots and heatmaps are commonly used. Clustering results are always show by grouping genes of clusters next to each other. Genes within a cluster should follow the

average expression pattern of the cluster. Some genes with unique expression patterns do not fit well in any group, and we should pay attention to them and to find if they are truly different expressed genes or just a result of experiment errors in the further analysis.

A red/green color scheme figure is most widely used. Red and green color can represent the two extremes of gene expression. And the color intensity represents the magnitude of deviation.

3. Research Objectives

The main objective is to propose a gene signature distinguishing GBM samples with IDH1 mutation from the counterpart without IDH1 mutation. Furthermore, the overall goal is to create a framework for identifying gene signature as well as analyzing the aberrance of pathways in GBMs. Therefore, several sub-tasks are required to be accomplished, including:

- Data collection
- Selection of appropriate statistical testing methods
- Execution of statistical tests and correction of p-values for the multiple testing problem
- Clinical data analysis
- GO pathways enrichment analysis
- Investigation of GO and KEGG pathways
- Proposing of gene signature validated by hierarchical clustering
- Visualization the analysis results

Ideally, the gene signature can be used to distinguish GBM samples with IDH1 mutations from those without IDH1 mutation. The established framework should be capable of identifying gene signatures correlated with a specific medical condition. In addition, this study intends to find the difference in biological pathways between the 2 types of GBMs. From the overall results, targets for future research can be aroused.

4. Material and methods

4.1 Scripts for researching

All scripts are written with R language.

4.2 Data collection

The processed gene expression data used in this study are download from TCGA website, which are obtained originally from microarray experiments using H113 platform. Different batches of data are collected and into one data matrix. And in this matrix, each row represents a gene, and each column corresponds one sample. In total, there are 12042 genes and 538 samples.

The list of samples with IDH1 mutation is from the paper "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1", which is available in TCGA website.

The clinical data about 579 GBM patients are also obtained from TCGA website.

4.3 Analysis about the normality of gene expression data

In order to check whether the expression value data for each gene are from a normally distributed population, Shapiro-Wilk test is executed[161]. If the p-value for a test is less than the predetermined confidence level (0.01), the corresponding data can be considered as not normally distributed.

The p-value for each test is gathered to make a histogram (figure.4.1). From this figure it is obvious that most gene expression value do not follow normal distribution. With this indication, non-parametric test methods are considered for the following analysis.

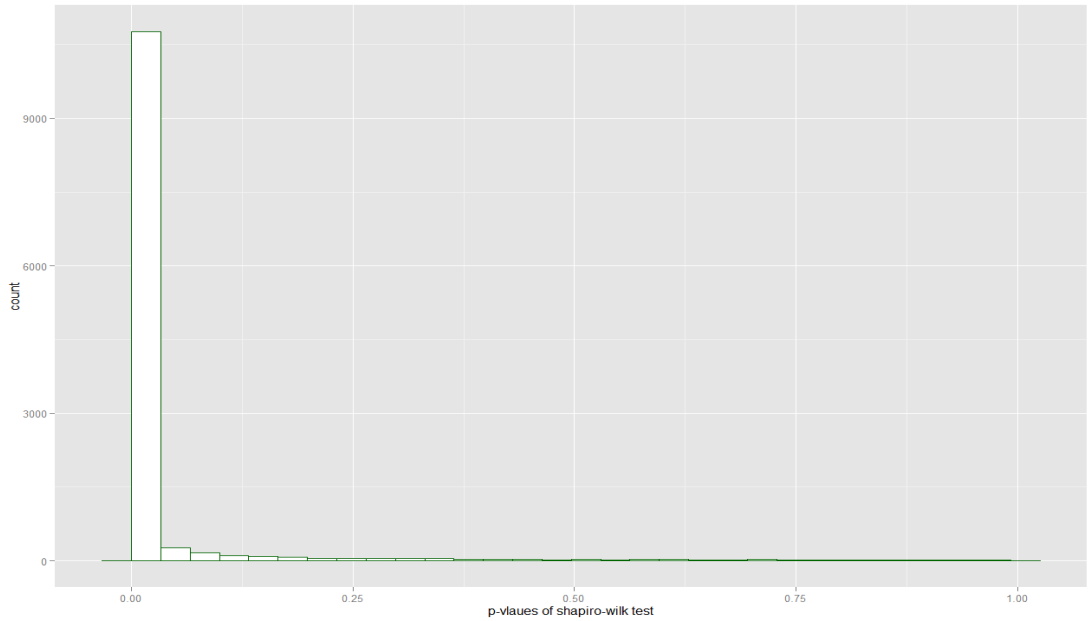

Figure 4.1. Histogram of the p-values for all genes from Shapiro-Wilk test

4.4 Identification of differently expressed genes

4.4.1 The fundamental approach

All the samples are divided as two groups, based on whether they have IDH1 mutations. Samples with IDH1 mutations are denoted as "IDH1+" group, while those without IDH1 mutations belong to "IDH1-" group.

The aim of hypotheses testing is to detect whether the two groups of samples represent two population with different gene expression levels. Thus the null hypothesis for each gene is "there is no difference of the expression values between the IDH1 mutation positive and negative samples". And the alternative hypothesis for each gene is "there is difference of the expression values between the IDH1 mutation positive and negative samples". The test is thus two-sided. The confidence level (α) is set as 0.01, i.e. if the p-value is less than 0.01, the null hypothesis can be rejected and the corresponding gene is identified as differently expressed.

4.4.2 Selection of methods for multiple testing procedure

A test statistic that discriminates between the hypothesis and the alternative should be selected. Since most data do not follow normal distribution and the samples can be assumed as independent, test statistics calculated from Mann–Whitney U test is chosen to perform the basic hypotheses testing. Besides, Mann–Whitney U test get less impact of outliers compared with t-test.

Furthermore, there are 14 samples with IDH1 mutations, while there are more than 500 samples in the other group. And it may cause the class imbalance problem during the analysis. Besides, there are 12042 genes to be tested, so the multiple testing problem is inevitable. To get reliable results, step-down maxT multiple testing procedures, a resampling method with strong controlling of family wise type I error rate, is used in the process of analysis.

Assuming that m genes for n samples are obtained from the microarray experiments, a $m * n$ data matrix X = (x_{ij}) can be constructed with rows corresponding genes and columns to samples. Additional information consists covariates Y describing whether an individual sample has IDH1 mutation, which can be represented as an indicator vector Y. Let $Y_j = 1$ when a sample j has IDH1 mutation, and $Y_j = 0$ otherwise (j=1,2,…,n). Therefore, the null hypothesis for a random gene i can be denoted as:

H_i: There is no association between X_i and Y

The alternative hypothesis is two-tailed: there is association between X_i and Y.

4.4.3 Calculation of statistics

As discussed before, Mann–Whitney U test is used to calculate the test statistics. The details and related formulas can be found in literature review section.

4.4.4 Permutation method for statistical testing[4]

In this study, resampling is achieved by permuting the columns in the expression data matrix which represent the samples of the experiments. This permutation of group

[4] This method is adapted from the paper "Resampling-based multiple testing for microarray data analysis"

labels (in this study, there are 2 kinds of groups) preserves gene-gene correlations and distributional characteristics of the gene expression levels, because the covariates Y is independent with gene expression levels. Since it is impractical to compute all the possible combinations for the samples, permutation tests are always performed by randomly selecting a large number of the possible arrangements for the data. In details, if B represent the total times of permutation, the test statistics of each gene are calculated and written as $t_{1,b}, t_{2,b}, \dots, t_{m,b}$ for bth time of permutation. Then the raw p-value of permutation for the two-sided test (denoted as p_i^*) is the proportion of the permutation test statistics which are as extreme as the original test statistics:

$$p_i^* = \frac{\#\{b: |t_{i,b}| \geq |t_i|\}}{B} \quad for\ i = 1,2, \dots m$$

The algorithm of computing permutation step-down maxT adjust p-values is shown as follow.

1. Calculate the original tests statistics for each hypothesis, and rank then in descending order: $|t_{s_1}| \geq |t_{s_2}| \geq \cdots \geq |t_{s_m}|$.

For each time of permutation, repeat steps 2-4 for B times.

2. Permutate the n columns in the gene expression data matrix.
3. Calculate the permutation test statistics $t_{1,b}, t_{2,b}, \dots, t_{m,b}$ for each hypothesis.
4. Compute the $u_{i,b} = \max_{l=1,2,\dots,m} |t_{s_l,b}|$ by $u_{m,b} = |t_{s_m,b}|$

$$u_{i,b} = \max(u_{i+1,b}, |t_{s_i,b}|) \quad for\ i = m-1, \dots, 1$$

,where b represents each time of the permutation and $u_{i,b}$ is the successive maxima of test statistics.

After repeating the permutation for B times, and enforcing the monotonicity constraints by setting

$$\tilde{p}_{s_1}^* \leftarrow \tilde{p}_{s_1}^*$$
$$\tilde{p}_{s_i}^* \leftarrow \max(\tilde{p}_{s_{i-1}}^*, \tilde{p}_{s_i}^*) \, for\ i = 2,3 \dots, m$$

the adjusted p-value can be estimated as:

$$\tilde{p}_{s_i}^* = \frac{\#\{b: u_{i,b} \geq |t_{s_i}|\}}{B} \quad for\ i = 1,2, \dots m$$

4.5 Pathways Enrichment Analysis

Data analysis of microarray experiments results in a list of differently expressed genes with statistical significance. The consequent task is to interpret the biological information from those genes. One way is to map the differentially expressed genes to onto known pathways or gene ontology, and then perform the enrichment analysis. To be specific, this study is to find out whether the set of genes identified as differently expressed between two groups of samples displays “enrichment” in some pathways or ontology. Such task is always achieved through using various statistical tests.

For the pathway enrichment analysis, the goal is to character the differently expression (DE) genes and identify the pathways those genes associate with. Particularly, we want to detect the significant enrichments of Gene Ontology (GO) categories within the list of DE genes which are found in previous analysis.

A $2 * 2$ contingency table can be made to describe the problem (table.4.1). The non-DE genes are those genes which are considered as non-differently expressed in the previous analysis. In this contingency table, "n" represents the total number of annotations in the GO database for the genes in the experiment. Obviously, the sum of first row "a+b" denotes the number of annotations related with genes of interest (corresponding to the Category 1 in the aforementioned model). For each GO term, the total number of annotations of the genes in the data can be represented the sum of first column "a+c". And "a" is the number of annotations of the genes of interest for the particular GO term.

Table 4.1 Contingency table for enrichment analysis using fisher exact test

	Genes are associated with a GO term	Genes are not associated with a GO term	Row Sums
DE Genes	a	b	a+b = n_1
non-DE Genes	c	d	c+d = n_2
Column Sums	a+c	b+d	a+b+c+d = n

For a certain GO term, the number of genes associated with it might be small, and Fisher exact test is employed to the enrichment analysis. Therefore the probability of getting such a result can be calculated with this equation:

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+c)!\,(b+d)!\,(a+b)!\,(c+d)!}{n!\,a!\,b!\,c!\,d!}$$

The null hypothesis is that: "there is no association between the genes of interest and pathway (GO term)". And the alternative hypothesis is: "the genes of interest are connected with the GO term".

The p-value (p_i) of a certain test is the sum of probabilities corresponding to the observed value "a" and more extreme cases:

$$p_i = 1 - \sum_{j=0}^{a-1} f(j; n, a+b, a+c)$$

If the p-value is very small, it is unlikely to observe such an enrichment by chance given that the differently expressed genes and the GO term are not related.

In order to perform the enrichment analysis, all the genes in the data will be mapped to the annotations of GO terms. Then the p-values of the each GO terms can be calculated. In this study, the confidence level is set as 0.01 ($\alpha = 0.01$). GO terms with a p-value less than 0.01 are regarded as significant (i.e. the GO terms have some connections with the differently expressed genes obtained from the experiments).

In this enrichment analysis the p-values are not adjusted. Usually, the raw p-values in enrichment analyses are not very extreme[162]. And if the p-values are adjusted to control the FWER or FDR, the results might be very conservative and no or very few GO terms can be identified as significant. In this case some interesting GO terms could be ignored and valuable information will be lost. Furthermore, many assumptions have been

adopted before conducting the enrichment tests. Hence, it is not enough to control the error rates only by considering the number of GO terms (tests).

4.6 Analysis of the KEGG pathways

To get insight of how the DE genes participate in some pathways, an analysis about relevant KEGG pathways is performed.

First of all, 3 KEGG pathways are selected and merged, and they are "glioma", "regulation of actin cytoskeleton", and "pathways in cancer" (their KEGG pathway ID are "05214", "04810", and "05200", respectively).

With the help of R package "KEGGgraph"[163], the 3 pathways are downloaded from KEGG site and parsed into graphs. And the 3 pathways are merged into one network because of the fact that some KEGG pathways embed other pathways. In addition, some pathways only record the genes involved in, but do not provide the relationships (edges) between those genes and others. However, such relationships can be found in other pathways. And merging pathways can solve this problem.

The genes are recorded as distinct identifiers in KEGG pathways, while the expression data from TCGA site adopt gene official symbols as identifier. Therefore, gene IDs in KEGG pathways are converted to gene symbols using R package "org.Hs.eg.db". And the DE genes from the expression data analysis can be identified whether to participate in the 3 KEGG pathways. And how those genes interact with others in KEGG network can be examined.

4.7 Visualization of pathways

Visualization is an efficient way to comprehend and learn from the data. In this study, the significantly enriched GO pathways and the 3 pathways related to glioma are visualized as graphs using Cytoscape and several R packages. Besides, the hierarchical clustering result of the DE genes and samples without IDH1 mutation is visualized as heatmap.

4.7.1 Visualization of GO pathways

Upon the enrichment analysis, a list of statistically significant GO terms is obtained. To explore how those GO terms are distributed over the GO system (the biological process ontology), a directed acyclic graph containing all the significant GO terms and their ancestors are created. There are 118 nodes and 223 edges in this graph. The nodes in the graph are labeled as their names, and all nodes representing the significant GO terms are square. The color of each node is linked to the p-value of each GO term in the enrichment analysis: the darker the color is, the smaller the corresponding p-value is.

Since the graph looks complicated, 3 sub-graphs harboring the most significant GO terms are made. The 3 simple sub-graphs incorporate similar information with the original graph, but they are more comprehensible.

4.7.2 Visualization of KEGG pathways

The 3 KEGG pathway are merged and a graph is generated. This graph have 467 nodes and 1898 edges representing genes and their relationships. Since the gene symbols are more readable, they are used as nodes labels in the network of KEGG pathways.

Different types of relationships between genes are represented by edges with different colors and shapes, and the detailed stipulations are elucidated in table.4.2.

Table.4.2 Stipulation for visualizing KEGG pathways interactions

Relationship Type	Line Type	Source Arrow	Target Arrow	Color
activation	solid	-	arrow	green
phosphorylation	sinewave	-	arrow	green
inhibition	dash dot	-	T arrow	red
expression	solid	-	delta arrow	blue
dissociation	dash dot	-	-	black
dephosphorylation	dot	-	arrow	green
compound	dot	-	-	black
binding/association	solid	arrow	arrow	cyan
indirect effect	equal dash	-	arrow	black
missing interaction	dot	-	-	dark red

There are 4 DE genes are involved in this merged pathway, and they are FGF17, MSN, ITGB8, and PDGFA. However, the graph is too large and complicated, and it looks like a hairball. Consequently, some sub-graphs are made for the simplification purpose. Nodes with different sizes are assigned to genes based on their expression values. Moreover, the nodes in the sub-graphs are colored corresponding to the log fold change of each gene between the two kinds of sample group.

4.8 Hierarchical clustering and visualization

After performing the multiple testing procedure, 58 DE genes are identified and 50 of them are participated in the enrichment analysis of GO. 2 new matrices consisting expression data of the 50 genes are formed, which contain all the samples with IDH1 mutation (14) or without IDH1 mutation (534) respectively. And the 2 matrices are merged into a large one, which is for clustering genes using hierarchical clustering based on Euclidean distance. Furthermore, all the IDH1 mutation absent samples are clustered with the same method. Finally a clustered matrix is obtain, with the rows representing genes and columns representing samples. And this clustered matrix is visualized as heatmap. In the heatmap, red color denotes the low expression values, while green color indicates the high expression values.

5. Results

5.1 Data collection

The gene expression data are from microarray experiments using the "HT_HG-U133A" platform. After downloading from TCGA site, data are assembled into one file. In total, 12042 genes expression data from 538 GBM samples are obtained. Among the 538 GBM samples, 14 samples harbor IDH1 mutation. All the data have been pre-processed.

The clinical data from TCGA site contains different types of clinical information from 579 samples, including "vital status", "age at initial pathologic diagnosis", "death days to", and "tumor status".

5.2 Multiple testing procedure

To perform the statistical testing, samples are divided into 2 groups according to whether they contain IDH1 mutation. To be simplified, the 2 groups are denoted as IDH1+ (with IDH1 mutation) and IDH1- (without IDH1 mutation).

The confidence level of testing is set as 0.01. After 100000 times permutation, there are 58 hypotheses (genes) whose step-down maxT adjust p-values are less than preset 0.01. The list of differently expression (DE) genes are presented in table 5.1, where the corresponding p-values, means of expression level as well as log fold changes (lfc) between the 2 groups are show. The genes are named after the official symbols. (Note that the expression data in TCGA site has been log-transformed, thus the lfc for each gene is calculated by subtracting the mean expression value of IDH1- from the counterpart of IDH1+.) Another information is about whether the DE genes participate in the following enrichment analysis of GO.

Table 5.1 Genes identified as differently expressed between IDH1+ and IDH1- group

Gene ID	p-value	Mean of expression value	Log fold change	Participant of enrichment analysis
C13orf18	0.00003	6.50481843505201	-1.7594423258852	FALSE
SLC2A10	0.00168	7.46581666859923	-1.9150559157706	TRUE
C1orf107	0.00011	6.02261090102387	-0.623579467055817	TRUE
SDF4	0.00013	7.73593932220155	-0.898515241468418	TRUE
HRH1	0.00735	7.04064086154501	-1.48060700681628	TRUE
KLHL26	0.00453	6.3338242722834	-1.1189940433284	FALSE
TRIM48	0.00245	3.97511774451691	0.490785176537107	FALSE
GALNS	0.00095	6.96020656346752	-0.808163801639173	TRUE
MSN	0.00665	9.66746011523786	-1.53046964713946	TRUE
LDHA	0.00487	12.3074759945108	-1.07775135394969	TRUE
SYNJ2	0.00212	5.12410183143534	-0.547907567331171	TRUE
PLA2G5	0.00035	6.34784245278702	-2.33045012483499	TRUE
EFEMP2	0.00735	7.69813379611282	-1.86564391945418	TRUE
GPR172A	0.00268	6.0666750020006	-0.635645803050076	TRUE
M6PRBP1	0.00194	8.59829097849664	-0.992722131599963	TRUE
AK3L1	0.0013	8.4941886078041	-1.6168604700219	TRUE
FGF17	0.00535	4.26248812444411	0.253443360938069	TRUE

OSBPL10	0.00103	6.56026011270817	-1.65829036087509	TRUE
ITGB8	0.00612	5.32601620181649	-0.832741876576378	TRUE
CHST2	0.00039	8.46073001506132	-1.59513521956752	TRUE
MYO1E	0.00083	5.83316532823624	-0.733524348553893	TRUE
PLAT	0.00117	8.59058151796819	-1.7702037187437	TRUE
CHST7	0.00008	6.79953922274088	-1.50447419773885	TRUE
PHLDA3	0.00222	6.08742984683793	-0.770160524581396	TRUE
SLC22A18	0.00076	6.26263608928778	-1.08909804675403	TRUE
FHL2	0.00255	6.45770539762117	-1.55737875797186	TRUE
ALDOA	0.00089	11.8938046037745	-0.764873469512798	TRUE
ANXA5	0.00586	10.9302304754921	-0.921580743285823	TRUE
ACRV1	0.00864	3.9897896643043	0.250108213784156	TRUE
BDH1	0.0043	5.69648233545281	-0.946421454400804	TRUE
ELOVL6	0.00114	5.74994645280678	-0.796107142706778	TRUE
DUSP5	0.00612	5.78823916743292	-1.10129642438526	TRUE
SPRY2	0.00627	9.00128139527228	-1.5027289996529	TRUE
NSUN5	0.00004	7.02347309368607	-1.16836321216677	FALSE
MEOX2	0.00055	6.53659028548913	-2.86618350197156	TRUE
C20orf23	0.00616	5.45255932048912	-0.856925090525791	TRUE
ARSJ	0.00759	5.44013646547739	-1.30774197914524	TRUE
CXCL14	0.00501	8.30057854103491	-3.10201048186541	TRUE
MRC2	0.00363	7.64166783714594	-1.18866001283532	TRUE
CD97	0.00334	6.1789474647951	-1.08767784430674	TRUE
OPLAH	0.00149	4.52847562466045	-0.518567413120151	TRUE
CYP27A1	0.00695	6.8420002532146	-0.762349168144615	TRUE
TMEM22	0.0013	7.62938687115094	-1.7821185913363	FALSE
ZNF492	0.00033	4.9128656793679	0.620298050262146	TRUE
ACSL3	0.00387	8.78019385729841	-1.00450522197869	TRUE
FLJ11286	0.00008	7.94001078532698	-1.49653649275423	FALSE
FLJ21963	0.00011	6.41905274095563	-1.80130464933032	FALSE
IMPACT	0.00457	6.89535157793583	-0.994913444437907	TRUE
TRIP6	0.00010	7.90373430689605	-1.66666272252316	TRUE
PDLIM4	0.00978	5.670215544313	-1.00220387359665	FALSE
PDGFA	0.00013	7.54141522216949	-1.67480178317232	TRUE
ELOVL2	0.00063	6.4971040348515	-1.86774149986604	TRUE
PMP22	0.00205	10.8124357948992	-1.47076270137476	TRUE
PIPOX	0.00010	7.66064053544618	-2.27058433235844	TRUE
STEAP3	0.00307	7.20480661809553	-1.58750087171272	TRUE
RAB36	0.00004	5.2719111443214	-0.836496692396398	TRUE
MOXD1	0.00689	7.57323012307391	-2.34576356887844	TRUE
CNKSR1	0.00227	4.40394753350249	0.365849465714881	TRUE

5.3 Analysis of clinical data

The clinical data contains information from 579 samples, and among them there are 12 IDH1+ samples and 530 IDH- samples whose gene expression data are used in previous statistical testing. After removing "not available" data, the remaining clinical data for 12 IDH1+ samples and 530 IDH- samples are visualized as boxplots (figure 5.1) and Kaplan–Meier plot (figure 5.2). In the same way, IDH1+ represents the samples with IDH1 mutation, while IDH1- means the samples without IDH1 mutation in the 2 figures.

The boxplots in figure 5.1 show the age at initial pathological diagnosis for samples. It is obvious that patients with IDH1 mutation are younger than those without such mutation. In the Kaplan–Meier plot, IDH1+ samples display a better overall survival than IDH1- samples. The clinical information shown in the 2 figures conforms to previous knowledge about IDH1 mutation.

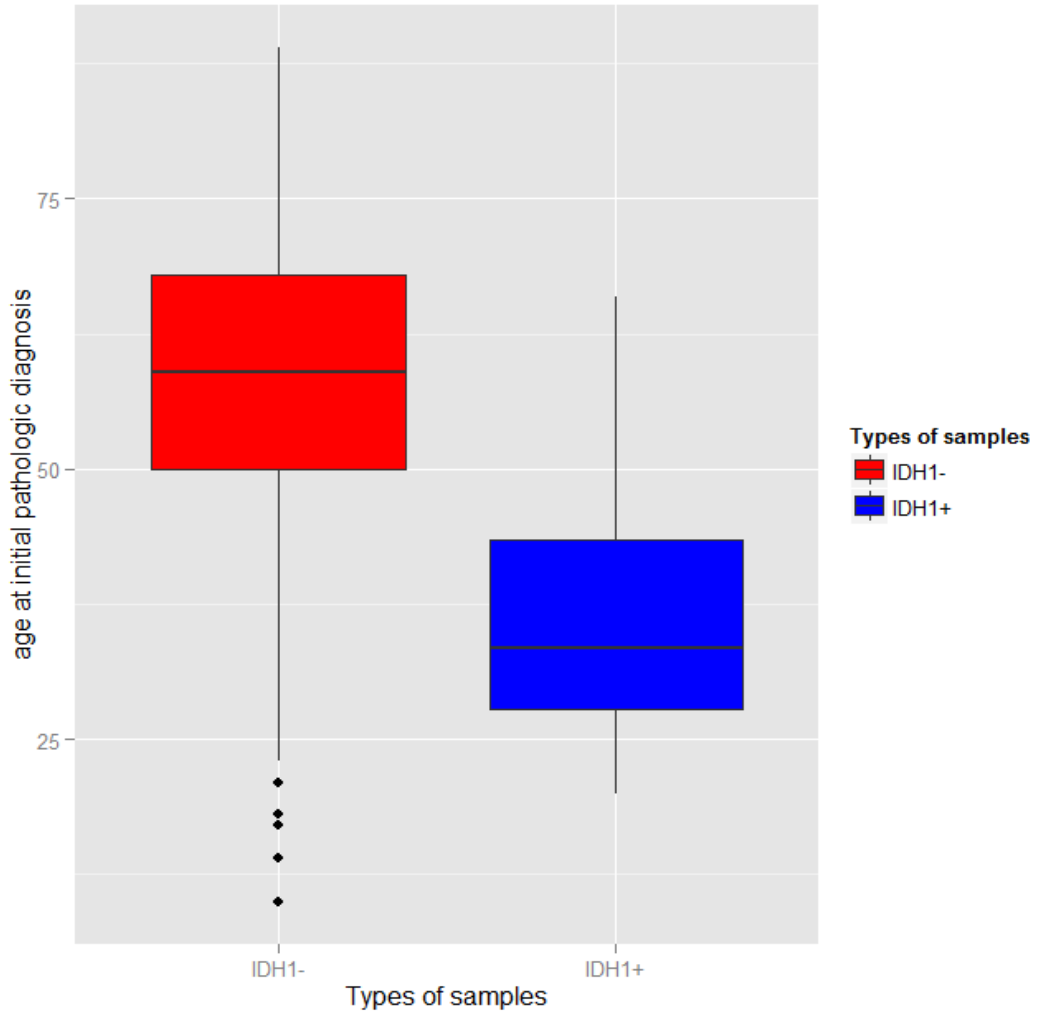

Figure.5.1 Boxplots about age at initial pathologic diagnosis based on 2 groups

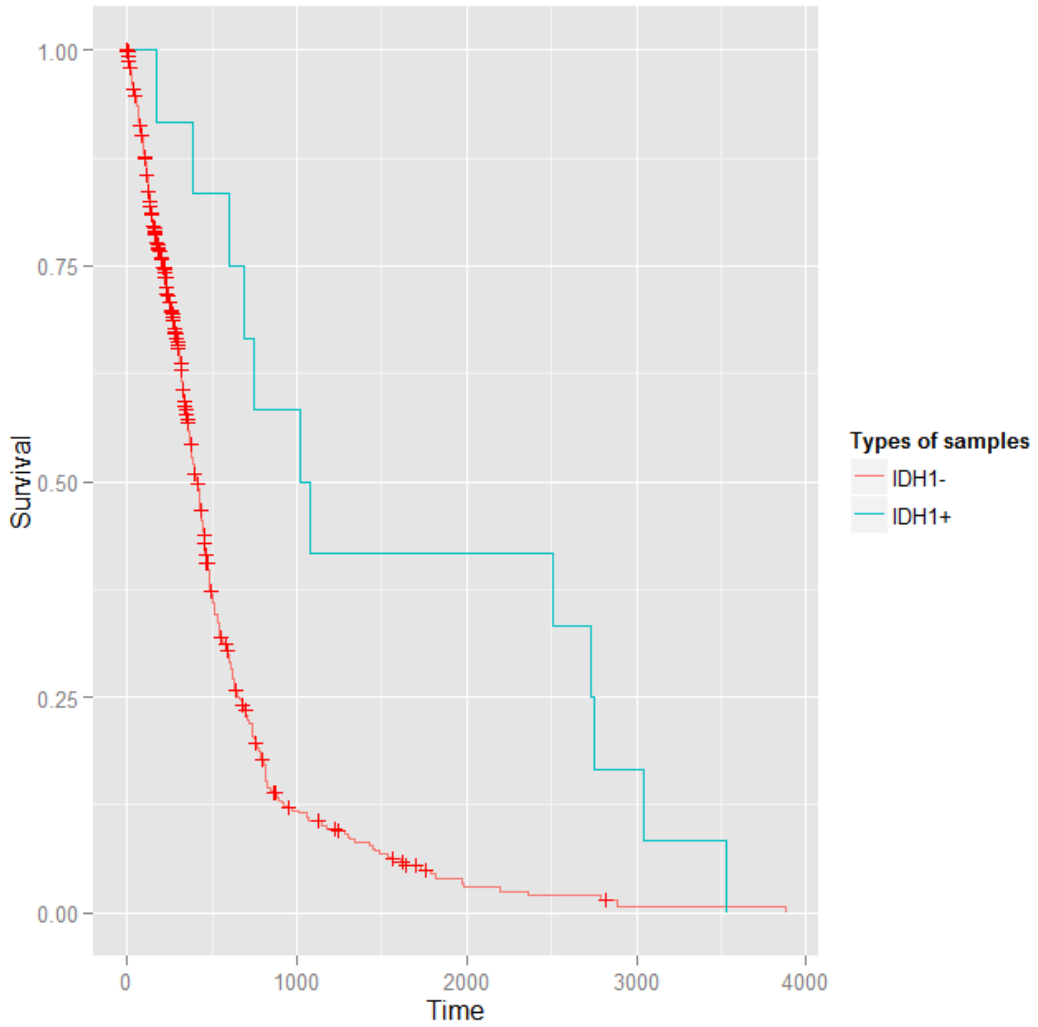

Figure.5.2 Kaplan–Meier plot for IDH1+ and IDH1- sample groups

5.4 Enrichment analysis of GO

Not all the genes from previous study participate in the enrichment analysis, because some genes are not found to be assigned to any BP GO terms, and the Entrez ID of some genes are unknown. In fact, 10515 genes out of all the 12042 genes can be used to perform this analysis, and 50 genes of them are DE genes.

The fisher exact test identifies 36 GO terms of the biological process, which DE genes are significantly enriched (p-value < 0.01). The table 5.2 shows the list of 36 GO terms as well as the p-values of fisher exact test. Also, the number of annotated genes and DE genes among them for each GO term is included in table 5.2. And the definition of each GO term as well as the symbols of DE genes annotated can be found in appendix2.

Table.5.2 Significant enriched GO terms

GO ID	Annotated	Significant	p-value	Term
GO:0045017	191	6	0.00027	glycerolipid biosynthetic process

GO:0030497	6	2	0.00033	fatty acid elongation
GO:0071071	7	2	0.00046	regulation of phospholipid biosynthetic process
GO:0044283	396	8	0.00050	small molecule biosynthetic process
GO:0044711	410	8	0.00064	single-organism biosynthetic process
GO:0048008	37	3	0.00070	platelet-derived growth factor receptor signaling pathway
GO:0050819	43	3	0.00109	negative regulation of coagulation
GO:0072330	169	5	0.00119	monocarboxylic acid biosynthetic process
GO:0016053	261	6	0.00140	organic acid biosynthetic process
GO:0046394	261	6	0.00140	carboxylic acid biosynthetic process
GO:0046486	264	6	0.00149	glycerolipid metabolic process
GO:0006044	13	2	0.00167	N-acetylglucosamine metabolic process
GO:0044255	726	10	0.00190	cellular lipid metabolic process
GO:0043436	861	11	0.00200	oxoacid metabolic process
GO:0006637	54	3	0.00212	acyl-CoA metabolic process
GO:0035383	54	3	0.00212	thioester metabolic process
GO:0006082	874	11	0.00225	organic acid metabolic process
GO:0006633	117	4	0.00226	fatty acid biosynthetic process
GO:0035338	16	2	0.00255	long-chain fatty-acyl-CoA biosynthetic process
GO:0035336	17	2	0.00288	long-chain fatty-acyl-CoA metabolic process
GO:0046949	18	2	0.00323	fatty-acyl-CoA biosynthetic process
GO:0035337	21	2	0.00439	fatty-acyl-CoA metabolic process
GO:1901071	21	2	0.00439	glucosamine-containing compound metabolic process
GO:0044281	2226	19	0.00479	small molecule metabolic process
GO:0050818	72	3	0.00479	regulation of coagulation
GO:0090407	450	7	0.00509	organophosphate biosynthetic process
GO:0010741	148	4	0.00525	negative regulation of intracellular protein kinase cascade
GO:0006629	984	11	0.00563	lipid metabolic process
GO:0046474	155	4	0.00618	glycerophospholipid biosynthetic process
GO:0044710	2650	21	0.00679	single-organism metabolic process
GO:0008543	164	4	0.00752	fibroblast growth factor receptor signaling pathway
GO:0008610	488	7	0.00786	lipid biosynthetic process
GO:0008654	171	4	0.00869	phospholipid biosynthetic process
GO:0051896	90	3	0.00889	regulation of protein kinase B signaling cascade
GO:0006040	31	2	0.00944	amino sugar metabolic process
GO:0042339	31	2	0.00944	keratan sulfate metabolic process

Many of the significant GO terms are about the reactions and pathways involved in lipid, glycerolipid, phospholipid, and glycerophospholipid. Secondly, many GO terms describe the pathways and reactions implicated in the formation and elongation of fatty acid and fatty-acyl-CoA. And there is also a GO term on the pathways involving thioester (acetyl-CoA is a derivative of thioester). Furthermore, several GO terms about

the metabolism of different types of organic acids are found. Of note, there are some GO terms about those molecular signals generated by fibroblast growth factor receptor (FGFR), platelet-derived growth factor receptor (PDGFR), protein kinase B (AKT), and other protein kinase. Other biological processes include the reactions and pathways involving small molecular, keratan sulfate, glucosamine and N-acetylglucosamine.
To visualize the significant GO terms and their location in the hierarchical GO system, a graph (figure.5.3) is generated. For the purpose of clarity, 3 sub-graphs (figure.5.4-figure.5.6) are made, and the details can be read easier. In these graphs, square nodes represent the significant GO terms. The smaller the p-value of a GO term is, the darker the color of the corresponding node is.

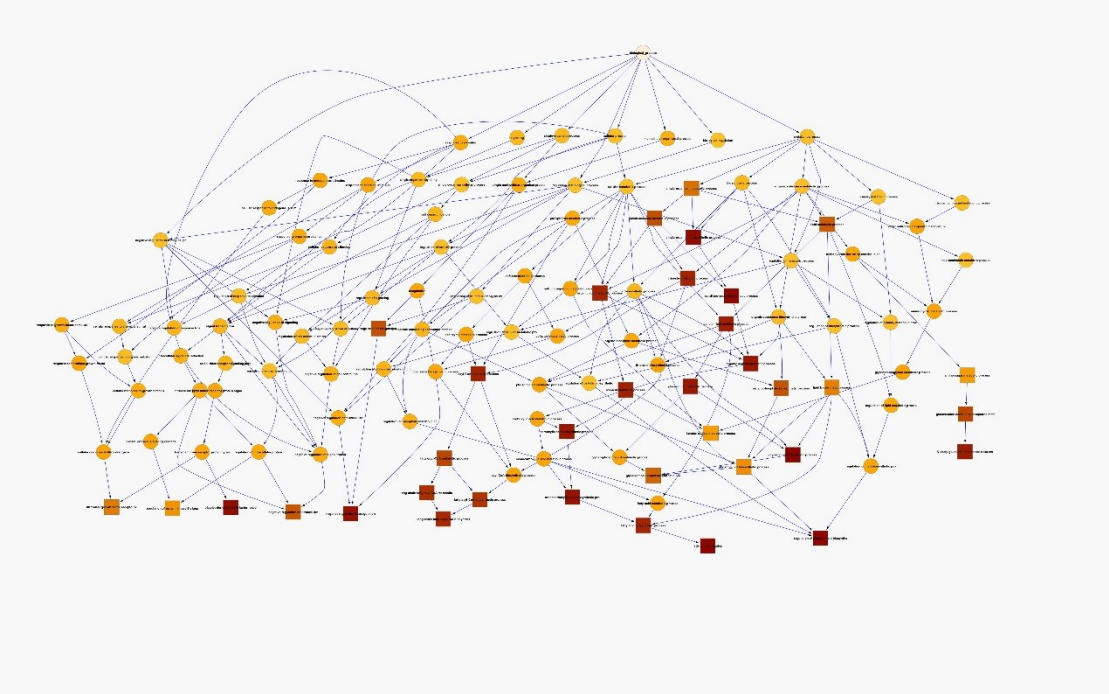

Figure.5.3 hierarchical graph of the GO biological process system

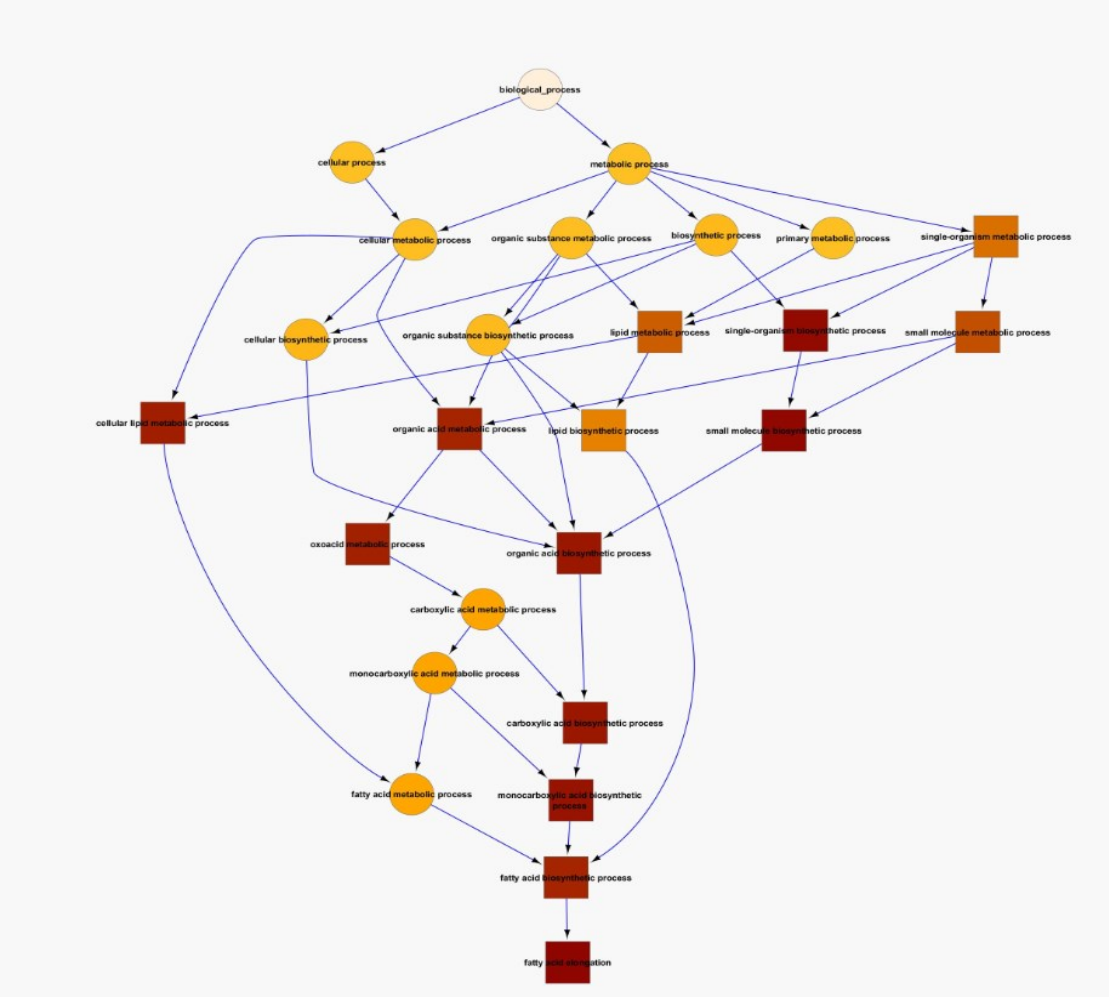

Figure.5.4 sub-graph of GO terms hierarchy focusing on organic acids metabolism

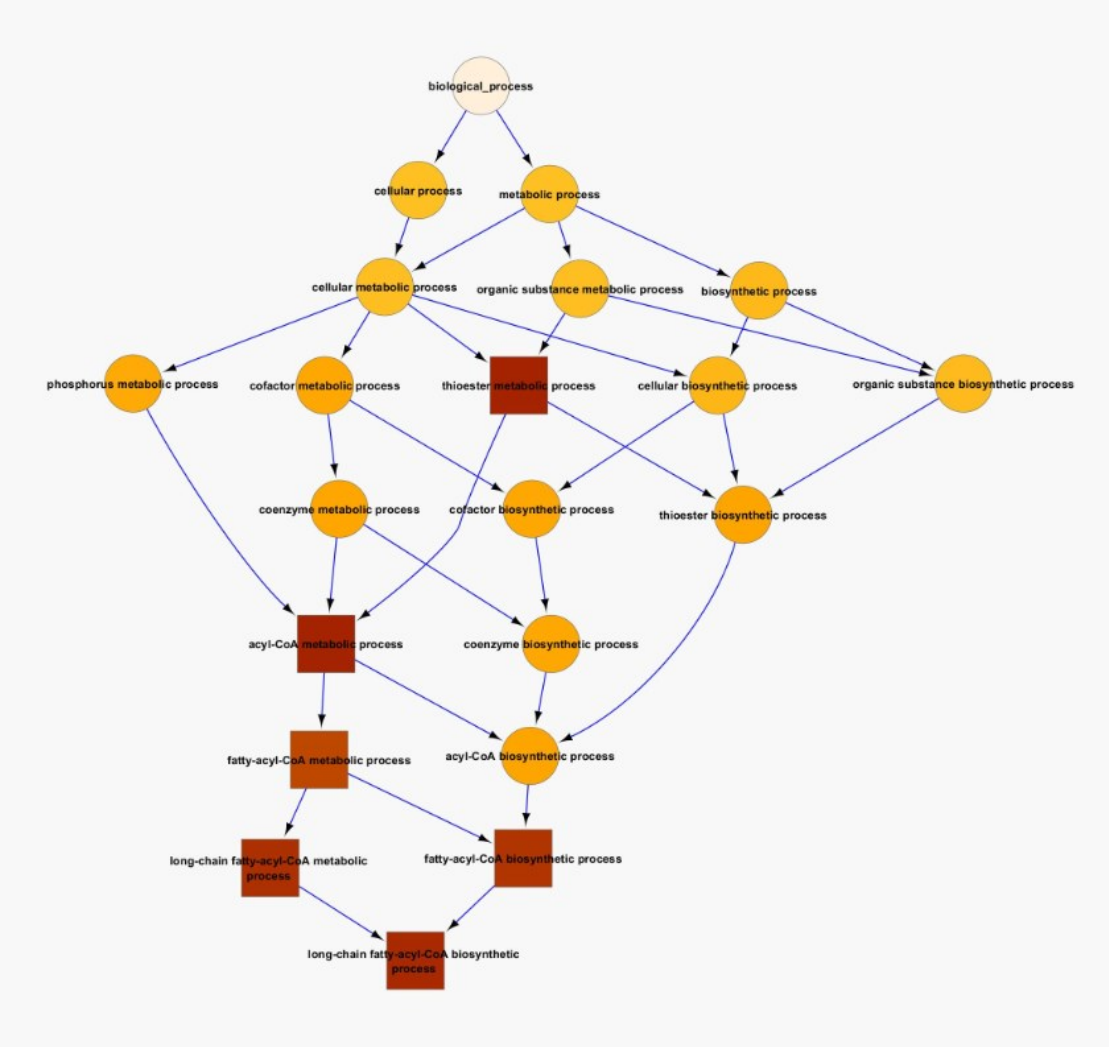

Figure.5.5 sub-graph of GO terms hierarchy focusing on acyl-CoA metabolism

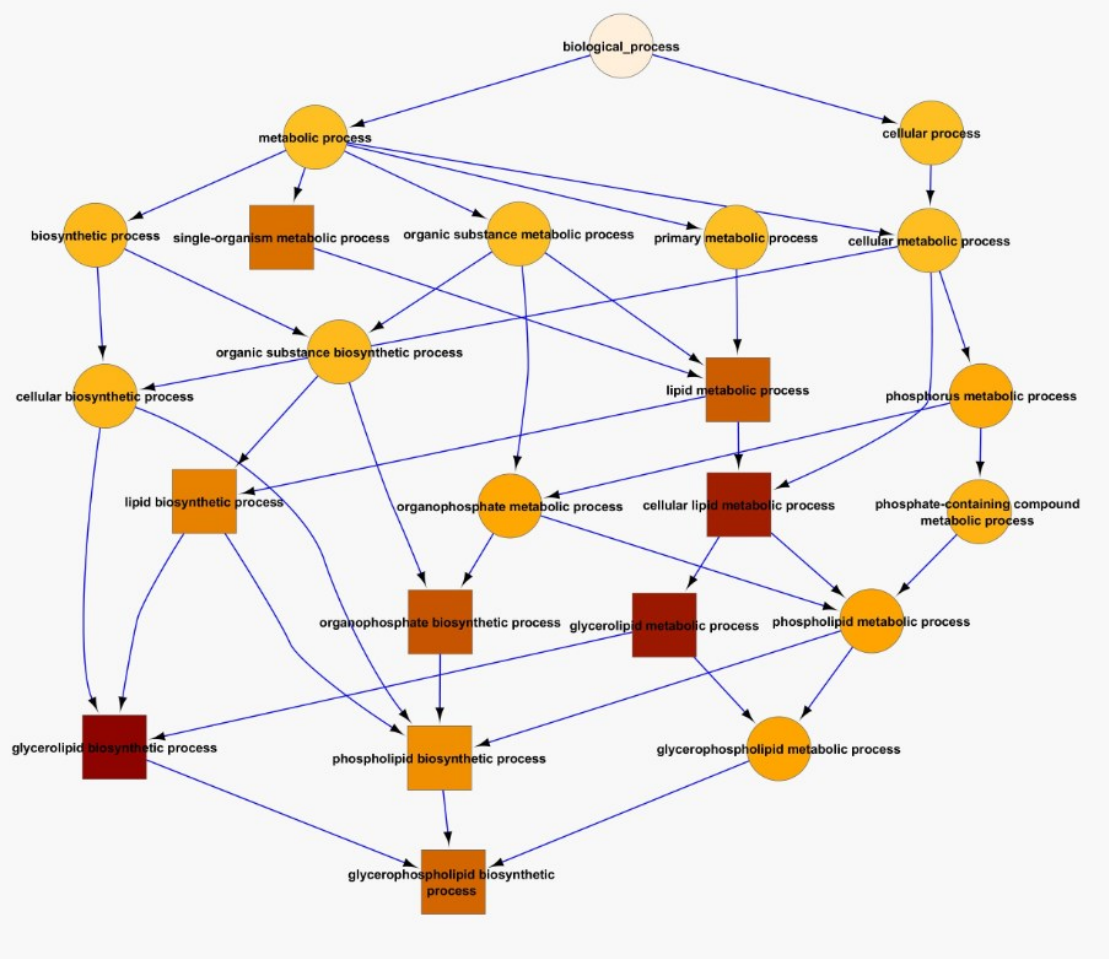

Figure.5.6 sub-graph of GO terms hierarchy focusing on lipid metabolism

5.5 Investigation of KEGG pathways

Three KEGG pathways, "glioma", "regulation of actin cytoskeleton", and "pathways in cancer", are selected and merged into one network. The merged network is visualized as a graph, containing 467 nodes (genes) and 1898 edges (different relationships). And 4 of the DE genes are found in this network, their symbols are FGF17, ITGB8, PDGFA and MSN. In addition, there are some genes connected with glioma in this network, but they are not found as DE genes in statistical testing procedure. To get an insight of those genes, a sub-graph from this network is drawn and shown in figure 5.7. And 4 sub-graphs focusing on the 4 DE genes and their neighbors are created (figure 5.8-5.11). Moreover, the size of each node reflects the corresponding gene expression level: genes higher expression level will have a larger size. And different colors of nodes represent the log fold change (lfc) between IDH1+ and IDH1- groups. Red and green color indicate the high and low extreme of lfc, respectively.

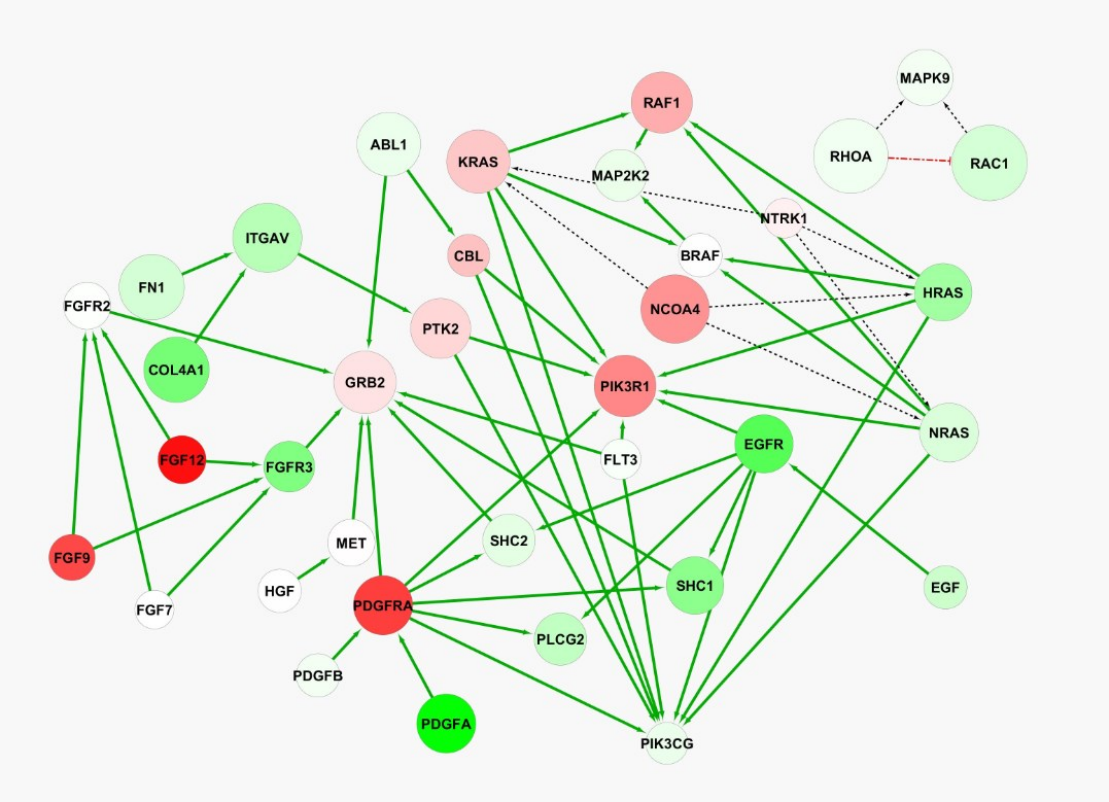

Figure.5.7 Important genes and interactions in glioma-related KEGG network

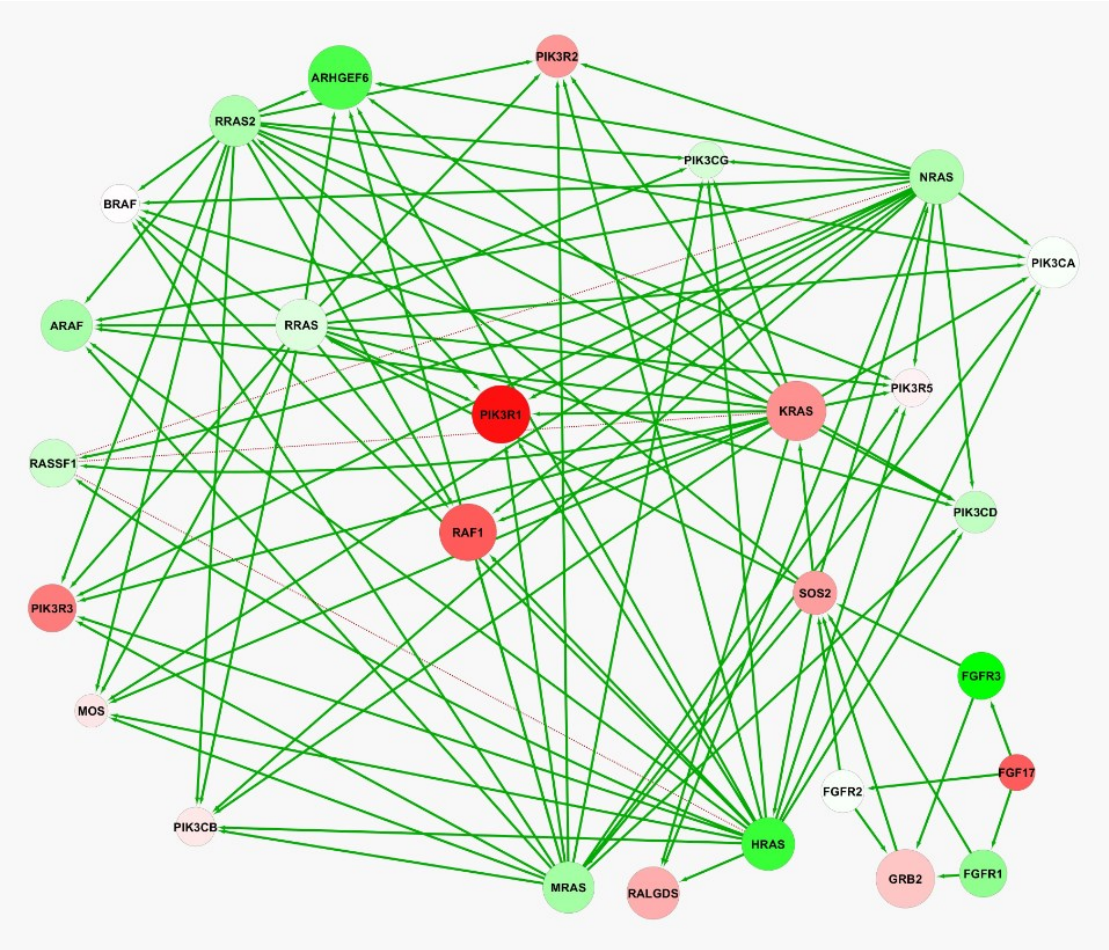

Figure.5.8 FGF17 and neighbor genes in glioma-related KEGG network

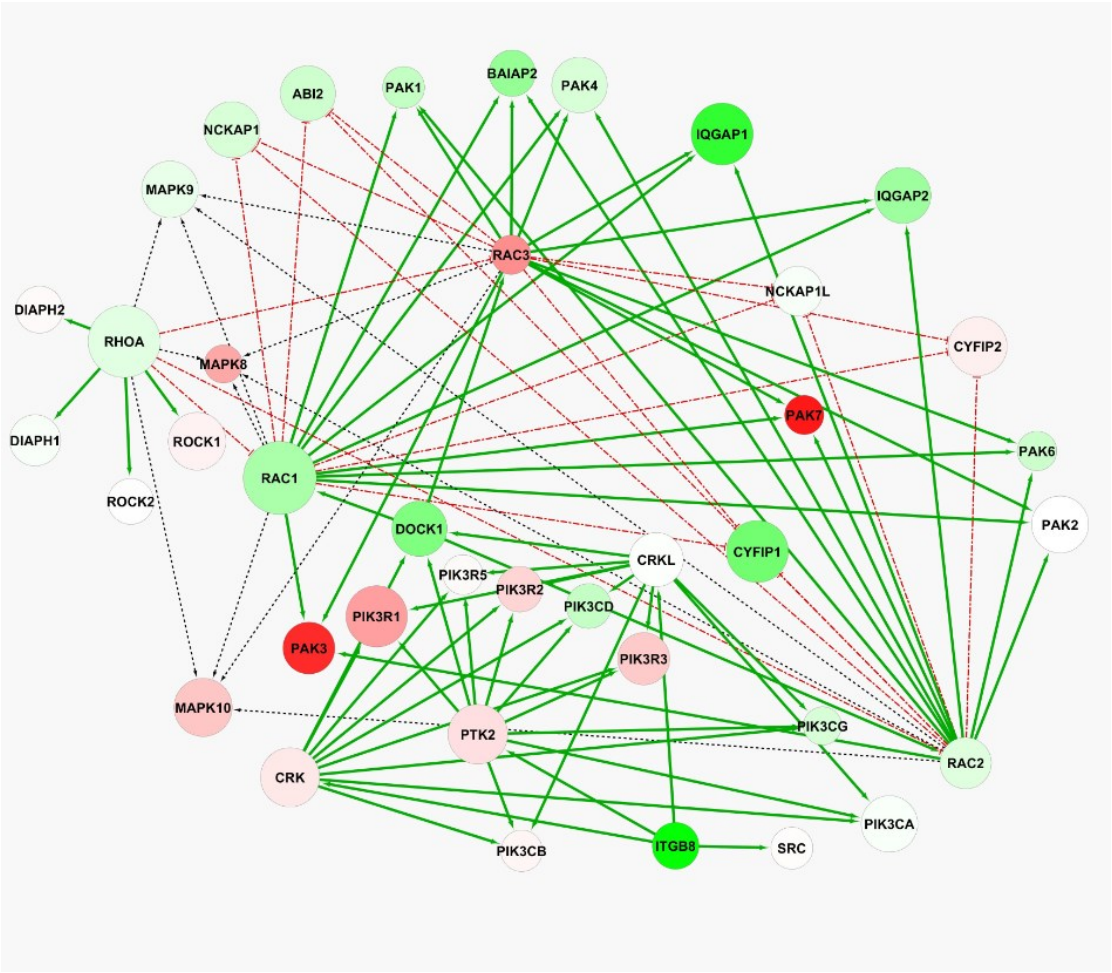

Figure.5.9 ITGB8 and neighbor genes in glioma-related KEGG network

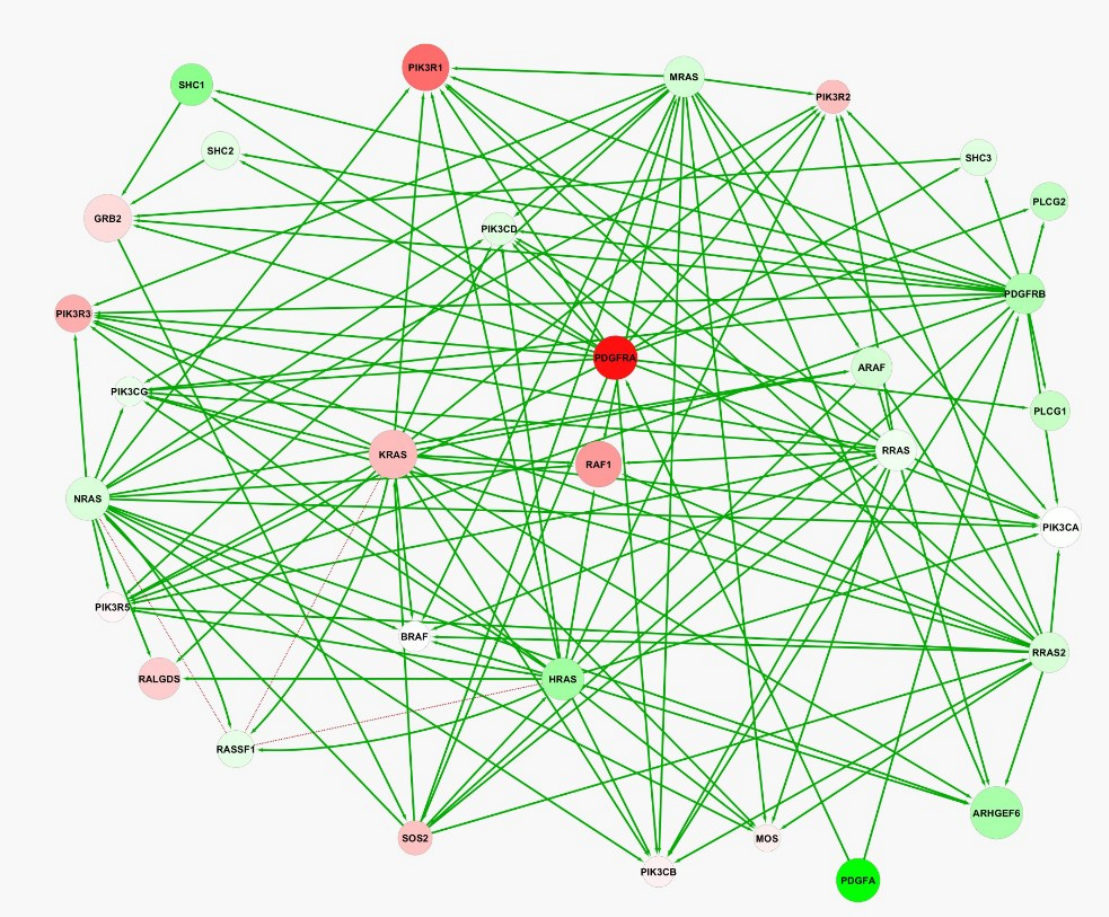

Figure.5.10 PDGFA and neighbor genes in glioma-related KEGG network

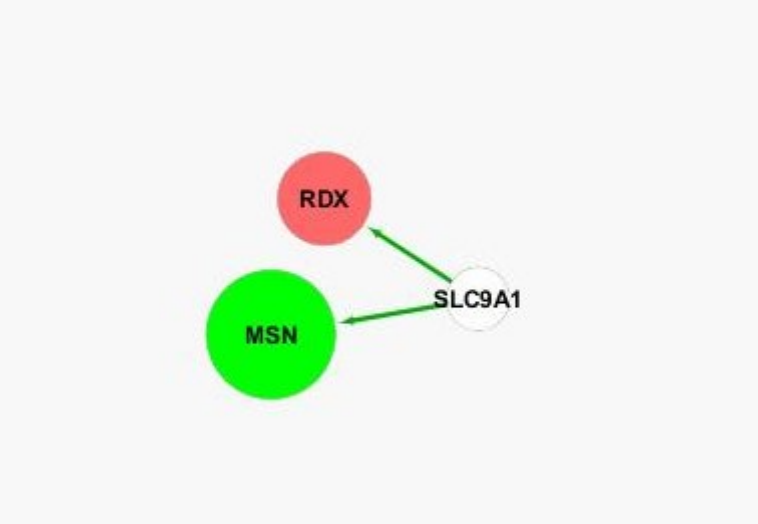

Figure.5.11 MSN and neighbor genes in glioma-related KEGG network

5.6 Clustering and heatmap

The clustering result based on DE genes and samples with or without IDH1 mutation is visualized as heatmap (figure 5.12). Since there are more than 500 samples, this heatmap is not so clear. Therefore, to "amplify" the boundary between IDH1+ group and IDH1- group, a heatmap containing the first 60 samples are drawn (figure 5.13). From the smaller heatmap, striking differences of gene expression between the 2 groups

of samples can be seen. Such marked boundary between the 2 groups confirms that the 50 genes work well as a gene signature.

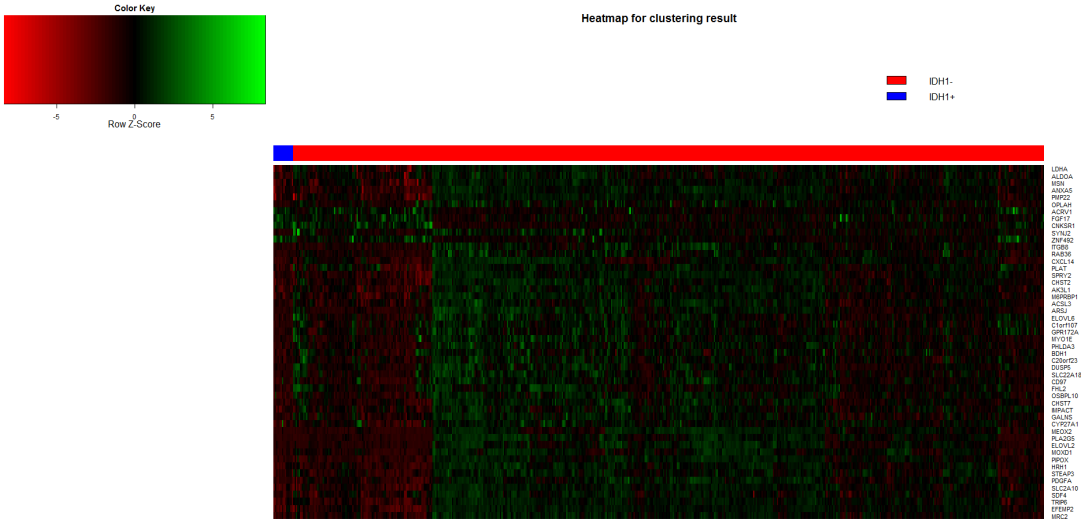

Figure.5.12 Heatmap of the clustering result based on the gene signature

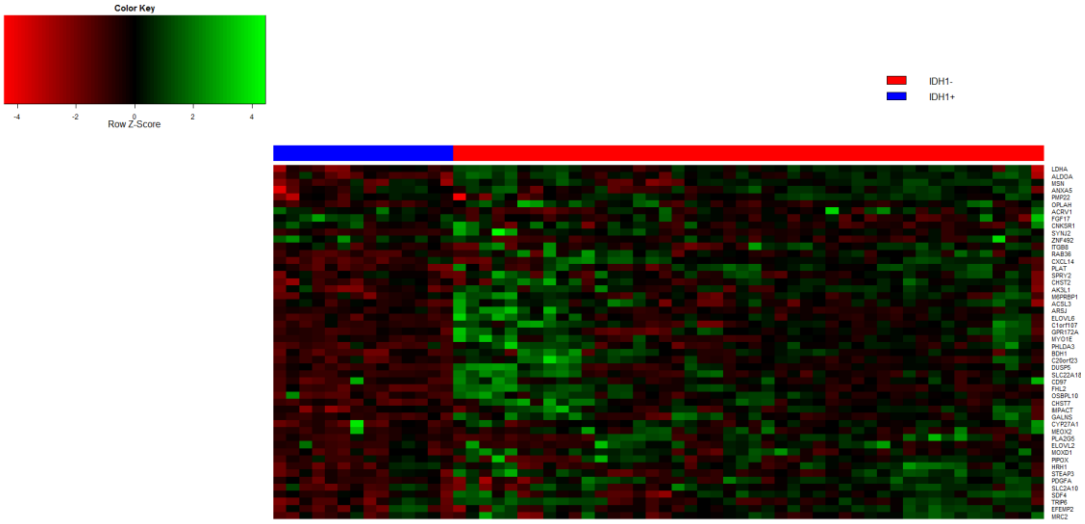

Figure.5.13 Heatmap of the clustering result of the first 60 samples

6. Discussion

Understanding tumorigenesis of GBM and classification is significantly advanced through identification gene signatures which display different expression patterns between different subtypes of GBM. And gene ontology enrichment analysis based on gene signatures enable researchers to find out the mechanism of the formation of different tumor.

In this study, 58 genes are found to be differently expressed (p-value < 0.01) between GBM samples with IDH1 mutation and those without the mutation. And a gene signature containing 50 genes is proposed after performing GO enrichment analysis. Furthermore, GO enrichment analysis also indicates the pathways or chemical reactions those gene and their products participate in. In details, the main pathways and chemical reactions involve in the metabolism of lipid, fatty acid, organic acid and thioester. In other words, most significant GO terms are about the metabolism reprogramming, which is one of the hallmark of cancer. The analysis of clinical data exhibits a result which is consistent with previous knowledge. Investigation of differently expressed (DE) genes in KEGG network show the roles and the change of expression levels of DE genes as well as their neighbors in glioma-related pathways.

6.1 Analysis of the results

6.1.1 Discussion about the data analysis of gene expression and clinical data

Mann–Whitney U test is selected to find out the DE genes with statistical significance, and resampling method (permutation) is used to ensure that the test is robust enough. The Step-down maxT multiple testing procedure is performed, and the adjusted p-values are obtained by controlling FWER. Finally, 58 genes are found with significantly differential expression (p-value < 0.01).

From the survival curve and boxplot about diagnosis ages, it is easy to find that samples with IDH1 mutation tend to be younger and survive for a longer time. Obviously, patients with IDH1 mutation display a feature similar with the Proneural subtype. Since the main characteristic of Proneural GBM is harboring frequent IDH1 mutation, the result of clinical data analysis conforms to the classification of GBM.

Most of DE genes have lower expression levels in the IDH1+ group. In fact, there are only 4 genes (FGF17, ACRV1, ZNF492, and CNKSR1) displaying a slightly higher expression level in samples with IDH1 mutation. All other genes express higher in those samples without such mutation. Combined with the clinical data, it is possible that most of DE genes are associated shorter survival time and more aggressive tumors in GBM patients.

And among them 50 genes are proposed as a gene signature and they are used for the clustering of samples. The clustering result show a striking boundary between the 2 groups, suggesting that samples can be divided into 2 groups on the basis of this gene signature. And GO enrichment analysis is conducted based on this gene signature.

6.1.2 Discussion about the GO enrichment analysis

In this study, 36 GO terms are found as significant, and they can be divided as several groups: lipid metabolism, coagulation process, the series of molecular signals, glucosamine-containing compounds metabolism, and keratan sulfate.

Elevated lipogenesis has been revealed as a main feature of cancer[164]. Especially, the lipid level in malignant gliomas tumor tissues are higher than normal ones[165]. As known, lipids are crucial for the formation of new cellular membranes, which is necessary for the rapidly growing and dividing cancer cells. Some types of lipids even act as regulator for signal transductions. And lipids are alternative energy resource for cells[166].

A study on GBM shows that EGFR signaling expedite the activation of SREBP-1[167], which is a master transcriptional regulator of fatty acid synthesis[168]. In other words, EGFR signaling can promote the transcriptional activation for some fatty acid synthase and increased the amounts of intracellular fatty acids. Moreover, high level of EGFR signaling makes the cells be more dependent on fatty acids synthesis. Besides, another study confirms that targeting fatty acid synthesis could be effective to block tumor cell growth[169].

4 genes (ELOVL6, ELOVL2, PLA2G5, and ACSL3) are annotated to the GO terms about fatty acids metabolism, and all of them express at a lower level in IDH1+ samples, compared with those without IDH1 mutation. The difference of expression from genes involved in lipid metabolism indicate that lipid synthesis is a potential target for designing new treatments. Apart from that, such a result is consistent with the previous study: IDH1 is implicated in lipid biosynthesis.

The GO terms about organic acids metabolism are also significantly enriched, and genes involved in the fatty acids metabolism are annotated in these GO terms as well. Hence it is possible that the metabolism of other types of organic acids except fatty acids differs between the 2 sample groups. To be specific, samples without IDH1 mutation may have elevated organic acids metabolism. In details, CYP27A1, OPLAH, LDHA, CHST2, CHST7, PIPOX and GALNS genes are involved in these GO terms.

In addition, several DE genes are assigned to the GO terms about glycerolipids and glycerophospholipids biological process, which may indicate that there are some difference for the metabolism of glycerolipids and glycerophospholipids between the 2 types of samples. Some researches has proposed that glycerophospholipid will be a novel drug target against cancer[170].

Moreover, GO terms of coagulation regulation are enriched with 3 genes (PLAT, PDGFA, and ANXA5) annotated. Previous studies had demonstrated that the activation of coagulation system prompts tumor growth and invasion in human glioma[171]. And scientists suggests that anticoagulation in patients with gliomas will have anticancer activity[172]. Coagulation is also connected with thrombembolic events, which is found in patients with primary and secondary brain tumors[173]. And the thrombembolic events is one of the factor that diminish the survival time of cancer patients. Taken together, the coagulation system may control the behavior of tumors and it can be a target of the novel therapy.

There are 2 signaling pathways (PDGFR and FGFR) enriched. In addition, the enriched GO terms include process regulating the protein kinase B signaling cascade. Such a result supports pathways mediated by PDGFR, FGFR, AKT and other protein kinases are crucial pathways for the tumorigenesis of GBM. And it has been confirmed by

various studies. Importantly, the result in this study indicates that there might be some discrepancy in these pathways between samples with or without IDH1 mutation.

Markedly, pathways about amino sugars (N-acetylglucosamine and glucosamine) are enriched, and 2 genes (CHST2 and CHST7) are assigned. The molecular beta1,6-N-acetylglucosamine (beta1,6-GlcNAc)-bearing N-glycans has been found in human gliomas, while it is absent in normal brain cells. And the expression of beta1,6-GlcNAc-bearing N-glycans is correlated with the invasivity of gliomas[174]. On the other hand, glucosamine is a prominent precursor in the biochemical synthesis of glycosylated proteins and lipids. And glucosamine is demonstrated to induce autophagic cell death in glioma cells[175]. The enriched GO terms imply the difference of animo sugars metabolism may exist between the two types of samples, and such a difference may be correlated with the disparity of lipids metabolism between the groups.

Another enriched GO term is about pathway and reactions of keratan sulfate (KS). And N-acetylglucosamine is an important residue of KS[176]. A study show that KS is highly expressed on a cell surface in a glioblastoma cell line, and the KS is detected as highly sulfated in glioblastoma cells[177]. Nevertheless, the structure and function of KS in glioblastoma remains obscure. Thus the finding in this study provides a cue for the future research about understanding glioblastoma.

The GO enrichment analysis using DE genes reveals that some pathways may differ between the IDH1+ and IDH1- samples. However, further studies are necessary to find the connections between those differences in pathways and IDH1 mutation as well as the functions of the DE genes. For example, it is worth discussing whether IDH1 mutation is the direct reason for the difference in these biological process. And there are some DE genes not annotated to any significant GO terms, but studying their functions are also worthwhile to understand how IDH1 mutant affect the glioma cells.

6.1.3 Discussion about the KEGG network analysis

In this study the KEGG pathways are combined to perform the analysis, so that all the information from the pathways is included, and the redundant one is removed.
Except genes roles in pathways, graphs representing KEGG network show some genes expression level and the log fold change between IDH1+ and IDH1- groups. There are 4 genes (PDGFA, FGF17, ITGB8 and MSN) in the KEGG network from the gene signature.

Literature review confirms the validity of this study. In IDH1+ group the expression of PDGFA gene is decreased, while PDGFRA is up-regulated. Since the PDGFA is positively correlated with tumor grade, it is possible that tumors with IDH1 mutation tend to be more indolent.

And the expression of FGF17 is elevated, while all the FGFR1-3 are down-regulated. Of note, the expression change of FGFR2 is slight compared with FGFR1 and FGFR3. Unlike FGFR1, which is abundant in malignant tumors, FGFR2 are only found in normal tissues and low-grade astrocytomas. Hence such a finding is in accord with that glioblastomas with IDH1 mutation is less aggressive than those without IDH1 mutation. And future studies are required to find out the reason for up-regulating FGF17 in IDH1+ tumors.

Expression of ITGB8 (the gene product is β8 integrin) is recently found to drive the invasive growth behaviors of GBM by interacting with RhoGDI1, and thus shorten the survival time of patients significantly[178]. And in this study, ITGB8 expresses at a lower level in IDH1+ tumors, which is conformity with the longer survival time and less invasiveness of this type of tumor. However, whether the lower expression of ITGB8 is directly associated with IDH1 mutation is not known.

The moesin encoded by MSN is a member of ERM family, which serve both as cross-linkers between plasma membranes and actin-based cytoskeleton and as regulators of signaling transduction of cytoskeletal remodeling[179]. The cytoskeleton plays a prominent role in the cellular morphogenesis[180]. MSN expression level is significantly higher in astrocytoma relative to normal brain and the MSN up-regulation is associated with the pathological grade of astrocytoma. Furthermore, MSN expression is identified as strongly negatively correlated with the patient survival[181]. Lately, a study demonstrates that moesin directly binds to microtubules in vitro and stabilizes microtubules at the cell cortex in vivo[182]. Compared with many genes, MSN expresses at a high level across all samples in this study. However, in line with the better survival, MSN gene has a significantly decreased expression in IDH1+ tumors.

Both ITGB8 and MSN have not been previously described as playing a role in development or progression GBM, so it is possible that other genes in this gene signature will be found as important in the tumourigenesis of GBM. For example, some researchers conjecture that up-regulation of CD97 promotes cellular invasion and migration in gliomas[183]. And CD97 expression is inversely correlated with survival time of GBM patients[184]. This gene signature may provide information for discovering the predictors of poor prognosis and targets for novel therapy.

In summary, this study proposes a gene signature to distinguish GBM samples harboring mutant IDH1 from the counterparts without mutant IDH1. Further researches are necessary to afford a more comprehensive understanding of the connection between this gene signature and IDH1 mutation in GBM. Hopefully, this study will facilitate the development of more effective diagnosis and treatments.

6.2 Known limitations and potential enhancements

First of all, the selection of methods for statistical test will affect the result. If other multiple testing procedures are chosen, the result of might be different. And in this study p-values are adjusted by controlling FWER, which is a conservative way[185], and some DE genes may be ignored. Therefore, less conservative methods can be adopted to attempt to identify genes which may differently expressed between the 2 groups.

Secondly, the samples with IDH1 mutation are much less than the samples without IDH1 mutation, which may lead to the imbalance problem. And the result of clinical data analysis thus does not have a statistical significance. In other words, more GBM samples with IDH1 mutation are required to get a more reliable result.

The selection of KEGG pathways is also crucial for the investigation of genes function in biological pathways. The three selected KEGG pathways only include 4 DE genes. All other genes from the gene signature are omitted in this section. In the graph representing KEGG pathways, the size of a node is assigned according to average expression level of the gene across all samples. Nevertheless, the heterogeneity between

samples is ignored. Therefore, it would be meaningful to make similar graphs for each samples, and novel ideas might be extracted from the comparison of these graphs.

7. Conclusion

This study proposes a gene signature which is correlated with the IDH mutation in GBM samples. And the gene signature are validated by hierarchical clustering. The relevant clinical data and pathways information are analyzed. As the result indicated, this gene signature is available for discerning IDH1 mutant samples from those GBMs with IDH1 mutation.

With the development of novel techniques and software, it is accessible to identify differently expressed genes linking to a medical condition, examine the enrichment states of those genes in different pathways, investigate the roles of the genes in biological pathways and visualize the data in an intuitive way. Importantly, this study illustrates a pipeline for identification gene signatures. With the help of this pipeline, it is possible to gain insights into the aberrances in genome and pathways of cancer for a particular condition. Moreover, future researches on IDH1 mutation in GBMs are provoked.

Reference

1. Meyer M a. Malignant gliomas in adults. *N Engl J Med*. 2008;359(17):1850; author reply 1850. doi:10.1056/NEJMc086380.
2. Omuro A, DeAngelis LM. Glioblastoma and other malignant gliomas: a clinical review. *JAMA*. 2013;310(17):1842-50. doi:10.1001/jama.2013.280319.
3. Liang Y, Diehn M, Watson N, et al. Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proc Natl Acad Sci U S A*. 2005;102(16):5814-9. doi:10.1073/pnas.0402870102.
4. Chang C, Xu K, Shu H. The role of isocitrate dehydrogenase mutations in glioma brain tumors. *Mol targets CNS tumors*. 2011. Available at: http://cdn.intechweb.org/pdfs/19945.pdf. Accessed May 5, 2014.
5. Parsons DW, Jones S, Zhang X, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science*. 2008;321(5897):1807-12. doi:10.1126/science.1164382.
6. Labussiere M, Sanson M, Idbaih A, Delattre J-Y. IDH1 gene mutations: a new paradigm in glioma prognosis and therapy? *Oncologist*. 2010;15(2):196-9. doi:10.1634/theoncologist.2009-0218.
7. Hahn W, Weinberg R. Rules for making human tumor cells. *N Engl J Med*. 2002;347(20):1593-1604. Available at: http://www.nejm.org/doi/full/10.1056/NEJMra021902. Accessed May 5, 2014.
8. Anand P, Kunnumakkara AB, Kunnumakara AB, et al. Cancer is a preventable disease that requires major lifestyle changes. *Pharm Res*. 2008;25(9):2097-116. doi:10.1007/s11095-008-9661-9.
9. Fidler I. The pathogenesis of cancer metastasis: the'seed and soil'hypothesis revisited. *Nat Rev Cancer*. 2003;3(June):1-6. Available at: http://www.nature.com/nrc/journal/v3/n6/abs/nrc1098.html. Accessed May 7, 2014.
10. Bengmark S. The natural history of primary and secondary malignant tumors of the liver. *Cancer*. 1969;23(1):198-202.
11. Kleihues P, Sobin LH. World Health Organization classification of tumors. *Cancer*. 2000;88(12):2887.
12. Louis DN, Ohgaki H, Wiestler OD, et al. The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol*. 2007;114(2):97-109. doi:10.1007/s00401-007-0243-4.
13. Pan E, Prados M. *Holland-Frei Cancer Medicine. 6th edition*.; 2003.
14. Gilbertson RJ. High-grade glioma: can we teach an old dogma new tricks? *Cancer Cell*. 2006;9(3):147-8. doi:10.1016/j.ccr.2006.02.024.
15. Furnari FB, Fenton T, Bachoo RM, et al. Malignant astrocytic glioma : genetics , biology , and paths to treatment. 2007:2683-2710. doi:10.1101/gad.1596707.instability.
16. Kleihues P, Ohgaki H. Primary and secondary glioblastomas: from concept to clinical diagnosis. *Neuro Oncol*. 1999;1(1):44-51. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1919466&tool=pmcentrez&rendertype=abstract.
17. Kleihues P, Soylemezoglu F, Schäuble B, Scheithauer BW, Burger PC. Histopathology, classification, and grading of gliomas. *Glia*. 1995;15(3):211-21. doi:10.1002/glia.440150303.
18. Phillips HS, Kharbanda S, Chen R, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell*. 2006;9(3):157-73. doi:10.1016/j.ccr.2006.02.019.
19. Hanahan D, Weinberg R. The hallmarks of cancer. *Cell*. 2000;100:57-70. Available at: http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.elsevier-e82bf1bf-57fd-3911-add9-8e1592a777a9/c/00816839.pdf. Accessed May 6, 2014.
20. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. *Cell*. 2011;144(5):646-674.
21. Verhaak RGW, Hoadley K a, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. 2010;17(1):98-110. doi:10.1016/j.ccr.2009.12.020.
22. AP B. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*. 1980;8:1499–1504.
23. Antequera F. Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci C*. 2003;60:1647-1658. doi:10.1007/s00018-003-.
24. Deaton A, Bird A. CpG islands and the regulation of transcription. *Genes Dev*. 2011:1010-1022. doi:10.1101/gad.2037511.1010.

25. Bogdanović O, Veenstra GJC. DNA methylation and methyl-CpG binding proteins: developmental requirements and function. *Chromosoma*. 2009;118(5):549–565.
26. Portela A, Esteller M. Epigenetic modifications and human disease. *Nat Biotechnol*. 2010;28(10):1057-68. doi:10.1038/nbt.1685.
27. Phillips T. The Role of Methylation in Gene Expression. *Nat Educ*. 2008;1:116.
28. Jones PA, Takai D. The Role of DNA Methylation in Mammalian Epigenetics. *Science (80-)*. 2001;293:1068-1070.
29. Cokus SJ, Feng S, Zhang X, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*. 2008;452(7184):215-9. doi:10.1038/nature06745.
30. AM K, YJ P, C P, HH S. Determination of genomic 5-hydroxymethyl-2'-deoxycytidine in human DNA by capillary electrophoresis with laser induced fluorescence. *Epigenetics*. 2011;6(5):560-5.
31. Kanwal R, Gupta S. Epigenetic modifications in cancer. *Clin Genet*. 2012;81(4):303-11. doi:10.1111/j.1399-0004.2011.01809.x.
32. Urdinguio RG, Sanchez-Mut J V, Esteller M. Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies. *Lancet Neurol*. 2009;8(11):1056-72. doi:10.1016/S1474-4422(09)70262-5.
33. Weller M, Stupp R, Reifenberger G, Brandes AA. MGMT promoter methylation in malignant gliomas: ready for personalized medicine? *Nat Rev Neurol*. 2010;6:39-51.
34. Cancer T, Atlas G. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008;455(7216):1061-8. doi:10.1038/nature07385.
35. Christensen BC, Smith A a, Zheng S, et al. DNA methylation, isocitrate dehydrogenase mutation, and survival in glioma. *J Natl Cancer Inst*. 2011;103(2):143-53. doi:10.1093/jnci/djq497.
36. Venneti S, Felicella MM, Coyne T, et al. Histone 3 lysine 9 trimethylation is differentially associated with isocitrate dehydrogenase mutations in oligodendrogliomas and high-grade astrocytomas. *J Neuropathol Exp Neurol*. 2013;72(4):298-306. doi:10.1097/NEN.0b013e3182898113.
37. Ornitz DM, Itoh N. Protein family review Fibroblast growth factors Gene organization and evolutionary history. 2001:1-12.
38. Turner N, Grose R. Fibroblast growth factor signalling: from development to cancer. *Nat Rev Cancer*. 2010;10:116-129.
39. Powers CJ, McLeskey SW, Wellstein a. Fibroblast growth factors, their receptors and signaling. *Endocr Relat Cancer*. 2000;7(3):165-97. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11021964.
40. DE J, J L, H C, S W, LT W. The human fibroblast growth factor receptor genes: a common structural arrangement underlies the mechanisms for generating receptor forms that differ in their third immunoglobulin domain. *Mol Cell Biol*. 1991;11:4627–4634.
41. Takahashi J a, Mori H, Fukumoto M, et al. Gene expression of fibroblast growth factors in human gliomas and meningiomas: demonstration of cellular source of basic fibroblast growth factor mRNA and peptide in tumor tissues. *Proc Natl Acad Sci U S A*. 1990;87(15):5710-4. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=54397&tool=pmcentrez&rendertype=abstract.
42. JA T, M F, K I, Y O, H K, M. H. Correlation of basic fibroblast growth factor expression levels with the degree of malignancy and vascularity in human gliomas. *J Neurosurg*. 1992;76(5):792-8.
43. Auguste P, Gürsel DB, Lemière S, et al. Inhibition of Fibroblast Growth Factor / Fibroblast Growth Factor Receptor Activity in Glioma Cells Impedes Tumor Growth by Both Angiogenesis-dependent and -independent Mechanisms Inhibition of Fibroblast Growth Factor / Fibroblast Growth Factor Receptor . 2001.
44. X S, FV M, GR M. Functions of FGF signalling from the apical ectodermal ridge in limb development. *Nature*. 2002;418(6897):501-8.
45. RS M, F Y, JM B, M T, W M, MS B. Fibroblast growth factor receptor gene expression and immunoreactivity are elevated in human glioblastoma multiforme. *Cancer Res*. 1994;54(10):2794-9.
46. Loilome W, Joshi AD, ap Rhys CMJ, et al. Glioblastoma cell growth is suppressed by disruption of Fibroblast Growth Factor pathway signaling. *J Neurooncol*. 2009;94(3):359-66. doi:10.1007/s11060-009-9885-5.
47. Fredriksson L, Li H, Eriksson U. The PDGF family: four gene products form five dimeric isoforms. *Cytokine Growth Factor Rev*. 2004;15(4):197–204.

48. Demoulin J-B, Montano-Almendras CP. Platelet-derived growth factors and their receptors in normal and malignant hematopoiesis. *Am J Blood Res*. 2012;2(1):44-56. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3301440&tool=pmcentrez&rendertype=abstract.
49. Nazarenko I, Hede S-M, He X, et al. PDGF and PDGF receptors in glioma. *Ups J Med Sci*. 2012;117(2):99-112. doi:10.3109/03009734.2012.665097.
50. Hermanson M, Funa K, Hartman M, et al. Platelet-derived Growth Factor and Its Receptors in Human Glioma Tissue : Expression of Messenger RNA and Protein Suggests the Presence of Autocrine and Paracrine Loops Platelet-derived Growth Factor and Its Receptors in Human Glioma Tissue : Expression o. 1992:3213-3219.
51. Lokker NA, Sullivan CM, Hollenbach SJ, Israel MA, Giese NA. Platelet-derived Growth Factor (PDGF) Autocrine Signaling Regulates Survival and Mitogenic Pathways in Glioblastoma Cells : Evidence That the Novel PDGF-C and PDGF-D Ligands May Play a Role in the Development of Brain Tumors Platelet-derived Growth Fact. 2002.
52. F DR, RS C, J Z, PM B. Platelet-derived growth factor and its receptor expression in human oligodendrogliomas. *Neurosurgery*. 1998;42(2):341-6.
53. M H, M N, C B, CH H, B W, K F. Endothelial cell hyperplasia in human glioblastoma: coexpression of mRNA for platelet-derived growth factor (PDGF) B chain and PDGF receptor suggests autocrine growth stimulation. *Proc Natl Acad Sci*. 1988;85(20):7748-52.
54. GJ T, C F, JE. DL. Transforming growth factors produced by certain human tumor cells: polypeptides that interact with epidermal growth factor receptors. *Proc Natl Acad Sci*. 1980;77(9):5258-62.
55. P von B, J S, K D, M W-K, E K. Correlation of TGF-alpha and EGF-receptor expression with proliferative activity in human astrocytic gliomas. *Pathol Res Pr*. 1998;194(3):141-7.
56. Hoi Sang U, Espiritu OD, Kelley PY, Klauber MR, Hatton JD. The role of the epidermal growth factor receptor in human gliomas: II. The control of glial process extension and the expression of glial fibrillary acidic protein. *J Neurosurg*. 1995;82(5):847-57. doi:10.3171/jns.1995.82.5.0847.
57. M M, JS K, PJ K, T Y. Transforming growth factor-alpha, epidermal growth factor receptor, and proliferating potential in benign and malignant gliomas. *J Neurosurg*. 1991;75(1):97-102.
58. Yamada N, Kato M, Yamashita H, et al. Enhanced expression of transforming growth factor-β and its type-I and type-II receptors in human glioblastoma. *Int J Cancer*. 1995;62:386–392.
59. R H, K K. Transforming growth factor beta1-specific binding proteins on human vascular endothelial cells. *Exp Cell Res*. 1992;201(1):119-25.
60. Hatanpaa K, Burma S. Epidermal growth factor receptor in glioma: signal transduction, neuropathology, imaging, and radioresistance. *Neoplasia (New York, NY ...*. 2010;12(9):675-684. doi:10.1593/neo.10688.
61. Heimberger AB, Suki D, Yang D, Shi W, Aldape K. The natural history of EGFR and EGFRvIII in glioblastoma patients. *J Transl Med*. 2005;3:38. doi:10.1186/1479-5876-3-38.
62. Sciences M, Group C, Hospital S. Amplified and rearranged epidermal growth factor receptor genes in human glioblastomas reveal deletions of sequences encoding portions of the N- and/or C-terminal tails. 1992;89(May):4309-4313.
63. L M, Z F, NH A, KK A. Epidermal growth factor receptor and tumor response to radiation: in vivo preclinical studies. *Int J Radiat Oncol Biol Phys*. 2004;58(3):966-71.
64. Johannessen C. The NF1 tumor suppressor critically regulates TSC2 and mTOR. *... Sci ...*. 2005;(24):13367-13371. Available at: http://www.pnas.org/content/102/24/8573.short. Accessed May 8, 2014.
65. FJ R, A P, DH G, et al. Gliomas in neurofibromatosis type 1: a clinicopathologic study of 100 patients. *J Neuropathol Exp Neurol*. 2008;67(3):240-9.
66. Chalhoub N, Baker S. PTEN and the PI3-kinase pathway in cancer. *Annu Rev Pathol*. 2009;(1):127-150. doi:10.1146/annurev.pathol.4.110807.092311.PTEN.
67. Koul D. PTEN signaling pathways in glioblastoma. *Cancer Biol Ther*. 2008;7(9):1321-5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18836294.
68. Nakada M, Kita D, Watanabe T, et al. Aberrant signaling pathways in glioma. *Cancers (Basel)*. 2011;3(3):3242-78. doi:10.3390/cancers3033242.
69. Sherr CJ, McCormick F. The RB and p53 pathways in cancer. *Cancer Cell*. 2002;2(2):103-12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12204530.
70. Ohgaki H, Kleihues P. Genetic pathways to primary and secondary glioblastoma. *Am J Pathol*. 2007;170(5):1445-53. doi:10.2353/ajpath.2007.070011.

71. Nakamura M, Watanabe T, Klangby U, et al. p14ARF deletion and methylation in genetic pathways to glioblastomas. *Brain Pathol*. 2001;11(2):159-68. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11303791.
72. Furnari FB, Fenton T, Bachoo RM, et al. Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev*. 2007;21(21):2683-710. doi:10.1101/gad.1596707.
73. James CD. P53 in Malignant Glioma: 20 Years Later and Still Much To Learn. *Neuro Oncol*. 2010;12(5):421. doi:10.1093/neuonc/noq037.
74. MH K, SN J, KH.Vousden. Regulation of p53 stability by Mdm2. *Nature*. 1997;387(6630):299-303.
75. Toledo F, Wahl GM. MDM2 and MDM4: p53 regulators as targets in anticancer therapy. *Int J Biochem Cell Biol*. 2007;39:1476–1482.
76. Y Z, Y X, WG Y. ARF promotes MDM2 degradation and stabilizes p53: ARF-INK4a locus deletion impairs both the Rb and p53 tumor suppression pathways. *Cell*. 1998;92(6):725-34.
77. Bagchi A, Papazoglu C, Wu Y, et al. CHD5 is a tumor suppressor at human 1p36. *Cell*. 2007;128(3):459-75. doi:10.1016/j.cell.2006.11.052.
78. Z Z, H W, M L, ER R, S A, R Z. Stabilization of E2F1 protein by MDM2 through the E2F1 ubiquitination pathway. *Oncogene*. 2005;24(48):7238-47.
79. Nakamura M, Yang F, Fujisawa H, Yonekawa Y, Kleihues P, Ohgaki H. Loss of heterozygosity on chromosome 19 in secondary glioblastomas. *J Neuropathol Exp Neurol*. 2000;59(6):539-43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10850866.
80. ME H, U S, A U, VI V. Uniform MDM2 overexpression in a panel of glioblastoma multiforme cell lines with divergent EGFR and p53 expression status. *Anticancer Res*. 2006;26(6B):4191-4.
81. Riemenschneider MJ, Büschges R, Wolter M, Bu R, Reifenberger J, Bostro J. Amplification and Overexpression of the MDM4 (MDMX) Gene from 1q32 in a Subset of Malignant Gliomas without TP53 Mutation or MDM2 Amplification Subset of Malignant Gliomas without TP53 Mutation or MDM2 Amplification 1. 1999;4:6091-6096.
82. Zheng H, Ying H, Yan H, et al. p53 and Pten control neural and glioma stem/progenitor cell renewal and differentiation. *Nature*. 2008;455(7216):1129-33. doi:10.1038/nature07443.
83. Engelman J a, Luo J, Cantley LC. The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. *Nat Rev Genet*. 2006;7(8):606-19. doi:10.1038/nrg1879.
84. Carracedo a, Pandolfi PP. The PTEN-PI3K pathway: of feedbacks and cross-talks. *Oncogene*. 2008;27(41):5527-41. doi:10.1038/onc.2008.247.
85. EC H, J C, C D, L S, RE S, GN F. Combined activation of Ras and Akt in neural progenitors induces glioblastoma formation in mice. *Nat Genet*. 2000;25(1):55-7.
86. DA G, DM S. Defining the role of mTOR in cancer. *Cancer Cell*. 2007;12(1):9-22.
87. P W, YZ H. PI3K/Akt/mTOR pathway inhibitors in cancer: a perspective on clinical progress. *Curr Med Chem*. 2010;17(35):4326-41.
88. Carracedo A, Ma L. Inhibition of mTORC1 leads to MAPK pathway activation through a PI3K-dependent feedback loop in human cancer. *J Clin …*. 2008;118(9):3065. doi:10.1172/JCI34739.tion.
89. Matthews K. Akt activation leads to poor prognosis in postoperative breast cancer patients. *Nat Clin Pract Oncol*. 2006;3:64-65.
90. H Z, N B, E W, et al. PDGFRs are critical for PI3K/Akt activation and negatively regulated by mTOR. *J Clin Invest*. 2007;117(3):730-8.
91. LS H, GM F, A G, et al. The TSC1-2 tumor suppressor controls insulin-PI3K signaling via regulation of IRS proteins. *J Cell Biol*. 2004;166(2):213-23.
92. SI W, J P, J L, et al. Somatic Mutations of PTEN in Glioblastoma Multiforme. *Cancer Res*. 1997;57(19):4183-6.
93. Quayle SN, Lee JY, Cheung LWT, et al. Somatic mutations of PIK3R1 promote gliomagenesis. *PLoS One*. 2012;7(11):e49466. doi:10.1371/journal.pone.0049466.
94. Carpten JD, Faber AL, Horn C, et al. A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. *Nature*. 2007;448(7152):439-44. doi:10.1038/nature05933.
95. Bos J. Ras oncogenes in human cancer: a review. *Cancer Res*. 1989:4682-4689. Available at: http://cancerres.aacrjournals.org/content/49/17/4682.short. Accessed May 9, 2014.
96. B A, A J, J L, Al E. *Molecular Biology of the Cell. 4th edition.*; 2002:Integrins. Available at: http://www.ncbi.nlm.nih.gov/books/NBK26867/.
97. H L, A B, SL Z. *Molecular Cell Biology. 4th edition.*; 2000:Section 20.5, MAP Kinase Pathways.
98. Srivastava A. Analysis of RAS Subfamily involved in Cancer via in silico Approach and Designing of Potent Drug against Cancer. *Helix*. 2013;1:226-230.

99. Fukushima Y, Ohnishi T, Arita N, Hayakawa T, Sekiguchi K. Integrin alpha3beta1-mediated interaction with laminin-5 stimulates adhesion, migration and invasion of malignant glioma cells. *Int J Cancer*. 1998;76(1):63-72. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9533763.
100. EC H, J C, C D, L S, RE S, GN F. Combined activation of Ras and Akt in neural progenitors induces glioblastoma formation in mice. *Nat Genet*. 2000;25(1):55-7.
101. Filipp F, Scott D, Ronai Z, Osterman A, Smith J. Reverse TCA cycle flux through isocitrate dehydrogenases 1 and 2 is required for lipogenesis in hypoxic melanoma cells. *Pigment Cell Melanoma Res*. 2012;25(3):375-83.
102. Reitman ZJ, Yan H. Isocitrate dehydrogenase 1 and 2 mutations in cancer: alterations at a crossroads of cellular metabolism. *J Natl Cancer Inst*. 2010;102(13):932-41. doi:10.1093/jnci/djq187.
103. KE Y, DP S. Cancer-associated isocitrate dehydrogenase mutations. *Oncologist*. 2012;17(1):5-8.
104. Yan H, Parsons D, Jin G. IDH1 and IDH2 mutations in gliomas. ... *Engl J* .... 2009;360(8):765-773. doi:10.1056/NEJMoa0808710.IDH1.
105. Schaap F, French P, Bovée J. Mutations in the Isocitrate Dehydrogenase Genes IDH1 and IDH2 in Tumors. *Adv Anat Pathol*. 2013;20(1):32-8.
106. B P, H Z, H Q, et al. A tale of two subunits: how the neomorphic R132H IDH1 mutation enhances production of αHG. *Biochemistry*. 2011;50(21):4804-12.
107. Zhao S, Lin Y, Xu W, Jiang W, Zha Z, Wang P. Glioma-derived mutations in IDH1 dominantly inhibit IDH1 catalytic activity and induce HIF-1α. *Science (80-)*. 2009;324(5924):261-265. doi:10.1126/science.1170944.Glioma-Derived.
108. JA L, RE L, P K, et al. (R)-2-hydroxyglutarate is sufficient to promote leukemogenesis and its effects are reversible. *Science (80-)*. 2013;339(6127):1621-5.
109. Ye D, Ma S, Xiong Y, Guan K-L. R-2-hydroxyglutarate as the key effector of IDH mutations promoting oncogenesis. *Cancer Cell*. 2013;23(3):274-6. doi:10.1016/j.ccr.2013.03.005.
110. Xu W, Yang H, Liu Y, et al. Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. *Cancer Cell*. 2011;19(1):17-30. doi:10.1016/j.ccr.2010.12.014.
111. Semenza GL. Targeting HIF-1 for cancer therapy. *Nat Rev Cancer*. 2003;3:721-732.
112. Blouw B, Song H, Tihan T, et al. The hypoxic response of tumors is dependent on their microenvironment. *Cancer Cell*. 2003;4(2):133-46. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12957288.
113. Williams SC, Karajannis M a, Chiriboga L, Golfinos JG, von Deimling A, Zagzag D. R132H-mutation of isocitrate dehydrogenase-1 is not sufficient for HIF-1α upregulation in adult glioma. *Acta Neuropathol*. 2011;121(2):279-81. doi:10.1007/s00401-010-0790-y.
114. Sasaki M, Knobbe CB, Munger JC, et al. IDH1(R132H) mutation increases murine haematopoietic progenitors and alters epigenetics. *Nature*. 2012;488(7413):656-9. doi:10.1038/nature11323.
115. Koivunen P, Lee S, Duncan CG, et al. Transformation by the (R)-enantiomer of 2-hydroxyglutarate linked to EGLN activation. *Nature*. 2012;483(7390):484-8. doi:10.1038/nature10898.
116. Querbes W, Bogorad RL, Moslehi J, et al. Treatment of erythropoietin deficiency in mice with systemically administered siRNA. *Blood*. 2012;120(9):1916-22. doi:10.1182/blood-2012-04-423715.
117. S I, L S, Q D, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*. 2011;333(6047):1300-3.
118. Ito S, D'Alessio AC, Taranova O V, Hong K, Sowers LC, Zhang Y. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*. 2010;466(7310):1129-33. doi:10.1038/nature09303.
119. Kim Y-H, Pierscianek D, Mittelbronn M, et al. TET2 promoter methylation in low-grade diffuse gliomas lacking IDH1/2 mutations. *J Clin Pathol*. 2011;64(10):850-2. doi:10.1136/jclinpath-2011-200133.
120. S S, JM X, DM S, et al. p53 interaction with JMJD3 results in its nuclear distribution during mouse neural stem cell differentiation. *PLoS One*. 2011;6(3):e18421.
121. Lu C, Ward PS, Kapoor GS, et al. IDH mutation impairs histone demethylation and results in a block to cell differentiation. *Nature*. 2012;483(7390):474-8. doi:10.1038/nature10860.
122. A L, CG da S, GC F, et al. Mitochondrial energy metabolism is markedly impaired by D-2-hydroxyglutaric acid in rat tissues. *Mol Genet Metab*. 2005;86(1-2):188-99.

123. Kim SY, Lee SM, Tak JK, Choi KS, Kwon TK, Park J-W. Regulation of singlet oxygen-induced apoptosis by cytosolic NADP+-dependent isocitrate dehydrogenase. *Mol Cell Biochem*. 2007;302(1-2):27-34. doi:10.1007/s11010-007-9421-x.
124. Balss J, Meyer J, Mueller W, Korshunov A, Hartmann C, von Deimling A. Analysis of the IDH1 codon 132 mutation in brain tumors. *Acta Neuropathol*. 2008;116(6):597-602. doi:10.1007/s00401-008-0455-2.
125. Nobusawa S, Watanabe T, Kleihues P, Ohgaki H. IDH1 mutations as molecular signature and predictive factor of secondary glioblastomas. *Clin Cancer Res*. 2009;15(19):6002-7. doi:10.1158/1078-0432.CCR-09-0715.
126. SJ B, SH K, SK K, JH P, SH. P. Distinct genetic alterations in pediatric glioblastomas. *Childs Nerv Syst*. 2012;28(7):1025-32.
127. M M, A P, V C, et al. IDH1 and IDH2 mutations, immunohistochemistry and associations in a series of brain tumors. *J Neurooncol*. 2011;105(2):345-57.
128. Losman J, Kaelin W. What a difference a hydroxyl makes: mutant IDH,(R)-2-hydroxyglutarate, and cancer. *Genes Dev*. 2013;27(8):836-852. doi:10.1101/gad.217406.113.whether.
129. Noushmehr H, Weisenberger DJ, Diefes K, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*. 2010;17(5):510-22. doi:10.1016/j.ccr.2010.03.017.
130. Christensen BC, Smith A a, Zheng S, et al. DNA methylation, isocitrate dehydrogenase mutation, and survival in glioma. *J Natl Cancer Inst*. 2011;103(2):143-53. doi:10.1093/jnci/djq497.
131. Turcan S, Rohle D, Goenka A, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature*. 2012;483(7390):479-83. doi:10.1038/nature10866.
132. H I, S M, H K. Can systems biology understand pathway activation? Gene expression signatures as surrogate markers for understanding the complexity of pathway activation. *Curr Genomics*. 2008;9(5):349-60.
133. Veer L van't, Dai H, Vijver M Van De. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415(345). Available at: http://www.nature.com/nature/journal/v415/n6871/abs/415530a.html. Accessed May 8, 2014.
134. Yeoh E-J, Ross ME, Shurtleff S a, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*. 2002;1(2):133-43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12086872.
135. Chibon F. Cancer gene expression signatures - the rise and fall? *Eur J Cancer*. 2013;49(8):2000-9. doi:10.1016/j.ejca.2013.02.021.
136. J.Sheskin D. *Handbook of Parametric and Nonparametric Statistical Procedures, 2nd Edition*.; 2000.
137. Rupert G, Miller J. *Simultaneous Statistical Inference*.; 1981.
138. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc*. 1995;1:289-300.
139. H. WP. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*.; 1993:42-43.
140. Good P. *Permutation, Parametric and Bootstrap Tests of Hypotheses 3rd Edition*.; 2005:57-62.
141. García V, Mollineda JSSRA, Sotoca RAJM. The class imbalance problem in pattern classification and learning.
142. Japkowicz N. Learning from Imbalanced Data Sets: A Comparison of Various Strategies. *Work Notes AAAI'00 Work Learn from Imbalanced Data Sets*. 2000:10–15.
143. Zhou Z, Member S, Liu X. Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. :1-14.
144. Wong G. *DNA microarray data analysis*.; 2005:15-19.
145. Wong G. *DNA microarray data analysis*.; 2005:48-49.
146. 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. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=100354&tool=pmcentrez&rendertype=abstract.
147. Hopcroft J, Kannan R. *Foundations of Data Science*.; 2011.
148. Harris M a, Clark J, Ireland a, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 2004;32(Database issue):D258-61. doi:10.1093/nar/gkh036.
149. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27-30. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract.
150. Gentleman R, Irizarry RA, Carey VJ, Dudoit S, Huber W. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*.; 2005.
151. JP M, S. R. Software and tools for microarray data analysis. *Methods Mol Biol*. 2011;784:41-53.
152. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001;98(9):5116-21. doi:10.1073/pnas.091062498.
153. Huang DW, Sherman BT, Lempicki R a. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211.
154. Shannon P, Markiel A, Ozier O, et al. Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. 2003;(Karp 2001):2498-2504. doi:10.1101/gr.1239303.metabolite.
155. Shannon PT, Grimes M, Kutlu B, Bot JJ, Galas DJ. RCytoscape: tools for exploratory network analysis. *BMC Bioinformatics*. 2013;14(1):217. doi:10.1186/1471-2105-14-217.
156. Scholtens D. *Graph Basics in R and Bioconductor*.(2007).
157. A Tutorial on Clustering Algorithms. Available at: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/index.html.
158. Yang ZR. *Machine Learning Approaches to Bioinformatics*.; 2010:55-58.
159. Yang ZR. Machine Learning Approaches to Bioinformatics. 2010:59.
160. Fraley C, Raftery AE. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *J Am Stat Assoc*. 2002;97(458):611-631. doi:10.1198/016214502760047131.
161. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika*. 1965;52(3/4):591-611.
162. Adrian A, Rahnenfuhrer J. *topGO: Enrichment analysis for Gene Ontology*.(2010).
163. Zhang J, Wiemann S. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics*. 2009;25(11):1470-1.
164. Swinnen J V, Brusselmans K, Verhoeven G. Increased lipogenesis in cancer cells: new players, novel targets. *Curr Opin Clin Nutr Metab Care*. 2006;9(4):358-65. doi:10.1097/01.mco.0000232894.28674.30.
165. Srivastava NK, Pradhan S, Gowda G a N, Kumar R. In vitro, high-resolution 1H and 31P NMR based analysis of the lipid components in the tissue, serum, and CSF of the patients with primary brain tumors: one possible diagnostic view. *NMR Biomed*. 2010;23(2):113-22. doi:10.1002/nbm.1427.
166. Guo D, Bell E, Chakravarti A. Lipid metabolism emerges as a promising target for malignant glioma therapy. *CNS Oncol*. 2013;2(3):289-299. doi:10.2217/cns.13.20.Lipid.
167. Deliang Guo, Robert M. Prins, [...] and PSM. EGFR signaling through an Akt-SREBP-1-dependent, rapamycin-resistant pathway sensitizes glioblastomas to antilipogenic therapy. *Sci Signal*. 2009.
168. Horton JD, Goldstein JL, Brown MS. Critical review SREBPs : activators of the complete program of cholesterol and fatty acid synthesis in the liver. 2002;109(9):1125-1131. doi:10.1172/JCI200215593.Lipid.
169. Zhao W, Kridel S, Thorburn a, et al. Fatty acid synthase: a novel target for antiglioma therapy. *Br J Cancer*. 2006;95(7):869-78. doi:10.1038/sj.bjc.6603350.
170. Dolce V, Rita Cappello A. Glycerophospholipid Synthesis as a Novel Drug Target Against Cancer. *Curr Mol Pharmacol*. 2011;4:167-175.
171. Rong Y, Post DE, Pieper RO, Durden DL, Meir EG Van. PTEN and Hypoxia Regulate Tissue Factor Expression and Plasma Coagulation by Glioblastoma and Plasma Coagulation by Glioblastoma. 2005:1406-1413.
172. Meehan2 DLOKR, Zacharski LR. The Coagulation System as a Target for the Treatment of Human Gliomas. *Semin Thromb Hemost*. 2002;28:19-28.
173. Perry J. Thromboembolic disease in patients with high-grade glioma. *Neuro Oncol*. 2012:73-80. Available at: http://neuro-oncologyoxfordjournals.newsofmedical.com/content/14/suppl_4/iv73.short. Accessed May 6, 2014.
174. Yamamoto H, Swoger J, Greene S, et al. Beta1,6-N-acetylglucosamine-bearing N-glycans in human gliomas: implications for a role in regulating invasivity. *Cancer Res*. 2000;60(1):134-42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10646865.

175. Hwang M-S, Baek W-K. Glucosamine induces autophagic cell death through the stimulation of ER stress in human glioma cancer cells. *Biochem Biophys Res Commun*. 2010;399(1):111-6. doi:10.1016/j.bbrc.2010.07.050.
176. O N, Uchimura K, Muramatsu H, et al. CARBOHYDRATES , LIPIDS , AND OTHER NATURAL PRODUCTS : Molecular Cloning and Characterization of Molecular Cloning and Characterization of an N -Acetylglucosamine-6- O -sulfotransferase *. 1998.
177. Hayatsu N, Ogasawara S, Kaneko MK, Kato Y, Narimatsu H. Expression of highly sulfated keratan sulfate synthesized in human glioblastoma cells. *Biochem Biophys Res Commun*. 2008;368(2):217-22. doi:10.1016/j.bbrc.2008.01.058.
178. Reyes SB, Narayanan AS, Lee HS, et al. αvβ8 integrin interacts with RhoGDI1 to regulate Rac1 and Cdc42 activation and drive glioblastoma cell invasion. *Mol Biol Cell*. 2013;24(4):474-82. doi:10.1091/mbc.E12-07-0521.
179. Ivetic A, Ridley AJ. Ezrin/radixin/moesin proteins and Rho GTPase signalling in leucocytes. *Immunology*. 2004;112(2):165-76. doi:10.1111/j.1365-2567.2004.01882.x.
180. Hall A. The cytoskeleton and cancer. *Cancer Metastasis Rev*. 2009;28(1-2):5-14. doi:10.1007/s10555-008-9166-3.
181. Wu M, Liu D-Y, Yuan X-R, et al. The expression of moesin in astrocytoma: correlation with pathologic grade and poor clinical outcome. *Med Oncol*. 2013;30(1):372. doi:10.1007/s12032-012-0372-z.
182. Solinet S, Mahmud K, Stewman SF, et al. The actin-binding ERM protein Moesin binds to and stabilizes microtubules at the cell cortex. *J Cell Biol*. 2013;202(2):251-60. doi:10.1083/jcb.201304052.
183. Chidambaram A, Fillmore HL, Van Meter TE, Dumur CI, Broaddus WC. Novel report of expression and function of CD97 in malignant gliomas: correlation with Wilms tumor 1 expression and glioma cell invasiveness. *J Neurosurg*. 2012;116(4):843-53. doi:10.3171/2011.11.JNS111455.
184. Safaee M, Clark AJ, Oh MC, et al. Overexpression of CD97 confers an invasive phenotype in glioblastoma cells and is associated with decreased survival of glioblastoma patients. *PLoS One*. 2013;8(4):e62765. doi:10.1371/journal.pone.0062765.
185. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. 2005.

Appendix1

Table A1.1 present the Entrez ID and names of genes in the gene signature

Table A1.1 gene ID and name of genes from the IDH1 gene signature

Gene Symbol	Entrez ID	Gene Name
SLC2A10	81031	solute carrier family 2 (facilitated glucose transporter), member 10
C1orf107	27042	digestive organ expansion factor homolog (zebrafish)
SDF4	51150	stromal cell derived factor 4
HRH1	3269	histamine receptor H1
GALNS	2588	galactosamine (N-acetyl)-6-sulfate sulfatase
MSN	4478	moesin
LDHA	3939	lactate dehydrogenase A
SYNJ2	8871	synaptojanin 2
PLA2G5	5322	phospholipase A2, group V
EFEMP2	30008	EGF containing fibulin-like extracellular matrix protein 2
GPR172A	79581	solute carrier family 52 (riboflavin transporter), member 2
M6PRBP1	10226	perilipin 3
AK3L1	205	adenylate kinase 4
FGF17	8822	fibroblast growth factor 17
OSBPL10	114884	oxysterol binding protein-like 10
ITGB8	3696	integrin, beta 8
CHST2	9435	carbohydrate (N-acetylglucosamine-6-O) sulfotransferase 2
MYO1E	4643	myosin IE
PLAT	5327	plasminogen activator, tissue
CHST7	56548	carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 7
PHLDA3	23612	pleckstrin homology-like domain, family A, member 3
SLC22A18	5002	solute carrier family 22, member 18
FHL2	2274	four and a half LIM domains 2
ALDOA	226	aldolase A, fructose-bisphosphate
ANXA5	308	annexin A5
ACRV1	56	acrosomal vesicle protein 1
BDH1	622	3-hydroxybutyrate dehydrogenase, type 1
ELOVL6	79071	ELOVL fatty acid elongase 6
DUSP5	1847	dual specificity phosphatase 5
SPRY2	10253	sprouty homolog 2 (Drosophila)
MEOX2	4223	mesenchyme homeobox 2
C20orf23	55614	kinesin family member 16B
ARSJ	79642	arylsulfatase family, member J
CXCL14	9547	chemokine (C-X-C motif) ligand 14
MRC2	9902	mannose receptor, C type 2
CD97	976	CD97 molecule
OPLAH	26873	5-oxoprolinase (ATP-hydrolysing)
CYP27A1	1593	cytochrome P450, family 27, subfamily A, polypeptide 1
ZNF492	57615	zinc finger protein 492

ACSL3	2181	acyl-CoA synthetase long-chain family member 3
IMPACT	55364	impact RWD domain protein
TRIP6	7205	thyroid hormone receptor interactor 6
PDGFA	5154	platelet-derived growth factor alpha polypeptide
ELOVL2	54898	ELOVL fatty acid elongase 2
PMP22	5376	peripheral myelin protein 22
PIPOX	51268	pipecolic acid oxidase
STEAP3	55240	STEAP family member 3, metalloreductase
RAB36	9609	RAB36, member RAS oncogene family
MOXD1	26002	monooxygenase, DBH-like 1
CNKSR1	10256	connector enhancer of kinase suppressor of Ras 1

Appendix2

Table A2.1 delineates the definition of the significant enriched GO term, and table A2.2 records the symbols of DE genes annotated.

Table A2.1 significant GO terms and definitions.

GO ID	Definition
GO:0045017	The chemical reactions and pathways resulting in the formation of glycerolipids, any lipid with a glycerol backbone.
GO:0030497	The elongation of a fatty acid chain by the sequential addition of two-carbon units.
GO:0071071	Any process that modulates the frequency, rate or extent of the chemical reactions and pathways resulting in the formation of phospholipids.
GO:0044283	The chemical reactions and pathways resulting in the formation of small molecules, any low molecular weight, monomeric, non-encoded molecule.
GO:0044711	A biosynthetic process - chemical reactions and pathways resulting in the formation of substances - involving a single organism.
GO:0048008	The series of molecular signals generated as a consequence of a platelet-derived growth factor receptor binding to one of its physiological ligands.
GO:0050819	Any process that stops, prevents, or reduces the frequency, rate or extent of coagulation.
GO:0072330	The chemical reactions and pathways resulting in the formation of monocarboxylic acids, any organic acid containing one carboxyl (-COOH) group.
GO:0016053	The chemical reactions and pathways resulting in the formation of organic acids, any acidic compound containing carbon in covalent linkage.
GO:0046394	The chemical reactions and pathways resulting in the formation of carboxylic acids, any organic acid containing one or more carboxyl (-COOH) groups.
GO:0046486	The chemical reactions and pathways involving glycerolipids, any lipid with a glycerol backbone. Diacylglycerol and phosphatidate are key lipid intermediates of glycerolipid biosynthesis.
GO:0006044	The chemical reactions and pathways involving N-acetylglucosamine. The D isomer is a common structural unit of glycoproteins in plants, bacteria and animals; it is often the terminal sugar of an oligosaccharide group of a glycoprotein.
GO:0044255	The chemical reactions and pathways involving lipids, as carried out by individual cells.
GO:0043436	The chemical reactions and pathways involving any oxoacid; an oxoacid is a compound which contains oxygen, at least one other element, and at least one hydrogen bound to oxygen, and which produces a conjugate base by loss of positive hydrogen ion(s) (hydrons).
GO:0006637	The chemical reactions and pathways involving acyl-CoA, any derivative of coenzyme A in which the sulfhydryl group is in thiolester linkage with an acyl group.
GO:0035383	The chemical reactions and pathways involving a thioester, a compound of general formula RC(=O)SR' in which the linking oxygen in an ester is replaced by a sulfur atom. They are the product of esterification between a carboxylic acid and a thiol.
GO:0006082	The chemical reactions and pathways involving organic acids, any acidic compound containing carbon in covalent linkage.
GO:0006633	The chemical reactions and pathways resulting in the formation of a fatty acid, any of the aliphatic monocarboxylic acids that can be liberated by hydrolysis from naturally occurring fats and oils. Fatty acids are predominantly straight-chain acids of 4 to 24 carbon atoms, which may be saturated or unsaturated; branched fatty acids and hydroxy fatty acids also occur, and very long chain acids of over 30 carbons are found in waxes.
GO:0035338	The chemical reactions and pathways resulting in the formation of a long-chain fatty-acyl-CoA any derivative of coenzyme A in which the sulfhydryl group is in

	a thioester linkage with a long-chain fatty-acyl group. Long-chain fatty-acyl-CoAs have chain lengths of C13 or more.
GO:0035336	The chemical reactions and pathways involving long-chain fatty-acyl-CoAs, any derivative of coenzyme A in which the sulfhydryl group is in a thioester linkage with a long-chain fatty-acyl group. Long-chain fatty-acyl-CoAs have chain lengths of C13 or more.
GO:0046949	The chemical reactions and pathways resulting in the formation of a fatty-acyl-CoA, any derivative of coenzyme A in which the sulfhydryl group is in thiolester linkage with a fatty-acyl group.
GO:0035337	The chemical reactions and pathways involving a fatty-acyl-CoA, any derivative of coenzyme A in which the sulfhydryl group is in thiolester linkage with a fatty-acyl group.
GO:1901071	The chemical reactions and pathways involving glucosamine-containing compounds (glucosamines).
GO:0044281	The chemical reactions and pathways involving small molecules, any low molecular weight, monomeric, non-encoded molecule.
GO:0050818	Any process that modulates the frequency, rate or extent of coagulation, the process in which a fluid solution, or part of it, changes into a solid or semisolid mass.
GO:0090407	The chemical reactions and pathways resulting in the biosynthesis of deoxyribose phosphate, the phosphorylated sugar 2-deoxy-erythro-pentose.
GO:0010741	Any process that decreases the rate, frequency or extent of a series of reactions, mediated by protein kinases, which occurs as a result of a single trigger reaction or compound.
GO:0006629	The chemical reactions and pathways involving lipids, compounds soluble in an organic solvent but not, or sparingly, in an aqueous solvent. Includes fatty acids; neutral fats, other fatty-acid esters, and soaps; long-chain (fatty) alcohols and waxes; sphingoids and other long-chain bases; glycolipids, phospholipids and sphingolipids; and carotenes, polyprenols, sterols, terpenes and other isoprenoids.
GO:0046474	The chemical reactions and pathways resulting in the formation of glycerophospholipids, any derivative of glycerophosphate that contains at least one O-acyl, O-alkyl, or O-alkenyl group attached to the glycerol residue.
GO:0044710	A metabolic process - chemical reactions and pathways, including anabolism and catabolism, by which living organisms transform chemical substances - which involves a single organism.
GO:0008543	The series of molecular signals generated as a consequence of a fibroblast growth factor receptor binding to one of its physiological ligands.
GO:0008610	The chemical reactions and pathways resulting in the formation of lipids, compounds soluble in an organic solvent but not, or sparingly, in an aqueous solvent.
GO:0008654	The chemical reactions and pathways resulting in the formation of phospholipids, any lipid containing phosphoric acid as a mono- or diester.
GO:0051896	Any process that modulates the frequency, rate or extent of the protein kinase B signaling cascade, a series of reactions mediated by the intracellular serine/threonine kinase protein kinase B.
GO:0006040	The chemical reactions and pathways involving any amino sugar, sugars containing an amino group in place of a hydroxyl group.
GO:0042339	The chemical reactions and pathways involving keratan sulfate, a glycosaminoglycan with repeat units consisting of beta-1,4-linked D-galactopyranosyl-beta-(1,4)-N-acetyl-D-glucosamine 6-sulfate and with variable amounts of fucose, sialic acid and mannose units; keratan sulfate chains are covalently linked by a glycosidic attachment through the trisaccharide galactosyl-galactosyl-xylose to peptidyl-threonine or serine residues.

Table A2.2 significant GO terms and annotated DE Genes

GO ID	DE genes
GO:0045017	"SYNJ2" "PLA2G5" "ELOVL6" "ACSL3" "PDGFA" "ELOVL2"
GO:0030497	"ELOVL6" "ELOVL2"
GO:0071071	"ACSL3" "PDGFA"
GO:0044283	"HRH1" "PLA2G5" "BDH1" "ELOVL6" "OPLAH" "CYP27A1" "ACSL3" "ELOVL2"
GO:0044711	"HRH1" "PLA2G5" "BDH1" "ELOVL6" "OPLAH" "CYP27A1" "ACSL3" "ELOVL2"
GO:0048008	"MYO1E" "PLAT" "PDGFA"
GO:0050819	"PLAT" "ANXA5" "PDGFA"
GO:0072330	"PLA2G5" "ELOVL6" "CYP27A1" "ACSL3" "ELOVL2"
GO:0016053	"PLA2G5" "ELOVL6" "OPLAH" "CYP27A1" "ACSL3" "ELOVL2"
GO:0046394	"PLA2G5" "ELOVL6" "OPLAH" "CYP27A1" "ACSL3" "ELOVL2"
GO:0046486	"SYNJ2" "PLA2G5" "ELOVL6" "ACSL3" "PDGFA" "ELOVL2"
GO:0006044	"CHST2" "CHST7"
GO:0044255	"SYNJ2" "PLA2G5" "ITGB8" "FHL2" "BDH1" "ELOVL6" "ARSJ" "ACSL3" "PDGFA" "ELOVL2"
GO:0043436	"GALNS" "LDHA" "PLA2G5" "CHST2" "CHST7" "ELOVL6" "OPLAH" "CYP27A1" "ACSL3" "ELOVL2" "PIPOX"
GO:0006637	"ELOVL6" "ELOVL2" "PIPOX"
GO:0035383	"ELOVL6" "ELOVL2" "PIPOX"
GO:0006082	"GALNS" "LDHA" "PLA2G5" "CHST2" "CHST7" "ELOVL6" "OPLAH" "CYP27A1" "ACSL3" "ELOVL2" "PIPOX"
GO:0006633	"PLA2G5" "ELOVL6" "ACSL3" "ELOVL2"
GO:0035338	"ELOVL6" "ELOVL2"
GO:0035336	"ELOVL6" "ELOVL2"
GO:0046949	"ELOVL6" "ELOVL2"
GO:0035337	"ELOVL6" "ELOVL2"
GO:1901071	"CHST2" "CHST7"
GO:0044281	"HRH1" "GALNS" "LDHA" "SYNJ2" "PLA2G5" "AK3L1" "CHST2" "CHST7" "FHL2" "ALDOA" "BDH1" "ELOVL6" "SPRY2" "ARSJ" "OPLAH" "CYP27A1" "ACSL3" "ELOVL2" "PIPOX"
GO:0050818	"PLAT" "ANXA5" "PDGFA"
GO:0090407	"HRH1" "SYNJ2" "PLA2G5" "AK3L1" "ALDOA" "ACSL3" "PDGFA"
GO:0010741	"PHLDA3" "DUSP5" "SPRY2" "TRIP6"
GO:0006629	"SYNJ2" "PLA2G5" "ITGB8" "FHL2" "BDH1" "ELOVL6" "ARSJ" "CYP27A1" "ACSL3" "PDGFA" "ELOVL2"
GO:0046474	"SYNJ2" "PLA2G5" "ACSL3" "PDGFA"
GO:0044710	"HRH1" "GALNS" "LDHA" "SYNJ2" "PLA2G5" "AK3L1" "ITGB8" "CHST2" "CHST7" "FHL2" "ALDOA" "BDH1" "ELOVL6" "SPRY2" "ARSJ" "OPLAH" "CYP27A1" "ACSL3" "PDGFA" "ELOVL2" "PIPOX"
GO:0008543	"FGF17" "SPRY2" "C20orf23" "PDGFA"
GO:0008610	"SYNJ2" "PLA2G5" "ELOVL6" "CYP27A1" "ACSL3" "PDGFA" "ELOVL2"
GO:0008654	"SYNJ2" "PLA2G5" "ACSL3" "PDGFA"
GO:0051896	"PHLDA3" "SPRY2" "PDGFA"
GO:0006040	"CHST2" "CHST7"
GO:0042339	"GALNS" "CHST2"