IDENTIFICATION AND CHARACTERIZATION OF GENE AND MICRORNA NETWORKS ASSOCIATED WITH CANCER SURVIVAL AND DRUG ABUSE

BY

KRISTIN RENEE DELFINO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Animal Sciences
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Professor Sandra L. Rodriguez-Zas, Chair
Professor Germán A. Bollero
Associate Professor Juan J. Loor
Professor Romana A. Nowak

Abstract

The study of the dysregulation of the transcriptome in diseases like cancer and drug abuse can offer insights into preventive and therapeutical remedies, as well as targets for future basic and applied research. The identification of reliable transcriptome biomarkers requires the simultaneous consideration of regulatory and target elements including microRNAs (miRNAs), transcription factors (TFs), and target genes. Previously, there has been limited validation of reported associations between these diseases and miRNAs, TFs, and target mRNA in independent studies. This may be due to several reasons. Few studies simultaneously analyze multiple miRNAs, TFs, and target mRNA. Also, most studies do not consider clinical or cohort-dependent factors when characterizing the associations between the transcriptome and disease. Lastly, most transcriptome studies tend to be small, and the individual analysis has limited statistical power to detect accurate and precise associations between transcripts and diseases. This thesis aims to address the previous limitations and identify replicable biomarkers of cancer and drug abuse.

Functional and network analyses were performed to identify and study targets of microRNA biomarkers associated with glioblastoma multiforme survival within and across race, gender, recurrence, and therapy cohorts. A Cox survival model was applied to profiles from 253 individuals, 534 microRNAs, and the results were confirmed using cross-validation, discriminant analyses, and cross-study comparisons. All 45 microRNAs revealed were confirmed in independent cancer studies, and 25 of those were further confirmed in glioblastoma studies. Thirty-nine and six microRNAs were associated with one and multiple glioblastoma survival indicators, respectively. Nineteen and 26 microRNAs exhibited cohort-dependent and independent associations with glioblastoma, respectively.

An approach integrating survival analysis, feature selection, and regulatory network visualization was used to identify reliable biomarkers of ovarian cancer survival and recurrence. Expression profiles of 799 miRNAs, 17,814 TFs and target genes and cohort clinical records on 272 patients diagnosed with ovarian cancer were simultaneously considered and results were validated on an independent group of 146 patients. This study confirmed 19 miRNAs previously associated with ovarian cancer and identified two miRNAs that have previously been associated with other cancer types. In total, the expression of 838 and 734 target genes and 12 and eight TFs were associated (FDR-adjusted P-value <0.05) with ovarian cancer survival and recurrence, respectively. The simultaneous analysis of co-expression profiles along with consideration of clinical characteristics of patients allowed reliable microRNA-transcription factor-target gene networks associated with ovarian cancer survival to be inferred.

Illicit drug exposure brings about changes in the brain transcriptome that result in the dysregulation of pathways. To detect the progression of drug exposure pathways, meta-analysis of five individual microarray experiments measuring gene expression in the brain of mice under acute and chronic drug exposure was performed. Functional analysis and network visualization offered insights into the network changes across drug exposure levels. Meta-analyses uncovered 263 and 2,641 genes differentially expression (FDR-adjusted P-value <0.1) between control and acute and chronic exposure, respectively. Individual genes in these processes have been previously associated with drug exposure and reward-dependent behaviors. The MAPK signaling pathway and the molecular functions of protein dimerization and leucine zipper transcription factor were enriched in response to acute exposure. This study was able to detect the progression of drug exposure pathways using meta, functional, and network analyses.

Table of Contents

List of Figures

List of Tables

List of Tables (con't)

Table **Page**

Chapter IV (con't)

List of Supplementary Figures

CHAPTER I: Literature Review

Gene expression, microRNAs, and transcription factors

Gene transcription and factors influencing transcript levels

Gene expression refers to the production of a protein or functional RNA from a gene. Several steps are involved in the process of transcribing DNA into messenger RNA (mRNA) and of translating the latter into proteins. The first major step in gene expression is transcription. Transcription is a process in which one DNA strand is used as a template to synthesize a complementary RNA (1). The DNA strand serving as the template is often referred to as the template strand, while the other DNA strand is termed the non-template or coding strand. Both the DNA coding strand and the RNA strand are complementary to the template strand and thus, have the same sequences except the Thymines in the DNA coding strand are replaced by the Uracils in the RNA strand (2). Structural changes in the chromatin are required to initiate transcription and elongation, and the resulted primary transcript is processed and transported from the nucleus to the cytoplasm.

Growth of a nucleic acid strand is always in the 5' to 3' direction (3). Enzymes, termed polymerases, are used to catalyze the synthesis of nucleic acid strands. Initiation of transcription on a chromatin template requires the enzyme RNA polymerase to bind at the promoter. Eukaryotic DNA is then unwound by a specific transcription factor, and RNA synthesis begins based on the sequence of the DNA template stand (2, 4).

There are three classes of RNA polymerases in eukaryotic cells that are specialized for the transcription of particular sets of genes. RNA polymerase I (RNA Pol I) is located in the nucleus and synthesizes ribosomal RNA (rRNA); RNA polymerase II (RNA Pol II) is responsible for mRNA synthesis and is involved in the transcription of all protein genes; and

RNA polymerase III (RNA Pol III) is located outside the nucleus and synthesizes transfer RNA (tRNA) and other small RNAs (2). None of the three RNA polymerases recognize their promoters directly. These polymerases depend on other proteins, or transcription factors, to recognize the promoter and bind to it. The binding region contains all binding sites necessary for RNA polymerase to bind and function to initiate transcription by guiding RNA Pol II to the start point. After the first bound is synthesized, RNA Pol II is released from the promoter to start transcript elongation (4). The capping of the 5' end and addition of a poly (A) tail added to the 3' end must occur for mRNA export. The majority of pre-mRNAs consist of coding and non-coding sections, termed exons and introns (1). During RNA splicing, a RNA-protein catalytical complex known as a spliceosome removes the introns and joins the exons together. Alternative splicing may also occur where some introns or exons may be either removed or retained in mature mRNA which creates a series of different transcripts originating from a single gene (3). Once the non-coding sections are removed and exons are spliced together, the mature mRNA is then transported from the nucleus to the cytoplasm where protein synthesis occurs (2).

Introduction to transcription factors

Transcription factors (TFs) are proteins needed for the initiation of transcription that are not part of RNA polymerase (3). Transcription factors create a structure at the promoter to provide the target that is recognized by the RNA polymerase. The RNA Pol II is associated with six general transcription factors, termed TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, where TF stands for transcription factor and II for the RNA Pol II (1, 2). In eukaryotes, promoters are very diverse, but around 20% of them are characterized by a sequence of seven bases (TATAAAA) termed the TATA box. TFIID consists of TBP (TATA-box binding protein) and

TAFs (TBP associated factors) (2). The role of the TBP is to bind the core promoter, the binding region containing all binding sites necessary for RNA polymerase to bind and function to initiate transcription by guiding RNA Pol II to the start point.

Transcription factors bind to *cis*-regulator DNA sequences and are responsible for either positively or negatively influencing the transcription of specific genes, essentially determining whether a particular gene will be turned on or off (5). Roughly 8% of genes in the human genome encode TFs (6). Transcription factors are classified according to their conserved sequences as well as their three-dimensional protein structure, including basic helix-turn-helix, helix-loop-helix, and zinc finger proteins (7).

Transcription factors influencing gene expression

The gene regulatory region that the TF targets can span dozens of kilobases (5). Regulation of gene expression by TFs requires the coordinated interactions of multiple proteins (7). Transcription factors have defined DNA-binding domains with up to 10^6-fold higher affinity for their target sequences than for the remainder of the DNA strand (7).

Besides regulation at the level of gene expression, TFs are also regulated via posttranslational events including protein phosphorylation, processing, and localization (5). Transcription factors are vital for many important biological processes (6). Examples include regulating muscle differentiation in embryonic development (myogenic differentiation; MYOD), helping the kidneys recover from water dehydration (nuclear receptor subfamily 2, group C, member 2; NR2C2), andinstigating oncogenesis (v-myc myelocytomatosis viral oncogene homolog avian MYC). Expression of many TFs is subject to microRNA regulation and the specific expression profiles of microRNAs are brought about in large part by TF-dependent

transcriptional control mechanisms, hence microRNAs and TFs are linked to one another in gene regulatory networks.

Introduction to microRNAs

Introduced in the early nineties, microRNAs (miRNAs) are small, non-coding RNA molecules ~22 nucleotides (nt) in length (**Figure 1.1**) that are highly conserved among species (8). Over 21,264 miRNAs from over 193 species are registered in the miRBase database (http://www.mirbase.org/) as of August 2012, and bioinformatics studies predict that up to 1,500 miRNAs may exist in humans alone (9). It is estimated that the transcription of about 1/3 of human genes is regulated (enhanced or weakened) by miRNA (8). The accepted nomenclature of miRNAs starts with miR- followed by a number identifier (e.g. miR-125), although there are a few exceptions. To distinguish miRNAs within a group of similar sequences, an additional letter following the miRNA number is used (e.g. miR-125b). In addition, miRNA that have identical mature sequences may be coded at several genomic loci by different precursor sequences. Thus, an additional number is added at the end of the sequence (e.g. miR-125b-1) to further distinguish these events (9).

Based on the genomic distribution, miRNA coding genes can be grouped into two classes, intergenic miRNAs and intragenic miRNAs (10, 11). About 42% of miRNAs are intergenic, or miRNA-coding genes located between protein-coding genes (10). Intragenic miRNAs, or miRNA-coding genes located within their host protein-coding genes, can be further subdivided into four subclasses: a) intronic miRNAs, located within introns of their host protein coding genes; b) exonic miRNAs, located within exons of host protein coding genes; c) 3'untranslated region (3'UTR) miRNAs, located within 3'UTR of host protein coding genes; and

d) 5'UTR miRNAs, located within 5'UTR of host protein coding genes (10). The majority of intragenic miRNA are intronic (44%), with the other three subclasses accounting for less than 10% of total miRNA (10).

MicroRNAs influencing gene expression

miRNAs are generated by two mechanisms (**Figure 1.2**). Intergenic miRNA genes are initially transcribed as long transcripts, termed primary miRNAs (pri-miRNAs), by RNA polymerase II or RNA polymerase III (12). These pri-miRNAs are then capped and polyadenylated. Intronic miRNAs, processed by sharing the same promoter and other regulatory elements of their host genes, are first transcribed along with their host genes by RNA polymerase II (10). Pri-miRNAs are processed in the nucleus by the RNase III enzyme Drosha and the double stranded RNA-binding domain protein DiGeorge Syndrome Critical Region 8 Protein (DGCR8)/Pasha to release a ~70nt precursor miRNA (pre-miRNA) product with a 1-4 nt 3' overhang (13). DGCR8 recognizes the junction between single-stranded RNA and double-stranded RNA, which allows for the cleavage of the pri-miRNA to begin (10). The specific RNA cleavage by Drosha predetermines the mature miRNA sequence and provides the substrates for further processing (14). Through an interaction of Drosha and DGCR8, the stem of the pri-miRNA is cleaved eleven nts away from the two single stranded segments (10). MicroRNA precursor-containing introns are derived from certain debranced introns that fold into hairpin structures with 5' monophosphates and a 3' 1-4 nt overhang, which mimic the structure of pre-miRNAs, allowing entry into the miRNA-processing pathway (10, 13).

The 3' overhang of the pre-miRNA (or miRNA precursor-containing introns) is then recognized by Exportin-5 (Exp-5) which transports it into the cytoplasm via a Ran-GTP-

dependent mechanism that releases the pre-miRNA from Exp-5 (14). In the cytoplasm, the pre-miRNA is cleaved by RNase enzyme III Dicer about two helical turns away from the ends of the pre-miRNA steploop producing a double stranded RNA. A helicase then unwinds the cleaved double stranded RNA in a strand specific direction (13). The unwound strand with relatively lower stability of base-pairing at the 5'end is subsequently incorporated into a RNA-induced silencing complex (RISC) and thus, becomes the active miRNA, while the other strand is typically degraded. RISC is composed of several proteins, including the transactivation-responsive RNA-binding protein (TRBP) and Argonaute (Ago) proteins (11). These proteins contain four domains: PAZ, which binds to the 3' end of the active miRNA, and the N-terminal, middle, and Piwi domains, which form a unique structure that creates grooves for target mRNA and miRNA interactions (10). Once incorporated into RISC, the mature miRNA guides the complex to target sequences through the binding of imperfect complementary sites within the 3' UTR of mRNA transcripts by the Watson-Crick base-pairing mechanism, with a 5'-end 2-8 nts that is exactly complementary to the recognition motif within the target (10, 12). This 5'-end 2-8 nt region known as the seed sequence is critical for miRNA actions (10, 15). Partial complementarity with the rest of the sequence of a miRNA also plays a role in producing post-transcriptional regulation of gene expression (11).

MicroRNA regulate gene expression by either translational repression or degradation of target mRNA, or both, depending on the degree of sequence complementarity (15, 16). The miRNA seed sequence is crucial for miRNA targeting and function, and is responsible for searching for complementarity to sequences in the 3'UTR of all target genes (13, 17). The overall degree of complementarity of the binding site, the number of recognition motifs corresponding to the seed sequence of the miRNA, and the accessibility of the binding sites all determine how

gene expression is regulated (10). The greater the degree of complementarity of accessible binding sites, the more likely a miRNA degrades the targeted mRNA (10, 17). MicroRNAs do not require perfect complementarity for functional interactions with mRNA targets, and thus a single miRNA can regulate multiple targets and conversely, multiple miRNAs are known to regulate individual mRNAs (13). A simple change of one miRNA can provoke a chain reaction of feedback pathways involving multiple miRNAs and affecting multiple target genes of the same or different pathways (10). Likewise, the deregulation of one single miRNA is enough to trigger global alterations of genetic programs implicated in cell proliferation, differentiation, survival, or invasiveness depicted in **Figure 1.3** (18).

In addition to up- or down-regulating the transcription of genes, the miRNAs themselves can be up-regulated or down-regulated. The combination of up and down regulation can help identify genes that are regulated by specific miRNAs and cellular processes that are affected by specific miRNAs (10, 19). Genome-wide microarray expression techniques are widely used to comprehensively assay the global miRNA expression profile (also termed miRNome) in samples (10).

MicroRNA and target genes

MicroCosm (http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/), a central online repository for miRNA nomenclature, sequence data, annotation and target prediction, is a web resource developed by the Enright Lab at the EMBL-EBI containing computationally predicted targets for miRNAs across various species (20). The miRNA sequences are obtained from the miRNA Registry, and most genomic sequences are obtained from EnsEMBL. MicroCosm uses a miRanda algorithm to identify potential binding sites for a given miRNA in a

genomic sequence by using a weighted scoring system and rewarding the 5' end of the miRNA. The entire process of assembling miRNAs, genomic sequences, cross species UTR alignments and miRanda analysis is performed in parallel on a high performance compute cluster. Target and sequence information can be downloaded by the user based on genome and format (21).

Complementing MicroCosm, CircuitsDB (http://biocluster.di.unito.it/circuits/index.php) is a public web application devoted to the study of interactions between transcriptional and post-transcriptional regulatory network integration in the human and mouse genomes based on bioinformatic sequence-analysis (22). The database was constructed using an *ab-initio* oligo analysis procedure for the identification of the transcriptional and post-transcriptional interactions. Currently, the focus pertains to the study of mixed miRNA/ TF Feed Forward Regulatory Loops (FFLs), which are regulatory circuits in which a master TF regulates a miRNA and, together with it, a set of joint target protein-coding genes (**Figure 1.4)**. CircuitsDB allows users to explore and directly investigate relationships in terms of their sequence and functional annotation through a bioinformatic sequence analysis pipeline applied to the human and mouse genomes. The main access point in CircuitsDB is a dataset of mixed FFLs where a researcher's entry point of interest can be a TF, miRNA, or gene. CircuitsDB also contains a transcriptional network and post-transcriptional network that users can explore where entry points can be a TF of interest, a gene, or a DNA oligo. Users can also access all data sets contained in CircuitsDB in plain text format and download the complete catalogue of mixed FFLs for the human and mouse genomes (22).

miRNAs exhibit important regulatory roles in a variety of biological processes, including development, cell proliferation, cell survival, and apoptosis (8, 23). Notably, most of these processes, when miss-regulated, can lead to cancer.

Review of complex diseases: glioblastoma multiforme, ovarian serous cystadenocarcinoma, and illicit drug abuse

Introduction to cancer

Cancer, also known as malignant neoplasm, is a class of diseases in which grouped cells display uncontrolled growth, invasion, and damage of nearby tissues or organs (24). The mass of cancer cells will eventually become large enough to produce lumps. These lumps are most commonly referred to as tumors and defined as a mass with a distinct growth pattern from adjacent normal tissues in the absence of stimuli (24, 25). Tumors can be classified as benign or malignant. Benign refers to a tumor that is localized and has not metastasized to another organ or invaded nearby tissue. Metastasis is where cancer cells break away from a tumor and enter the bloodstream or lymphatic system to form secondary tumors in other parts of the body (24). In general, benign tumors are usually not harmful, as they grow slowly and can usually be removed without return. However, depending on the size and weight, some benign tumors can press on nearby organs, blood vessels, or nerves causing problems. Malignant tumor cells are abnormal and invade and destroy surrounding tissues, possibly even metastasizing to various other tissues and organs in the body (26). On the molecular level, human tumors manifest a complex interplay of multiple, nonrandom genetic events that encompass activation of proto-oncogenes and inactivation of tumor suppressor genes, which in turn lead to aberrant expression of growth factor receptors and their ligands thus, promoting the initiation and progression of cancer (26, 27). The fundamental mechanisms underlying the genetic basis of cancer are constantly being defined and involve alterations in three general categories of genes: 1) proto-oncogenes, which are involved in growth promotion (defects leading to cancer are gain of function); 2) tumor

suppressor genes, which are negative regulators of growth (a loss of function gives rise to cancer); and 3) DNA repair genes (28).

Proto-oncogenes and oncogenes are a class of genes that encode for proteins that function to positively promote cell proliferation. The normal (non-mutant) versions are commonly referred to as proto-oncogenes while the mutant versions or inappropriately active forms are known as oncogenes (26). Mutations converting proto-oncogenes into their oncogenic forms are usually gain-of-function mutations, and these mutations include point mutations; structural alterations, such as insertions, deletions, inversions, and translocations; gene amplification; and hypomethylation of transcription regulatory elements (24, 25). Genes that repair DNA or caretaker genes help maintain fidelity of the genome and, when functioning abnormally, can result in a mutator phenotype (i.e., an enhanced frequency of unrepaired mutations) and in turn, a predisposition to cancer (27).

Different types of cancers have unique attributes, different growth and proliferation rates, and various responses to therapies. Cancers are defined as primary when they are developed in the same site without evidence of a previous neoplasm, and secondary if originated from antecedent malignant cells (24).

Brain cancers are classified according to The World Health Organization (WHO) by grade, histology, and group. The WHO grading system has four classes according to the level of malignancy of the brain neoplasm: I, II, III, and IV (29). Grade I tumors are characterized by a low proliferation rate and are usually cured with surgical resection. Grade II tumors have an increased capacity for recurrence compared to grade I tumors, allowing for more aggressive types of cancer to develop. Grade III tumors, compared to grade II tumors, are usually defined by a higher progression of the neoplasm. The most aggressive and fatal of tumors that reproduce

rapidly and invade other tissues is grade IV (29, 30). This grading system is applied to all types of primary malignant neoplasms of the central nervous system. Among the categories of brain neoplasms, astrocytomas constitute one of the largest groups and are derived from glial cells (30). The most aggressive form is the WHO grade IV astrocytoma known as glioblastoma multiforme (GBM).

Ovarian tumors are classified according to cell type and are stratified as benign, borderline, or malignant based on the degree of cellular proliferation, nuclear atypia, and presence or absence of stromal invasion (28). The WHO's classification system subdivides ovarian tumors into five main categories: epithelial tumors, sex cord-stromal tumors, germ cell tumors, metastatic tumors, and other (31). These categories are further grouped into histological types such as serous, mucinous, endometrioid, clear cell, transitional cell tumors, mixed epithelial tumors, and others (31). Ovarian tumors are classified into two grades: type 1 tumors, that are low grade and slowly developing, and type 2 tumors, which are high grade and rapidly progressive (32).

Impact of brain cancer on the population

Glioblastoma multiforme is the most common and aggressive of primary brain tumors, with an incidence of 3.55 new cases per 100,000 Caucasians per year (8). Glioblastoma multiforme consists of a genotypically and phenotypically divergent population of cells that are highly malignant and infiltrate the brain extensively (33). The survival median is less than one year with essentially no long term survival (33). Glioblastoma multiforme is given a WHO grade of IV due to the capacity to quickly proliferate, invade, and progress in the brain.

Glioblastomas can be classified as primary, manifesting de novo, or secondary, progressing to GBM from lower grade gliomas. Primary GBM accounts for 60% of the cases and typically develops in patients over the age of forty-five (33, 34). Secondary GBM develops from a malignant transformation of a previously diagnosed low-grade tumor (cancers that have clinical, radiologic, or histopathologic evidence of malignant progression from a preexisting lower-grade tumor), is usually more common in younger patients (33). The histopathologic findings of primary and secondary GBMs are indistinguishable, and the prognosis does not appear to be different after adjustment for age. Glioblastoma multiforme is slightly more common in men than women, with a male-to-female ratio of 3:2 (35). These high grade astrocytomas are also more common in Caucasians compared to African Americans, Latinos, and Asians (35).

Despite major improvements in neuroimaging, neurosurgery, radiation and chemotherapy techniques, the overall prognosis of GBM has changed little in the past two decades. The topographically diffuse nature of the disease makes it difficult to completely extract in surgery. Glioblastoma multiforme's high level of cellular heterogeneity is associated with therapeutic resistance (8). The understanding of the genetic basis of this malignancy is important for the development of effective therapies.

Glioblastom multiforme is multiforme, as its name suggests, in various ways: grossly, showing regions of necrosis and hemorrhage; microscopically, with regions of pseudopalisading necrosis, pleomorphic nuclei and cells, and microvascular proliferation (33). Likewise, there are multiple genomic events associated with GBM such as deletions, amplifications, and mutations of individual genes. For example, the cyclin-dependent kinase inhibitor 2A/ cyclin-dependent kinase inhibitor 2B (*Cdkn2a/Cdkn2b*) locus is deleted in 46.4% of all GBM cases, and epidermal

growth factor receptor (*Egfr*) is amplified in 35.7% of all GBM cases (36). Tumor protein 53 (*TP53*) is the most commonly mutated gene in human cancer, with about 40% of GBM patients displaying such mutations (37).

Impact of ovarian cancer on the population

Ovarian cancer, a cancer that begins in the ovary of a women's reproductive system, is a potentially fatal threat to women's lives. Ovarian cancer is the leading cause of death from a gynecologic malignancy among women in the United States and is the fifth leading cause of cancer deaths among women over all (28). In 2013 approximately 22,240 women in the United States will receive a new diagnosis of ovarian cancer and about 14,230 will die from ovarian cancer (38). Generally, less than half (45%) of ovarian cancer patients survive more than five years after initial diagnosis (39).

One of the biggest detriments to effective treatment for ovarian cancer is the failure to reliably identify early stage disease (28). Due to the asymptomatic nature of the disease in initial stages and nonspecific symptoms, early detection is difficult (28). As a result of a lack of early specific symptoms and signs, newly diagnosed patients may have already developed advanced cancer where the cancer has disseminated beyond the ovary, at which point the five year survival rate is less than 20% (40). Clearly, an effective form of screening and early detection ovarian cancer would significantly impact outcome.

Epidemiologic studies have shown that endocrine, environmental, and genetic factors are important in the carcinogenesis of ovarian cancer (28). Age is the most significant risk factor of ovarian cancer, with incidence rates increasing with each decade of life and peaking in the middle to late seventies (28, 41). Ovarian steroids, estrogen and progesterone, are important factors for tumor growth (42). Biochemical and molecular studies have shown that leiomyomas

(pelvic tumors) have significantly increased levels of both estrogen and progesterone when compared to normal tissue (42). Approximately 10% of ovarian cancers arise in the setting of known genetic predisposition, with the majority of cases associated with mutations in the breast cancer type 1 and 2 (*BRCA1* and *BRCA2*) genes (27-28, 41-43).

Although cancers may arise from most of the many cell types of the ovary, the vast majority are believed to originate from the cells covering the ovarian surface. Accounting for 90% of ovarian cancers, epithelial ovarian cancer is a heterogeneous group of neoplasms and is divided into histologic subgroups, each with their own underlying molecular genetic events (28, 41). Among them, the serous type accounts for 75-80% of epithelial ovarian carcinomas (41). Epithelial tumors arise from the surface epithelium or from the crypts or inclusion cysts developed from this surface epithelium. Common epithelial ovarian tumors are classified according to their cell type, grade, and stage (28). Although ovarian cancer is the most lethal of gynecologic malignancies, relatively little is known about the molecular genetics of ovarian cancer's initiation and progression (41).

Association between transcriptome profile and brain and ovarian cancer

Molecular genetics of glioblastoma multiforme

A deeper understanding of the molecular mechanisms behind GBM must be gained to identify potential targets for therapeutic intervention and to develop more optimized and effective treatment strategies. Glioblastoma multiforme is a complex and genetically unstable tumor characterized by a multitude of chromosomal gains and losses, gene mutations and amplifications, epigenetic dysregulation, and aberrant post-translational modifications (44). In tumors, multiple modes of gene perturbation exist including sequence mutations, copy number alterations, gene fusion events, or epigenetic changes. (45). Molecular studies thus far have

identified at least three important genetic events in human GBMs: 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 (RB) tumor suppressor pathways (33, 45, 46).

RTK Genes

The receptor tyrosine kinases are a transmembrane protein family that plays crucial role in tumor growth, survival, metastasis, dissemination and angiogenesis (47). The most common *RTK* target of mutation in GBM is amplification of *Egfr*, often coupled with intragenic deletion resulting in a constitutively activated form (34). *Egfr* is involved in control of cell proliferation and is frequently activated in primary glioblastoma multiforme (48). *Egfr*-mediated signaling plays a critical role in normal brain development (44). Potential ways in which *Egfr* might directly influence GBM include gene amplification and mutation, as well as wild type and mutant receptor over-expression (44).

PI3K Pathway

The PI3K-AKT pathway is a frequent target of disruption in GBM. Major downstream effects of PI3K/AKT activation include cell growth, proliferation, survival, and motility, which are all factors that drive tumor progression (44, 45). Known alterations in the PI3K/AKT pathway include frequent alterations or homozygous deletion of phosphatase and tensin homolog (*PTEN)*, mutations in *AKT*, and alterations to the components of the PI3K complex (45).

PTEN has been shown to be a critical tumor suppressor gene that is commonly inactivated in GBM by deletion, mutation, or attenuated expression to abolish the major negative regulator restraining PI3K activation (46, 49). *PTEN* is a cellular phosphatase that turns off signaling pathways. When phosphatase activity is lost due to mutation, signaling pathways can

become activated constitutively, resulting in aberrant proliferation. The AKT pathway is regulated by *PTEN* (46). Loss of functional *PTEN* leads to increased activity of the AKT and mTOR kinase pathways, which promotes tumor cell survival and proliferation through phosphorylation and activation of several downstream mediators (46). Phosphoinositide-3-kinase, regulatory subunit 1 (alpha) *(PIK3R1*) encodes a regulatory subunit of PI3K and is important in down-regulating AKT-mediated cell signaling (44).

p53 and RB

The p53 tumor suppressor pathway prevents the propagation of unstable genomes by activating the expression of downstream genes that inhibit growth and invasion, and thus functions as a tumor suppressor. This pathway is frequently altered (inactivated) in GBM (45, 50). Nearly all GBM tumors contain alterations within the p53 pathway (~86%), including mutations and deletions of *Tp53* and amplifications of MDM2 oncogene, E3 ubiquitin protein ligase (*MDM2*) and Mdm4 p53 binding protein homolog (mouse) (*MDM4*) (45, 48).

The *Tp53* gene encodes tumor protein p53, which responds to diverse cellular stresses to regulate target genes that induce cell cycle arrest, apoptosis, and DNA repair (50). Malfunction of *Tp53* inhibits apoptosis of malignant cells. Mutations in the *Tp53* gene once thought to predominate in secondary GBM have now been found with surprisingly high frequency in primary GBM (44). *MDM2/MDM4* amplification or over-expression constitutes an alternative mechanism to escape from p53-regulated control on cell growth by binding to the p53 gene and diminishing its tumor suppressor function (49, 50).

Nearly all GBMs universally circumvent cell cycle inhibition through genetic alterations to the RB pathway (45, 48). Alterations include mutations in the tumor suppressor gene retinoblastoma 1 (*RB1*), amplification of cyclin-dependent kinase 4 and 6 (*Cdk4* and *Cdk6*), and

homozygous deletions of cyclin-dependent kinase inhibitor 2A, 2B, and 2C (*Cdkn2a*, *Cdkn2b*, and *Cdkn2c*) (45, 49). *Cdk4* and *Cdk6* are responsible for the phosphorylation of RB gene product. Mutations in this gene were found to be associated with tumorigenesis of a variety of cancers. The *Cdkn2a* gene is frequently mutated or deleted in a wide variety of tumors and is known to be an important tumor suppressor gene (49). The *Cdkn2b* and *Cdkn2c* genes encode a cyclin-dependent kinase inhibitor, which forms a complex with *Cdk4* or *Cdk6*, and prevents the activation of the CDK kinases. Thus, the encoded proteins of *Cdkn2b* and *Cdkn2c* function as cell growth regulators that control cell cycle G1 progression (28, 49).

Currently, genetic alterations on DNA mismatch repair (MMR) genes and the O-6-methylguanine-DNA methyltransferase (*MGMT*) gene are of interest in regards to therapeutic research in GBM. The DNA MMR system is a highly conserved biological pathway that corrects errors generated during DNA replication (51). The primary function of the MMR system is to eliminate base-base mismatches and insertion-deletion loops which arise during replication of DNA (51). MMR defects are produced by mutations in any of the genes, such as mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (*MLH1*), mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli) (*MSH2*), and mutS homolog 6 (E. coli) (*MSH6*), and inactivation of this system leads to an altered gene and protein expression (51). *MGMT* is a DNA repair enzyme that removes alkyl groups from guanine residues (34, 48). One of the most important biomarkers for GBMs is the methylation status of *MGMT*, which predicts sensitivity to temozolomide (48). Temozolomide (TMZ) is the most commonly used and most effective chemotherapy for GBM (44). The benefit of TMZ is superior in patients with low levels of *MGMT* in their tumor (44). *MGMT* removes the methyl groups from the O-6 position of guanine, thus reducing TMZ's effectiveness (44). Transcriptional silencing of the *MGMT* gene

by promoter hypermethylation is seen in ~50% of GBMs and is linked to prolonged survival (34). MMR deficiency and *MGMT* methylation together, in regards to treatment, exert a powerful influence on the overall frequency and pattern of somatic point mutations in GBM tumors (48). A better understanding of the mechanisms driving tumor initiation and progression and subsequent discovery and advancement of targeted therapies for cancer offers new hope for molecularly targeted treatment of GBM.

Molecular genetics of ovarian cancer

Ovarian cancer is the most lethal of gynecologic malignancies, yet relatively little is known about the molecular genetics of its initiation and progression. Ovarian cancer arises from genetic alterations resulting in different patterns of expression, such as amplification and deletion (28). Alterations in tumor suppressor genes such as *Tp53* and *RB*, and other oncogenes such as v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog (*KRAS*) and v-myc myelocytomatosis viral oncogene homolog (*c-Myc)* have been shown to play an important role in ovarian cancer development (43).

Around 54% of ovarian cancers exhibit chromosomal variants including amplifications of 8q, 1q, 20q, and 19q, and ~50% have deletion of 13q, 4q, and 18q (43). The deletion of 13q results in the inactivation of the *RB* gene, the negative regulator of cell growth and suppressor of tumorigenesis (28, 43). This effect may be potentiated in ovarian cancer by the amplification of *c-Myc*, an antagonize tumor suppressor function of the *RB* gene (28). Over-expression of *c-Myc* is associated with stage III disease, suggesting a role in disease progression (28).

The genes *RAS*, v-akt murine thymoma viral oncogene homolog 2 (*AKT2*), and *PIK2CA* code for protein products that act within signal transduction pathways and are oncogenes

involved in the pathogenesis of ovarian cancer (28). The AKT2-PTEN-PI3K pathway is important in the processes of cell cycle, apoptosis, and RAS signaling, and appears to be frequently disrupted in ovarian tumors by alterations in *PTEN*, *PIK2CA*, and *AKT2* (28). *PTEN* is down-regulated in a proportion of ovarian tumors, and an inverse correlation between *PTEN* expression and activated *AKT2* expression has been found (28). Mutations in the *KRAS* gene have been reported at a fairly high frequency in ovarian tumors (28). Individuals that are heterozygotes for the *BRCA* gene have a 20-40% lifetime risk of ovarian cancer (52). The loss of wild-type BRCA allele and mutational inactivation of the *Tp53* gene are used to search for preclinical genetic evidence of ovarian tumorigenesis (52).

Egfr is over-expressed in 70% of ovarian tumors and is associated with advanced disease, poor prognosis, and chemo-resistance (53). Platelet-derived growth factor and its receptor (*Pdgf-Pdgfr*) and v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog (*c-KIT*) are over-expressed in 70% of ovarian cancers (53). The expression of *Pdgfr* is associated with shorter overall survival, and abnormal *c-KIT* expression, yet not mutations, have been associated with ovarian tumors (28). A possible new target oncogene v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog avian *Src*), which has been linked to drug resistance and survival of ovarian cancer cell lines in translation studies, has been identified as being over-expressed in ovarian cancer (54).

The identification of candidate oncogenes and tumor suppressor genes has provided a foundation for further studies to assess the relative importance of these genes and pathways in the development of ovarian cancer.

Role of microRNA in cancer

Cancer is a genetically complex disease that is caused by the accumulation of mutations, which ultimately lead to the deregulation of gene expression and uncontrolled cell proliferation (24). A possible link between miRNAs and cancer was reported in chronic lymphocytic leukemia, where miR-15 and miR-16 were found to be down-regulated in the vast majority of tumors (13). Recent studies have shown that miRNA are aberrantly expressed in a variety of cancers (9). Over 50% of annotated human miRNA genes are located in fragile chromosomal regions which are susceptible to amplification, deletion, or translocation in the process of cancer development and progression (14). Possible mechanisms leading to abnormal expression of miRNA in cancer that have been reported include chromosomal rearrangements, genomic copy number changes, defects in miRNA biogenesis pathway, regulation by transcriptional factors, and epigenetic modifications (13).

Epigenetic regulation of miRNA expression has been reported in various cancers. Between 20% and 40% of miRNAs are located close to CpG islands making them susceptible to epigenetic silencing (19). Epigenetic gene silencing is associated with aberrant hypermethylation performed by one of several DNA methyltransferases of CpG dinucleotides within the gene promoters. Inhibition of these DNA methyltransferases leads to the reversal of aberrant hypermethylation (55). Global hypomethylation or aberrant hypermethylation of gene promoter CpG islands result in tumor cell genomic instability and gene silencing, particularly of tumor suppressor genes (56).

Evidence shows that miRNAs have the ability to function as oncogenes or tumor suppressors, and expression profiling has revealed characteristic miRNA signatures in a variety of human cancers (14). To date, significant miRNA expression changes have been observed in

every type of tumor analyzed by profiling experiments (39). MicroRNA profiles are able to distinguish between normal and cancer tissues, separate different cancer types, and stratify the cancer differentiation state with high accuracy (15). In fact, miRNA profiling outperformed cDNA microarrays in the classification of tumors of unknown primary (11).

Transcription factors, microRNA and target genes and pathways of cancer

Transcription factors and miRNAs are important components of gene regulatory networks that control the expression of genomic information. By binding to discrete *cis*-regulatory elements, individual TFs and miRNAs can control dozens, if not hundreds, of target genes, and together they generate a complex combinatorial code (57). Transcription factors and genes containing binding sites for TFs have a high probability of being targeted by miRNAs. MicroRNA and TFs are linked to one another in gene regulatory networks.

Transcription factors act as both oncogenes and tumor suppressors, thus, making them targets for the development of anticancer drugs (16). Cancer cells avoid apoptosis by the activation of oncogenes and the loss of tumor suppressor genes. By controlling the up or down regulation of expression of specific genes, TFs and miRNA play a role in pathways affected by cancer. MicroRNA may function as a oncogene or as a tumor suppressor depending upon the cell type specific microenvironment, which may provide a different repertoire of available target genes (16). A miRNA can act as a tumor suppressor when its function loss can initiate or contribute to the malignant transformation of a normal cell. Loss of function can be attributed to several mechanisms, including genomic deletion, mutation, epigenetic silencing, and/or miRNA processing alterations (11, 14). For example, miR-16-1 negatively regulates cell growth and cell cycle progression and induces apoptosis in several human cancer cell lines (11, 39). A miRNA

can act as a oncogene by promoting proliferation, inhibiting apoptosis, or inducing tumor angiogenesis (11). There is strong evidence that miR-21 functions as an oncogene as this miRNA is up-regulated in a wide variety of hematological malignancies and solid tumors (39). There is evidence that miRNA may play a dual role as both a tumor suppressor and oncogene, (which has been previously described in protein-coding genes involved in the pathogenesis of cancer such as *Tp53*), depending on the tissue and its transcriptome, including miRNA targets expressed in particular tissue (11). A particular type of cancer may be associated with the dysregulation of several distinct miRNAs and conversely, dysregulation of one miRNA can be associated with several cancer types (14). A single miRNA can have many targets that are involved in different oncogenic pathways, such as miR-181 that targets the genes B-cell CLL/lymphoma 2 *(Bcl-2,* apoptosis), T-cell leukemia/lymphoma 1A (*TCL-1,* AKT pathway), and CD69 molecule (adhesion). Basically, modulating the level of a single miRNA can affect many pathways at the same time (11).

MicroRNA profiling may be an invaluable tool that can be used to classify tumors that represent diagnostic challenges, and the discovery of distinctive miRNA signatures will likely improve the molecular classification of cancer (11, 14). The role of miRNAs as potential oncogenes and tumor suppressors has generated great interest in using them as targets for cancer therapies (15). Certain miRNA signatures are correlated with the prognosis of cancer and in the future could possibly be used to determine the specific course of treatment (15). In order to globally observe and identify miRNAs and their associated cancer modules, a generation of a cancer-(gene/transcription factor)-miRNA network is crucial.

Illicit drugs and addiction

Introduction into drug abuse

Drug addiction is defined as the loss of control over drug use or the compulsive seeking and taking of drugs despite adverse consequences (58). Drug addiction is characterized by three phases: preoccupation/anticipation, where one has a compulsive need to seek and take a drug; binge/intoxication, where one consumes the drug and has a loss of control over the amount of drug consumed; and withdrawal/negative affect, where one has periods of attempted abstinence closely followed by relapse to drug taking behavior (59, 60). Drugs of abuse produce both acute and chronic changes in brain function. Acute drug use is defined as consumption of a high dose of a drug on one occasion, normally corresponding to recreational or casual drug users (59, 61-63). Chronic drug use is defined as repeated use of large enough doses of a drug to maintain an excessive drug concentration in the body over a long period of time, more commonly associated with addicts (59, 61-63). Acute exposure to drugs can cause alterations in the gene transcriptions which are related to addictive properties of various drugs (63). Chronic drug use by humans resulting in addiction can be observed for most drugs of abuse. Drug addiction is a brain disease (63). Periods of prolonged drug use result in cellular and neurological adaptations in brain reward systems resulting in addiction (59).

Impact of drug abuse on the population and transcriptome

Substance abuse and addiction have a devastating effect on individuals and their families, as well as pose a worldwide threat to public health. Between 99,000 and 253,000 deaths globally were attributed to illicit drug use in 2010 (64). Global prevalence of HIV, hepatitis C, and hepatitis B is heightened among drug users (65). At least 15.3 million people have drug use

disorders (65). Illicit drug use also puts a heavy financial burden on society, with relation to drug related crime and drug treatment (65). Accidental and intentional injury, drug induced psychotic symptoms, and increased risk for heart, liver, and lung disease can all be consequences of substance abuse. Addiction is a psychiatric disease attributable to biological and environmental factors.

Genetic factors contribute around 40-60% of vulnerability to drug addiction, with environmental factors accounting for the rest (58). Currently, there is a limited understanding of the underlying mechanisms of drug addiction. To study and understand the neurobiology of compulsive drug intake, it is necessary to employ an animal model that accurately recapitulates aspects of the disorder seen in human addicts**.** The mice striatum is a well-established system to study gene expression changes associated with illicit drug challenges (63).

Changes in gene expression have been documented in the brain after administration of drugs of abuse (63, 66, 67). Brain regions that control rewards sensitivity, motor function, and habit learning are the major neural target sites of addictive drugs (63). For example, the use of cocaine, a potent pharmacological stimulus that exerts widespread effects in striatal neurons, results in the blocking of the dopamine transporter in the striatum increasing dopamine levels and contributing to the reward-related effects (59). Further research must be done in the identification and evaluation of coordinated gene networks and transcriptional signatures associated with drug addiction. New opportunities arising from the analysis of these networks include identifying novel relationships between genes and signaling pathways, connecting biological processes with the regulation of gene transcription, and associating genes and gene expression with diseases. A broader understanding of genetic changes in the brain from short and long term drug abuse can be obtained with gene expression data.

Pathways statistically significantly enriched in addiction-related genes can be clustered into two categories: 1) upstream events of drug addiction including crosstalk among *MAPK* signaling, insulin signaling, and calcium signaling; and 2) downstream effects including regulation of glycolysis metabolism, regulation of the actin cytoskeleton, and apoptosis (58). The *MAPK* signaling pathway plays a role in regulating synaptic plasticity related to long lasting changes in both memory function and addictive properties (58). The *GnRH* signaling pathway is involved in regulating stress and may be involved in the regulation and control of certain emotional behaviors in addiction (58). As technologies continue to improve, more and more genes and pathways may emerge as being linked to addiction.

Transcriptomic analysis of complex diseases

Platforms to measure gene expression

DNA microarrays are a commonly used to detect the presence and abundance of labeled nucleic acids in samples on a high throughput level. Microarrays allow the simultaneous measurement of the expression of thousands of genes across treatments, developmental stages, or other conditions. The probes in the microarray can be oligonucleotides (or cDNAs) that are a perfect complementary match to a segment of the gene of interest. Gene expression describes the transcription of information contained within DNA into mRNA that is then translated into the proteins critical for the functions of cells (50). By examining the amounts of mRNA that are produced by a cell, scientists are able to identify which genes are expressed, which in turn reveals how cells adapt to changes within and outside of the organism due to treatments or other conditions. Changes in gene expression levels can be a suitable indicator of the changes in the abundance of protein (68). Data from multiple microarray experiments are available in public databases like the Gene Expression Omnibus (GEO; http:// www.ncbi.nlm.nih.gov/geo), and the

most commonly used microarray platform is the Affymetrix *in-situ* synthesized oligonucleotide array (69).

Affymetrix *In-Situ* Synthesized Oligonucleotide Array Technology

Oligonucleotide probes used in microarray platforms are short (i.e. 20 to 80 nucleotides long) segments of RNA or DNA synthesized by cleaving longer segments, or by polymerizing individual nucleotide precursors. During *in-situ* synthesis, oligos are built up base-by-base on the surface of the array (68). This occurs by a covalent reaction between the 5' hydroxyl group of the sugar of the last nucleotide to be attached and the phosphate group of the next nucleotide. A protective group on the 5' position of each added nucleotide prevents the addition of more than one base during each round of synthesis. Before moving on to the next round of synthesis, the protective group is converted to a hydroxyl group (70). Different methods can be used for this deprotection, with our interest focused on the Affymetrix technology.

Affymetrix GeneChip microarrays are high density oligonucleotide gene expression arrays widely used in genomics studies (70). For deprotection, a photodeprotection method using masks is employed. Basically, masks allow light to pass to appropriate features, while keeping it from other features, in order to convert the protective group on the terminal nucleotide into a hydroxyl group where bases can then be added. Each step of synthesis requires a different mask. Photodeprotection has a coupling efficiency of about 95%, meaning about 95% of nucleotides are successfully added at each step of synthesis (69, 71). The longer the oligonucleotide synthesized, the worse the yield of full length oligos. The composition of the final population of oligonucleotides is influenced by capping, which prevents further synthesis on a failed oligonucleotide resulting in all oligonucleotides on a feature to have the same start,

but be of different lengths (72). On an Affymetrix microarray, each gene is represented by a probe set consisting of 11 different pairs of 25 base pair oligos covering features of the transcribed gene. A probe set can consist of a series of probe pairs and represents an expressed transcript (72). Each pair consists of a perfect match (PM) and a mismatch (MM) oligonucleotide (although the latest versions of Affymetrix are excluding MM). The PM probe exactly matches the sequence of a particular standard genotype, while the MM differs in a single substitution in the central, 13th base and is designed to distinguish noise caused by non-specific hybridization from the specific hybridization signal (71). A single sample that has been previously labeled with a fluorescent dye (biotin) is hybridized to the microarray probes, the microarray is scanned, and an image of fluorescence intensities is obtained. Gene expression is determined by comparing the signal intensity from hybridization to probes complementary to the gene being measured with the signal intensity from hybridization to probes that contain mismatches; the signal from the mismatch probes are thought to represent cross-hybridization (73).

Affymetrix Microarray Experiment Protocol

There are four basic steps in using a microarray to measure gene expression in a sample: sample preparation and labeling; hybridization; washing; and image acquisition (72). Affymetrix platform experiments a specific protocol, which allows for easier comparison of results between laboratories than other platforms. First, RNA must be extracted from the tissue of interest. For labeling, one must construct a biotin-labeled complementary RNA for hybridizing to the GeneChip(69). The GeneChip anchors tens of thousands of closely arrayed oligonucleotide probes on to the surface of solid substrates, where the process of recognition can be highly parallel (74). In hybridization, DNA probes form heteroduplexes with labeled DNA (or RNA)

via Watson-Crick base pairing (71). Slides are washed after hybridization to ensure only the target specifically bound to features on the array (DNA/RNA we are trying to measure) are left and to reduce cross-hybridization. Affymetrix has integrated its image-processing algorithms into the GeneChip experimental process. Overall, oligonucleotide arrays are powerful tools for monitoring gene expression.

Microarray data processing and normalization

Quality control of the gene expression data involves the transformation of the image scanned from the arrays into reliable mRNA expression data. Hence, processes regarding the measurement of probe fluorescence in the array and data transformation are crucial steps that must be performed prior to statistical analyses in microarray experiments. This review of processing steps centers on the in-situ Affymetrix GeneChip platform (69).

During hybridization, fluorescence signals are emitted from the GeneChips and images are scanned at the end of the process and stored as .DAT files. This raw data file contains details regarding the image size, technical information and the pixel intensities from the whole array. The estimation of the intensity of each spot is stored in .CEL files after the grid alignment, to localize each probe cell in the array, and the computation of the 75th percentile of the intensity from the pixels in the corresponding cell or spot. In this manner, the correct grid alignment is important to avoid errors when summarizing the information from .DAT to .CEL files. In addition to this factor, other events can affect the overall quality of image acquisition, such as background variation and flagging. Pixels in high-intensity parts tend to lose signals to surrounding areas. If these neighboring pixels have too low of a signal, they will record the intensity from different pixels resulting in blurry images (71). The background signal must be subtracted from the feature in order to have a better estimate of the cell hybridization. Flagged

features have image problems and must be removed from the analysis. Some cells can have a higher pixel standard deviation compared to the mean (bad feature), higher background than foreground signal (negative feature) or even have a very low signal (dark feature). These estimates can be automatically removed from the data by software or manually by the user. Like in the other steps of gene expression analysis, image acquisition needs standardization of the procedures to yield high quality results (72).

Prior to the statistical analyses, the probe intensity data must be normalized in order to remove systematic errors and bias originated from microarray experimentation. Normalization attempts to remove sources of variation allowing the realization of meaningful biological comparisons. A widely used technique to normalize microarray data is the GC-Robust Multichip Average or GC-RMA (71, 75). . This method takes probes sequences into consideration and uses quantile normalization and median polish to process the expression data (76). The GC-RMA method usually yields more precise results and is more accurate than other normalization methods.

Microarray technology is susceptible to technical error even after following the recommended experimental protocols. Some of the systematic sources of variation can be removed through the process of normalization or preprocessing of the microarray gene expression intensities before analysis. Examples of these sources of technical error include arrays with higher overall expression levels than others with comparable samples due to scanning or labeling, and probes appearing more fluorescent than others due to the nucleotide composition. The most common normalization techniques applied to Affymetrix microarray experiments is the quantile normalization followed by the GC-RMA. The quantile normalization results in all the arrays in an experiment having the same distribution, thus the same mean, variance, quantiles,

and percentiles. This normalization maintains the within gene changes in expression across treatments or samples but removes all systematic differences between arrays across genes. The more general RMA normalization provides a single gene expression value per gene and microarray by combining all the probe values within a gene (75). The GC-RMA adjusts the expression intensity by the composition of bases and location of the bases on the probe. The rationale for this adjustment is that the hybridization bond between the complementary probe and sample bases depends on the labeling and base (72, 75). The label typically binds to Cytosine C base, and this can interfere with the Cytosine-Guanine bond. However, the C- G bond is stronger than the Adenine-Thymine bond. An adjustment formula was developed to account for these factors in the normalization of the gene expression measurements. The normalizations of the microarray experiments in my studies are implemented in Beehive (http://stagbeetle.animal.uiuc.edu/Beehive). Beehive is a public integrated suite of web-based tools to study gene expression data from microarray experiments (77).

Data processing and normalization are critical steps to remove systematic errors and bias from the data and allow a more adequate subsequent modeling and testing of gene expression across conditions. In many microarray experiments, the ultimate goal is to identify genes that are differentially expressed across conditions. Here, gene expression is considered the response variable, and the explanatory variables are the condition(s) (factors or covariates of interest). These models are analyzed by using a one or multi-way ANOVA model or regression model. The results of these models provide the ranking of the genes in terms of P-value and fold changes, which allows for the identification of differentially expressed genes.

Multiple testing corrections adjust p-values derived from multiple statistical tests to correct for occurrence of false positives. For example, in the stringent Bonferroni correction, the

p-value of each gene is multiplied by the number of genes in the gene list (78). The false discovery rate (FDR) is the expected fraction of statistically significant results that are mistakenly declared significant. In the false discovery rate (FDR) correction, *N* p-values are sorted in ascending order (i.e., $p_1,p_2,\ldots,p_n$). Then *k* denotes the largest index *I* for which $p_i \leq di/N$ for all $i=1,\ldots,k \leq N$. All tests with p-values $p_1,p_2,\ldots,p_k$ are then declared significant (78). If the corrected p-value is still significant, the gene is still significant. In some microarray experiments, an additional objective is to identify biomarker genes that can be used to accurately classify samples across conditions or predict phenotypes (e.g. survival, growth). Here, the condition is actually the response variable, while the expression of one or multiple genes is the explanatory variable.

Linear mixed model analysis

Standard linear model

The general linear model is used to determine the strength of association between a response variable *y* and one or many continuous or discrete explanatory variables *x*, as well as predict for values of *x*. Explanatory variables can be discrete, taking one of several distinct values, or continuous, taking any value within a continuous range. A basic linear effects model can be written as (79):

$$\Upsilon_i = \beta_0 + \beta_1 \mathrm{X}_i + \varepsilon_i$$

where Y_i is the value of the response variable in the *i*th trial; β_0 and β_1 are parameters; X_i is a known constant, namely the value of the predictor variable in the *i*th trial; ε_i is a random error term with mean $E\{\varepsilon_i\} = 0$ and variance $\sigma^2\{\varepsilon_i\} = \sigma^2$; ε_i and ε_i are uncorrelated so that their covariance is zero (i.e., $\sigma\{\varepsilon_i, \varepsilon_j\} = 0$ for all $i, j; i \neq j$; and $i=1,\ldots,n$ (79). The parameters β_0 and β_1 are the regression coefficient. β_1 is the slope of the regression line indicating the change in the

mean of the probability distribution of *Y* per unit increase in *X*. The parameterβ_0 is the *Y* intercept of the regression line and gives the mean of the probability distribution of *Y* at *X*=0 (79).

Survival analysis

Survival analysis, also known as time-to-event analysis, is a statistical methodology to model the probability of occurrence of an event at a specific time or hazard (80). In survival analysis, the outcome variable is the time until an event occurs, which can be in years, months, weeks, or even days. The event is any designated experience of interest that may happen to an individual, most commonly death or disease incidence. Survival analysis allows for a study to start without all experimental units enrolled and to end before all experimental units have experienced an event through censoring. Censoring is used when one does not know the exact survival time (80). Subjects can be right, left, or interval censored. For example, if a person does not experience the event before the study ends right censoring is used; if an event occurred before the start of the study left censoring is used; or if both of the previous examples occur, interval censoring is used.

Survival data is generally described and modeled in terms of survival and hazard functions. Given that *T* denotes the survival time, where $T \geq 0$, and *t* is any specific value of interest for *T*, the probability that a person survives longer than some specified time *t*, i.e. $P(T>t)$, is defined as the survival function $S(t)$ (81):

$$S(t) = P(T > t)$$

Theoretically, the survivor function is graphed as a smooth curve with t ranging from zero to infinity **(Figure 1.5)**. For example, at the beginning of a study (time $t=0$) the probability of surviving past time zero is one since no one has yet gotten to the event. If the study period increased without limit (time t=infinity), eventually no one would survive resulting in the

survivor curve falling to zero. However, in practice obtained graphs are usually step functions **(Figure 1.6)**. This is due to the fact that the study period is never infinite in length and the estimated survival function may not go all the way down to zero at the end of the study (80).

Unlike the survival function which concentrates on not failing (event not occurring), the hazard function focuses on failing, or on the event occurring. The hazard function *h(t)* gives the instantaneous potential at time *t* for getting an event, like death or disease, given survival up to time *t* and the expression is (81, 82):

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

The hazard formula gives the probability that a person's survival time, *T*, will lie in the time interval between *t* and *t+Δt*, given that the survival time is greater than or equal to *t* (80). The division of the numerator (conditional probability) and the denominator (*Δt*=a small time interval) gives the probability per unit time. The hazard function is graphed with *t* ranging from zero to infinity and can start anywhere and go up or down in any direction over time.
For example, the hazard function 1) may be constant, where no matter what value of *t* is specified, *h(t)* equals the same value; 2) may increase over time, whereas survival time increases, potential for the event occurring increases; 3) may decrease over time, whereas survival time increases, potential for the event occurring decreases; or 4) may first increase and then decrease **(Figure 1.7)**.

In the context of identification of transcript profiles associated with survival in cancer patients (GBM or ovarian cancer), the explanatory variables in the survival model include the expression levels of miRNA or genes, cohort information such as therapy, gender and race and potential interactions between transcript and cohort information. The Cox proportional hazards model is a technique for investigating the relationship between survival time and variables. This

model can be used to identify expression profiles associated with an event or hazard, independent or dependent on clinical factors, through an interaction term. The model is:

$$h\{(t),(z_1 z_2 \ldots z_m)\} = h_0(t) * \exp(b_1 * z_1 + \cdots + b_m * z_m)$$

where *h(t,...)* denotes the event or hazard, given the values of the *m* biomarker expression, clinical factors, and interaction levels *(z_1, z_2,... z_m)* and the corresponding survival time *(t)*. The term *$h_0(t)$* denotes the hazard when all expression levels are equal to zero and all factors are at base level. Once the transcriptome profiles statistically and biologically associated with survival are identified, functional analysis can be used to pinpoint biological processes and pathways enriched among the transcripts. In addition to studying known pathways, gene networks can be inferred based on the co-expression of the transcripts associated with survival.

Feature selection

Two widely used feature selection approaches for model selection in regression include stepwise selection and forward selection. Stepwise selection is an automatic search procedure that develops a sequence of regression models, at each step adding or deleting an x variable until a final best subset of variables is developed and a best model reveled (79). Two different significance levels must be chosen, one for entry into the model and one for deletion from the model. Basically, the process is one of alternation between choosing the least significant variable to drop and then re-considering all other dropped variables for re-introduction into the model. The algorithm is described in terms of *t** statistics and associated p-values. In the first step, a simple linear regression model is fit for each of the *P*-1 potential *X* variables. The *t** statistic with the largest value is the candidate for the first addition

$$t_k^* = \frac{b_k}{s\{b_k\}}$$

The variable is added only if the t^* value exceeded a predetermined level or if the corresponding p-value is less than a predetermined α value (79). The regression routine now fits the regression model with two *X* variables, based again on the t^* value or p-value, and so on. After each new variable is added, the model is scanned to examine whether any other *X* variables in the model should be dropped. In summary, explanatory variables enter the model one at a time if they surpass a p-value threshold. After an explanatory variable is added, all pre-existing explanatory variables are tested to see if they are still significant at another p-value threshold. If any pre-existing explanatory variables are no longer significant, the variable is removed from the model. This process is repeated until no further explanatory variables can be added or removed from the model.

Forward selection is essentially a simplified version of the stepwise regression. At each step, each variable that is not already in the model is tested for inclusion with the most significant of the variables being added to the model. In forward selection, the step of testing whether a variable once entered into the model should be dropped is omitted (79). Hence, once a variable enters the model, it stays in the model. Variables are continually added until none of the remaining variables are significant enough to be added. A drawback of this selection is an addition of a new variable may render one or more of already included variables non-significant.

Meta-, functional and network analysis of transcriptome profiles

Meta-analysis

The integration of gene expression information from multiple microarray studies via meta-analysis can aid in the characterization of gene expression profiles. The integration of various studies can produce a more accurate identification of specific transcriptome biomarkers and pathways than a single study alone. Model-based meta-analysis is when linear models are

used to combine indicators of expression patterns from individual studies (such as fold changes or differences between mean groups, standardized estimates, or normalized values) and associated test-statistics or functions are then used to evaluate the expression pattern across these studies (83). Gene expression patterns can be accurately characterized with increased statistical power through the meta-analysis of microarray experiments (83). Examples of meta-analysis include comparison of lists of genes with differential expression, consideration of gene expression data across treatments and experiments, and combination of p-values or estimates (84). The strength of a meta-analysis is its ability to combine the results from various small studies that may have been underpowered to detect a statistically significant difference. However, challenges can sometimes occur with different conditions among studies.

A deeper understanding of the molecular relationship of tumors is necessary in order to derive insights into diagnosis, prognosis, and treatment. Due to the large number of factors influencing differentially expressed genes, it is not surprising that gene lists from independent studies generally show little overlap. In order to compile the most accurate and robust list of relevant genes, meta-analysis of multiple independent publically available data sets can be performed. Meta-analytic statistics provide a combined ranking of significant genes while not allowing for strong p-values from any individual study to dominate the results (85). The general framework of meta-analysis can be used to distill relevant information from a vast amount of published literature and available data sets (86). The aggregation of data from various experiments that is obtained with meta-analysis enhances both the precision and accuracy of pooled results (86). For example, the gene vascular endothelial growth factor A (*Vegfa*), which is known to be important in gliomas and generates more than 750 pertinent citations on its own, is sixth in the meta-analysis list but does not even fall among the top 30 genes on any individual

lists from each study (85). Meta-analysis results are likely to be more accurate and stable because of the greater sensitivity and specificity that result from the integration of data.

Functional analysis

High-throughput technologies such as microarrays, genome wide association studies, and next generation sequencing produce huge amounts of data of unachievable interpretation without the application of automatic procedures for functional profiling (87). Extensive bioinformatic analysis of microarray data can be used to characterize differences in transcript profiles and uncover predominant signaling molecules (such as cytokines) in one tissue relative to another which allows for the identification of potential targets among genes that are more highly expressed in order to discover novel inter-tissue signaling networks (88). Advanced computational methods are needed for the integration, mining, comparative analysis, and functional interpretation of high-throughput data. Bioinformatics methods using the biological knowledge accumulated in public databases has made it possible to systematically dissect large gene lists in an attempt to assemble a summary of the most enriched and pertinent biology. Sources of biological information to categorize the differentially expressed transcripts include functional (Gene Ontology (GO), http://www.geneontology.org/; Kyoto Encyclopedua of Genes and Genomes (KEGG), http://www.genome.jp/kegg/; Biocarta, http://www.biocarta.com/), regulatory (Transfac, http://www.gene-regulation.com/pub/databases.html; Jaspar, http://jaspar.genereg.net/), miRNAs, text-mining, and protein-protein interactions (89).

The consideration of themes or annotation categories shared by genes identified from analysis of microarray experiments helps facilitate interpretation by allowing one to discuss groups of genes instead of just individual genes. The goal of functional analysis is to identify

categories differentially represented (or enriched) among genes. Fisher's hypergeometric test tests for the enrichment (depletion) of a category among a list of significant differentially expressed genes by generating all possible contingency tables with marginal column and row totals (90). The hypergeometric distribution describes the number of successes in a sequence of *n* draws from a finite population without replacement. For *N* genes in a platform, *K* genes are in the significant list. The hypergeometric distribution describes the probability that in a subsample of *n* genes belonging to a category, *k* genes are in the significant list. For example, there are $\binom{N}{n}$ possible samples without replacement; $\binom{K}{k}$ ways to obtain *k* significant genes in a category; and $\binom{N-K}{n-k}$ ways to fill the rest of the sample where $\binom{K}{k} = K! \div (k!\,(K-k)!)$ (90). Thus the hypergeometric probability is (90):

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

For analysis of all genes in a platform (instead of a smaller list of significant genes), Gene Set Enrichment Analysis can be performed. Gene Set Enrichment Analysis (GSEA) is a powerful analytical method for interpreting gene expression data at the level of gene sets (91). Genes are first ranked in a list based on the correlation between their expressions. An enrichment score (ES) is calculated that reflects the degree to which a set of genes is overrepresented at the extremes (top or bottom of list). The goal of GSEA is to determine whether members of a gene set are randomly distributed among the gene list or if they are primarily found toward the top or bottom of the differential expression list. Sets related to phenotypic distinction should show the latter distribution (91). An advantage of GSEA is its ability to detect biological processes, such

as metabolic pathways, transcriptional programs, and stress responses, that are distributed across an entire network of genes which are subtle at the level of individual genes (91).

Babelomics (http://www.babelomics.org) is a website that aims to provide the scientific community with an advanced set of tools for the functional profiling of high-throughput transcriptomic, genomic, and proteomic data (89). Babelomics consists of a complete suite of methods for the analysis of gene expression data that include normalization, pre-processing, differential gene expression, predictors, clustering, and large scale genotyping assays. These analysis facilities are integrated and then connected to multiple options for the functional interpretation of experiments. Various methods of functional enrichment can be used to understand the functional basis of an analyzed experiment in Babelomics.

The Database for Annotation, Visualization, and Integrated Discovery (DAVID) Knowledgebase (http://david.abcc.ncifcrf.gov/) is a large gene-centered knowledgebase that integrates the most useful and highly regarded heterogeneous annotation resources in a centralized location with improved cross-referencing capability between NCBI and UniProt systems (92). DAVID provides a comprehensive set of functional annotation tools for researchers to understand biological meaning behind large lists of genes. The DAVID Knowledgebase is built around a single-linkage method to agglomerate tens of millions of gene/protein identifiers from a variety of public genomic resources into DAVID gene clusters, thus improving cross-reference capability and enabling more than 40 publicly available functional annotation sources to be comprehensively integrated and centralized (92). This helps to aid researchers in focusing on data analysis or the core development of new high-throughput functional data mining algorithms. For any given gene list, DAVID tools are capable of identifying enriched biological themes (predominantly GO terms); discovering enriched

functional-related gene groups; clustering redundant annotation terms; visualizing genes on BioCarta and KEGG pathway maps; listing interacting proteins; exploring gene names and redirecting to related literatures; and converting gene identifiers from one type to another (92). The identification of enriched functional categories relies on a EASE score that is based on Fisher's test with a higher level of stringency. In addition to reporting the functional categories that are enriched among a list of genes uploaded, DAVID includes an algorithm that further clusters functional categories with similar gene composition. This additional clustering level minimizes the risk of over-counting related enriched categories that span across levels of the GO hierarchy or across databases (92). The entire DAVID Knowledgebase is freely downloadable in simple, pair-wise, text format files.

KEGG, or the Kyoto Encyclopedia of Genes and Genomes, is a database resource for understanding high-level functions and utilities of the biological system. It consists of the molecular building blocks of genes and proteins and chemical substances that are integrated with the knowledge on molecular wiring diagrams of interaction, reaction, and relation networks. KEGG is a reference knowledgebase for the integration and interpretation of large scale molecular data sets generated through genome sequencing and other experimental technologies. The KEGG database is divided into three categories, which are subdivided into sixteen databases. The first category contains systems information with databases including: KEGG PATHWAY, KEGG BRITE, KEGG MODULE, KEGG DISEASE, KEGG DRUG, and KEGG ENVIRON. Genomic information makes up the second category with KEGG ORTHOLOGY, KEGG GENOME, KEGG GENES, and KEGG SSDB as databases. The final category is chemical information and includes databases: KEGG COMPOUND, KEGG GLYCAN, KEGG REACTION, KEGG RPAIR, KEGG RCLASS, and KEGG ENZYME. Each of the databases

contain various data objects for computer representation of biological systems. Object identifiers are used to distinguish the database entry of each KEGG object and consist of a database dependent prefix and a five digit number (93).

Repositories of microRNA, transcription factors, Gene Ontology, pathway information

Network visualization

Visualization is a technique used to graphically represent sets of data. When data is large or abstract, visualization can help make the data easier to read or understand (94). Network visualization allows for the studying of networks of interactions within various biological systems by using tools to assist in modeling of network data.

Cytoscape (http://www.cytoscape.org/) is an open source software platform for visualizing complex networks designed to fill the strong demand for an easy, powerful, general-purpose, and extensible network interaction modeler (95). This resourcesupports the integration of large biological data sets and assists in the exploration of network models that are major parts of systems biology. By working as a web service client, Cytoscape can directly connect to external databases and import network and annotation data. Currently, external databases supported include Pathway Commons (http://www.pathwaycommons.org/pc/), IntAct (http://www.ebi.ac.uk/intact/), BioMart (http://www.biomart.org), NCBI Entrez Gene (http://www.ncbi.nlm.nih.gov/gene), and Protein Identifier Cross-Reference (PICR, http://www.ebi.ac.uk/Tools/picr/). Basic visualization consists of a set of nodes connected by edges that describe the type of interaction occurring between them (95). By attaching annotations to a network, a user can access these annotations to alter visualization of a network through various colors, shapes, images, sizes, or layout. Networks can be exported as

publishable-quality images, and PDF, EPS, SVG, PNG, JPEG, and BMP files are all supported (95).

Visual depiction of networks can be simplified with the use of certain plugins. The GPML-plugin (http://apps.cytoscape.org/apps/gpmlplugin) can be used in Cytoscape and is a converter between Cytoscape networks and the GPML (GenMAPP Pathway Markup Language) pathway format (96),. Via an import menu, this plugin allows a webservice client to directly search and open pathways represented on WikiPathways (http://www.wikipathways.org) as well as save the GPML pathways (97, 98), WikiPathways is an open, public platform dedicated to the curation of biological pathways by and for the scientific community (97, 98).

Another tool for gene network building and visualization is the BisoGenet plugin from the Cytoscape software (99). The BisoGenet system consists of three tiers: a data tier, a middle tier, and a client tier. In the data tier, genomic information, protein-protein interactions, protein-DNA interactions, gene ontology, and metabolic pathways are all stored in a database (99). The middle tier consists of a global network of all data on bioentities and their relationships from this database (99). The client tier is the actual Cytoscape plugin which allows for user input, communication, visualization and analysis of the resulting network (99). BisoGenet's database, SysBiomics, integrates data from multiple public domain datasets such as BIND, HPRD, Mint, DIP, BioGRID or Intact NCBI, UniProt, KEGG, and GO. With BisoGenet one can build and visualize biological networks along with distinguishing the relations between genes and their products.

Network reconstruction

An effective approach to rapidly and maximally leverage available transcriptome and phenotypic (e.g. cancer survival) data is needed. This will augment our understanding of the disease to develop the best diagnostic and treatment practices for patients with specific tumor characteristics, in terms of genetics (the potential for disease outcome), disease biology (how the potential has played out up to the point of measurement), and the connections between these and the clinical outcome. The main challenge in systems biology is uncovering the complex gene regulatory network that renders the phenotype or disease state of a biological system in response to various environmental cues (100). A proper incorporation of multi-source biological knowledge is beneficial for network reconstruction. The complex interaction between genes and the environment that govern the cellular response cannot be understood at the level of individual components of the network, but emerges through the intricate interplay between genes, miRNAs, and TF (100). An approach to tackle the integration of transcriptomic and phenotypic data is reverse-engineering. This approach encompasses network inference directly from a data set that generated probabilistic network models consisting of nodes and directed edges between then nodes (32). Subsequently, the models can be used to simulate transcriptome profiles and make quantitative outcome predictions (101). The edges represent statistical relationships that can be interpreted as casual when perturbations are used in the experiment. Genetic variation across tumors serves as perturbations that explain variation in molecular measurements and survival outcomes and helps to anchor casual influences in the model (32). The nodes consist of the measured molecules in the experiment (in this case, miRNAs, genes, survival outcome, and other relevant clinical covariates such as tumor grade). These models have the ability to automatically span the course of millions of possible scenarios to determine and distinguish the key driver

genes and pathways (potential targets) from the passenger entities with respect to a given genetic background (32). This approach of analyzing experimental and clinical data sets allows for building models capable of subject-specific predictions.

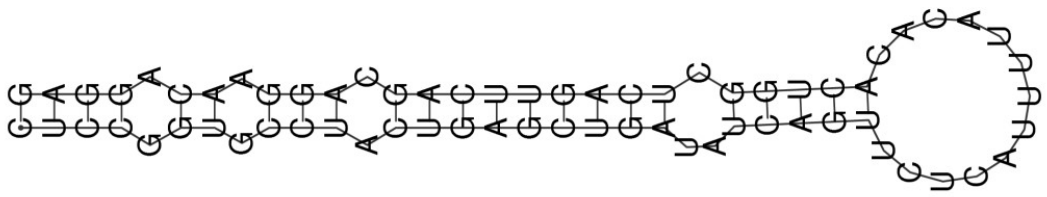
Cellular phenotypes are determined by the dynamical activity of large networks of co-regulated genes, and thus dissecting the mechanisms of phenotypic selection requires clarifying the function of individual genes in the context of the networks in which they operate. ARACNE, An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context (http://wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE) , is a novel information-theoretic algorithm for the reverse engineering of transcriptional networks from microarray data (102). This method uses an information theoretic approach to eliminate the vast majority of indirect interactions typically inferred by pair-wise analysis. Relationships between nodes (e.g. transcripts, proteins) are identified based on the capability of the co-variation (e.g. co-expression) patterns to augment the information (assessed as a function of the log-transform probability) in the data, conditional on other nodes. ARACNE focuses on identifying direct transcriptional interactions in cellular networks in order to elucidate functional mechanisms that underlie cellular processes. ARACNE has a low computational complexity, does not require discretization of expression levels, and does not rely on a priori assumptions (102).

Thesis motivation

The study of the dysregulation of the transcriptome in diseases like cancer and drug abuse can offer insights into genes, mRNA and proteins that can be the target of preventive and therapeutical remedies. Several reasons may be behind the limited validation of reported associations between these diseases and miRNAs, TFs, and target mRNA in independent studies.

First, few studies simultaneously analyze multiple miRNAs, TFs, and target mRNA. Second, most studies do not consider clinical or cohort-dependent factors when characterizing associations between the transcriptome and disease. Lastly, most transcriptome studies are small and the individual analysis has limited statistical power to detect accurate and precise associations between transcripts and diseases. The overall goal of this thesis is to address the previous limitations and identify replicable biomarkers of cancer and drug abuse.

Figures



```
((((.((.(((.(((((((((..(((((.............)))))..)))))))))..))).))..))))
CUCCGGUGCCUACUGAGCUGAUAUCAGUUCUCAUUUUACACACUGGCUCAGUUCAGCAGGAACAGGAG
```

Figure 1.1 Typical hairpin structure and corresponding secondary structure sequence of miRNA precursor. Base pairings are represented by complementary parentheses and non-paring bases by dots. Human miRNA mir-24-1 is shown.

Source: Brameir and Wiuf (2007) (103)

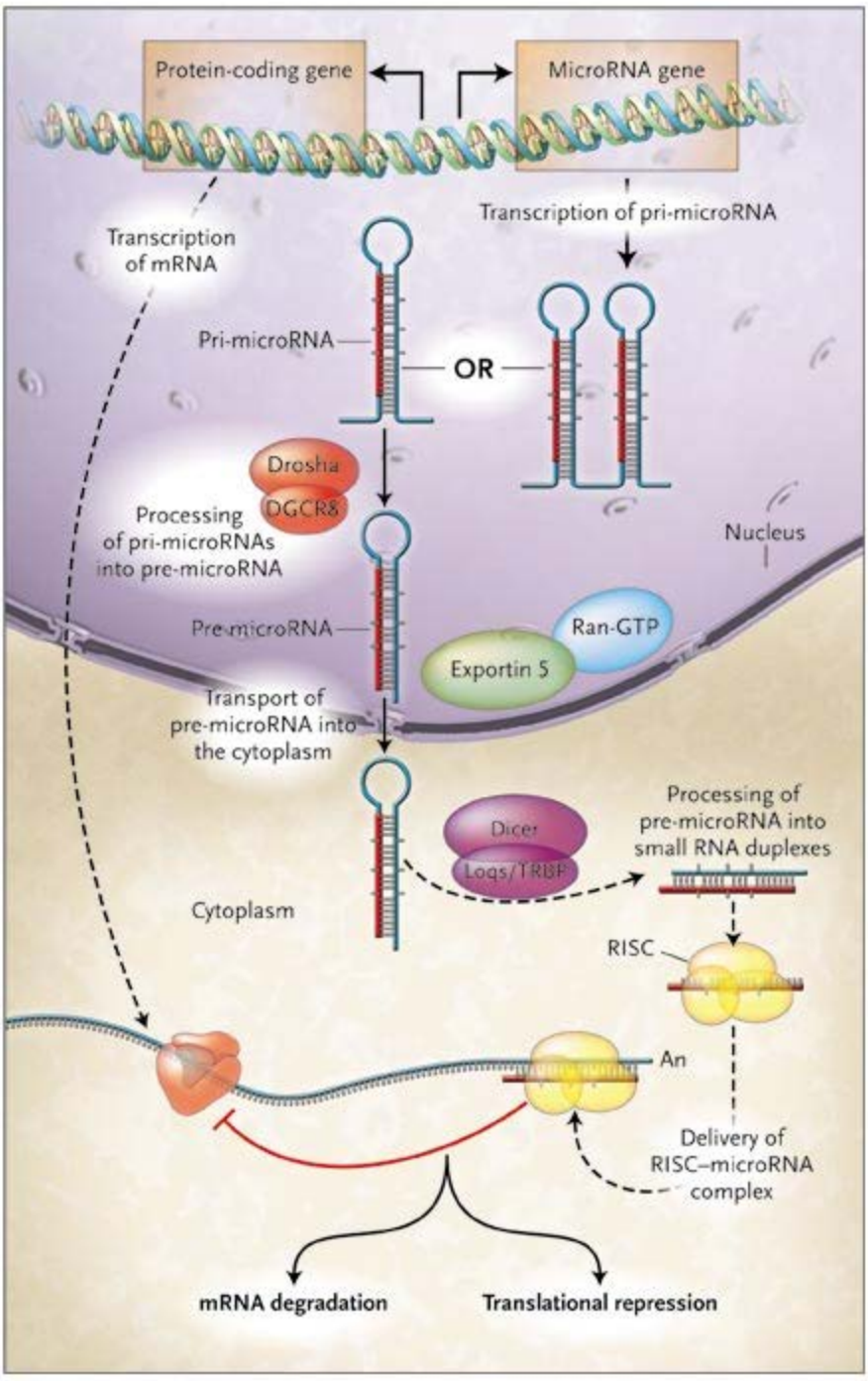


Figure 1.2 Biogenesis of miRNA and miRNA-mediated gene regulation in animal cells.

Source: Chen (2005) (104)

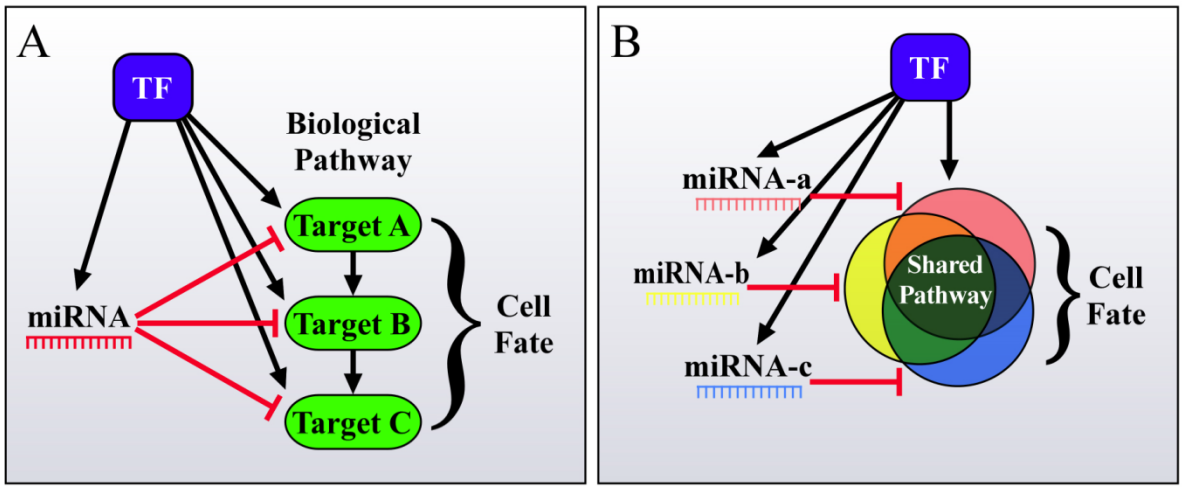


Figure 1.3 A) miRNAs and TFs in FFLs tend to mutually target genes from the same pathway. **B)** Co-regulated miRNAs and miRNA families co-target many genes in the same pathway resulting in a significant total output which has a major effect on cell fate.

Source: Shalgi et al. (2009) (105)

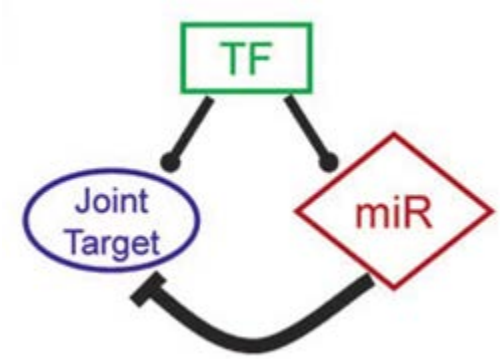


Figure 1.4 Feed-forward loops**.** Representation of a typical mixed feed-forward loop. In the square box, TF is the master transcription factor; in the diamond box, miR represents the miRNA involved in the circuit; in the round box, the Joint Target is the joint protein-coding target gene. Inside each circuit, ⊣• indicates transcriptional activation/repression, and ⊣ indicates post-transcriptional repression.

Source: Re et al. (2009) (106)

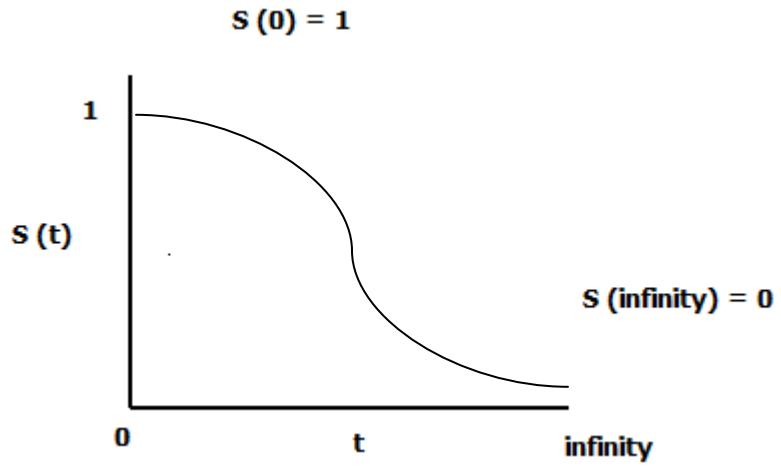


Figure 1.5 Probability of survival across time *t*. Theoretical situation where probabilities are non increasing, with $S(t) = 1$ when $t = 0$, and $S(t) = 0$ when $S(t) = \infty$.

Source: Kleinbaum (1996) (80)

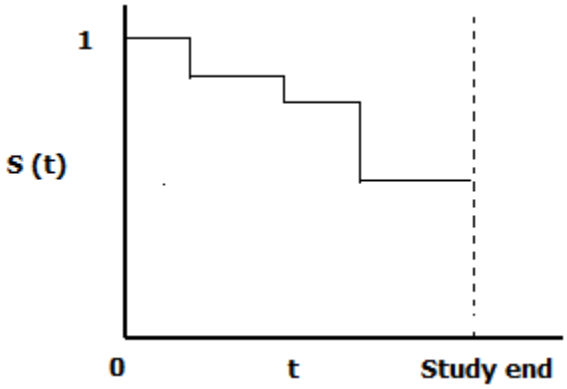


Figure 1.6 Probability of survival across time *t*. Practical situation where not all individuals show occurrence of the event and, thus, the estimate survival function does not go to zero at the end of the study.

Source: Kleinbaum (1996) (80)

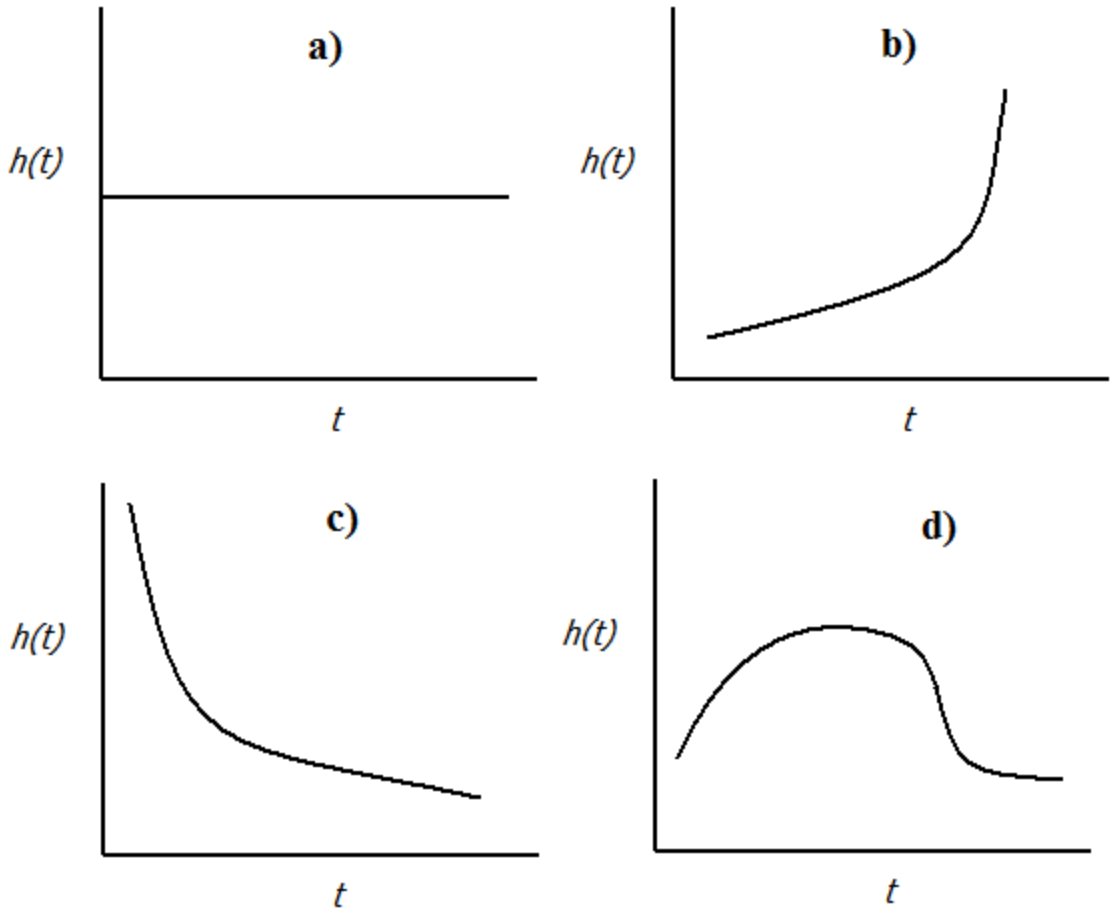


Figure 1.7 Hazard functions across time. For example, the hazard function **a**) may be constant, where no matter what value of t is specified, h(t) equals the same value; **b**) may increase over time, whereas survival time increases, potential for the event occurring increases; **c**) may decrease over time, whereas survival time increases, potential for the event occurring decreases; or **d**) may first increase and then decrease.

Source: Kleinbaum (1996) (80)

Chapter II

Therapy-, Gender- and Race-specific microRNA Markers, Target Genes, and Networks Related to Glioblastoma Recurrence and Survival*

ABSTRACT

Aim: To identify and study targets of microRNA biomarkers of glioblastoma survival across events (death and recurrence) and phases (life expectancy or post-diagnostic) using functional and network analyses. **Materials and methods:** MicroRNAs associated with glioblastoma survival within and across race, gender, recurrence, and therapy cohorts were identified using 253 individuals, 534 microRNAs, Cox survival model, cross-validation, discriminant analyses, and cross-study comparison. **Results:** All 45 microRNAs revealed were confirmed in independent cancer studies, and 25 of those were further confirmed in glioblastoma studies. Thirty-nine and six microRNAs (including hsa-miR-222) were associated with one and multiple glioblastoma survival indicators, respectively. Nineteen and 26 microRNAs exhibited cohort-dependent (including hsa-miR-10b with therapy and hsa-miR-486 with race) and independent associations with glioblastoma, respectively. **Conclusion:** Sensory perception and G protein-coupled receptor processes were enriched among microRNA gene targets also associated with survival, and network visualization highlighted their relations. These findings can help to improve prognostic tools and personalized treatments.

*K.R. Delfino, N.V.L. Serao, B.R. Southey, and S.L. Rodriguez-Zas. Cancer Genomics Proteomics. 2011 Jul-Aug;8(4):173-183.

Introduction

Glioblastoma multiforme (World Health Organization glioma grade IV) is a primary and aggressive cancer. Glioblastoma patients have a median survival of less than one year, and the incidence of glioblastoma varies among cohort groups, such as race and gender (8, 107). Some genes and microRNAs, small non-coding RNA molecules that can affect the post-transcriptional regulation of genes, exhibit abnormal expression patterns in glioblastoma (108, 109). Data and methodological limitations have prevented the identification of consistent microRNA biomarkers of glioblastoma survival that could be used to develop effective prognosis and diagnostic tools and therapies. Data limitations mostly encompass small data sets with unknown or restricted representation across cohort groups and consideration of a single glioblastoma survival indicator. Methodological limitations include arbitrary discretization of response (*e.g.* high and low survival) and explanatory (*e.g.* high or low expression level) variables (110), single-microRNA analysis (17, 111), pre-selection of microRNAs, and use of approaches that cannot accommodate the multifactorial nature of the disease.

The main objective of this study was to identify microRNAs that are reliable indicators of glioblastoma survival and recurrence using survival analysis. The study also aimed at extending the findings to microRNA target genes, their biological processes, molecular functions, and networks. Another goal was to identify and profile cohort-dependent associations between microRNAs and glioblastoma that can be used in personalized therapies.

Materials and methods

Survival, cohort, recurrence, and microRNA information from 253 individuals diagnosed with glioblastoma and death and recurrence records between the years 1990 and 2008 was considered. Surgical samples corresponded to newly diagnosed glioblastoma cases, had a minimum of 80% tumor nuclei and a maximum of 50% necrosis (112). The data was obtained from The Cancer Genome Atlas (TCGA) December 2009 data freeze (113). Cohort factors were gender (male or female), race (white

Caucasian or not), therapy received (radiation therapy alone, RX; chemotherapy plus radiation and no targeted therapy, CRN; chemotherapy plus radiation and targeted therapy, CRT; and all other therapies including no therapy, OTHER), and the detection of glioblastoma recurrence or progression after the original diagnosis (progression/recurrence or not).

Prognostic microRNA biomarkers for two events (death and recurrence) and two phases (from birth to event or from diagnostic to event) were studied through three complementary glioblastoma survivals: life expectancy (years from birth to death associated with glioblastoma), post-diagnostic glioblastoma survival (months from glioblastoma diagnostic to death), and post-diagnostic glioblastoma recurrence or progression (or post-diagnostic recurrence hazard, encompassing the months from glioblastoma diagnostic to reports of progression or recurrence). The last two indicators are also known as overall survival (OS) and progression-free survival (PFS) and offer complementary information to life expectancy (LE). The models used to describe the three indicators are specified in terms of hazard (instead of survival) and thus, hazard or survival is used where appropriate. **Table 2.1** summarizes the number and distribution of individuals studied across levels of the covariates considered in the model. The median age at diagnosis was 55.7 years. Expression levels of 534 microRNAs were measured using the Agilent 8 × 15K Human microRNA platform (Agilent Technologies, http://www.genomics.agilent.com/). The data was quantile-normalized (at the probe level), collapsed within microRNA, and log 2-transformed following the procedures described in Beehive (114).

A glioblastoma hazard predictive model that simultaneously considered all microRNAs and cohort information was used to identify general and personalized (or cohort-independent and -dependent, respectively) biomarkers. This model overcame the limitation of previous studies that ignored the simultaneous association between multiple microRNAs and glioblastoma hazard by analyzing only one microRNA at a time or ignoring cohort information that may result in cohort-dependent relationships between glioblastoma events and microRNA profiles by predicting hazard

within clinical factor levels. The factorial and complex nature of the data, including more microRNAs than individuals and a large number of cohort variables, required the novel integration of multiple statistical and computational tools into a biomarker identification pipeline.

Statistical computing method

A Cox survival model together with leave-one-out cross-validation (LOOCV) and discriminant analyses were used to identify microRNA expression profiles associated with glioblastoma survival. This model accommodates censored data resulting from individuals that are alive or that do not have a recurrence record at the end of the period analyzed. The Cox proportional hazard model assumes a parametric model to test the association between the covariates and the hazard ratio of the event. The model does not require the specification of a baseline hazard rate or an estimate of absolute risk, and thus, this non-parametric component of the model does not require the specification of the absolute shape of the curve formed by the two hazard rates (e.g. female versus male) over time. The test of no association between the microRNAs or cohort prognostic markers and the hazard ratio between gender, race, therapy, or recurrence groups and the 95% confidence interval limits follow a Chi-square distribution. The Cox model assumes proportional hazards across the period studied. This assumption can be expressed as parallel survival functions across the levels of the microRNA or cohort variables in the model. This assumption was tested for each of the three hazards considered, and there was no indication of significant departure from the assumption. Furthermore, visualization of the survival and residuals did not suggest departure from the model assumptions. There was no indication of significant departure from the proportional hazards assumption, also confirmed by the overlap on microRNAs between survival indicators.

A multi-step strategy was undertaken to identify and validate microRNA prognostic markers of glioblastoma survival or recurrence. Cohort variables, microRNAs, and interaction terms were included

simultaneously in a Cox model, and a combination of forward and stepwise model selection methods were used to identify association for each survival.

Three model selection methods were used because of their complementary advantages. In the first method, a stepwise model selection was run on all microRNAs and included cohort variables (entry into the survival predictive model conditional on the other biomarkers already in the model or entry P-value < 0.3 and remain in the survival predictive model after consideration of other biomarkers or remain P-value < 0.1). Next, a forward model selection was applied to all microRNAs and included cohort variables (entry P-value < 0.3). The significant microRNAs from the previous stepwise and forward model were identified and included in a model that was subjected to a stepwise selection (entry P-value < 0.3 and remain P-value < 0.1). This step allowed us to identify broad or general associations between microRNA profiles and glioblastoma hazards. The relaxed P-value threshold allowed us to detect the microRNA that may first have weak associations among large sets of microRNAs but stronger associations become apparent as the set is streamlined.

In the first step of the analysis, the Cox model was used to identify the cohort groups and microRNA profiles associated with the three complementary glioblastoma hazards. The complete microRNA set was analyzed using the Cox model because of the goal of identifying systematic and cohort-dependent microRNA associated with the three glioblastoma events. The associations between cohort and microRNA biomarkers and prognosis, while adjusting for the other systematic variables, were assessed. This strategy allowed to ensure a sufficient number of observations per cohort level in the model and to uncover prognostic microRNAs that exhibit unique patterns in cohort groups with moderate representation in the complete data set, such as individuals receiving uncommon therapies.

In the second step, the interaction between the microRNAs remaining in the model after the biomarker selection and cohort indicators was evaluated using the stepwise approach. This step allowed the identification of cohort-dependent associations between microRNA profiles and the hazard

of death or glioblastoma progression. The first and second steps were implemented separately for all of the microRNA measured in the platform and for 14 microRNAs that have been reported to be associated with glioblastoma (hsa-miR-21, hsa-miR-221, hsa-miR-222, hsa-miR-181a, hsa-miR-181b, hsa-miR-7, hsa-miR-128, hsa-miR-124, hsa-miR-137, hsa-miR-451, hsa-miR-10b, hsa-miR-129, hsa-miR-139 and hsa-miR-218) (8).

In the third step, the results from the analyses of all microRNAs and of previously reported microRNAs were combined and further streamlined using the stepwise method. The combination of the biomarkers identified from considering all microRNAs in the platform and only microRNAs known to be associated with glioblastoma facilitated the detection of novel microRNA biomarkers and verification of known biomarkers. The output of this comprehensive methodological tactic was a cohort and microRNA index to prognosticate glioblastoma. The association between the glioblastoma hazards and the cohort factors and microRNA expression profiles was visualized by plotting the probability of survival predicted from the Cox model against time.

Following common practice, the resulting microRNAs were evaluated using a LOOCV approach (115-117) and classification analyses (118-121). LOOCV is recommended especially for data sets of limited size, providing an almost unbiased estimator and identifying the same best classifiers as other X-fold training-test data partitions (115-117). For the X-fold validation approach, the specification of suitable training and testing data sets would have required at least 160 patients in each data set (5 patients $\times$ 2 races $\times$ 2 genders $\times$ 4 therapies $\times$ 2 recurrence groups) and only 253 patients were available. Use of smaller data sets would have lead to low power and biased findings because of the ill-representation of patients across cohort groups. Patients were classified into high and low survival groups using the median time at the glioblastoma event (death or recurrence) as a cutoff and removing patients with unclear hazard within one unit of the median. Only non-censored records were used to avoid biased classification estimates. Preliminary results from linear and quadratic

discriminant, logistic, and k nearest-neighbor analyses were consistent, and quadratic discriminant results are presented.

Validation of the results from the Cox model, LOOCV, and classification analyses on an independent data set was not feasible because no other data set has information on gender, race, therapy, recurrence, and age that would allow testing the cohort-dependent microRNAs identified in this data set. Thus, a two-fold approach was used to offer corroboration of our findings. First, the microRNA biomarkers identified in this study were searched against the glioblastoma multiforme and cancer literature based on independent data sets. Second, the expression profile of the targets genes of the microRNAs were analyzed (122). The gene targets corresponding to the microRNAs associated with glioblastoma survival were obtained from MicroCosm (20, 21). Expression measurements for the target genes were available from the same patients using the Affymetrix HT HG-U133A platform. The normalization and Cox survival models used for the gene targets were the same as described for the microRNAs. The target genes subsequently used had a significant association (P-value < 0.001) with either glioblastoma OS, PFS, or LE (122). Functional Gene Ontology (GO) and KEGG Pathway analysis of the significant target genes of the significant microRNAs was undertaken (93, 123). The enrichment of functional categories was evaluated using Fisher's exact (two-tailed) test and false discovery rate (FDR) multiple test adjustment (89). Network visualization was accomplished by depicting all pair-wise relationships between target genes using the BisoGenet plugin from the Cytoscape software (99). BisoGenet's database, SysBiomics, integrates data from multiple public domain datasets such as BIND, HPRD, Mint, DIP, BioGRID or Intact NCBI, UniProt, KEGG, and GO. Based on this information, a global network of relations among microRNA target genes was created and visualized using Cytoscape. The network was inferred using only significant target genes (circular network nodes) of significant microRNAs associated with either glioblastoma survival. Only interactions (network edges) connecting two target genes directly or through an intermediate gene (square gray node) were portrayed to facilitate the visualization of relationships and minimize the

incorporation of relationships not relevant to the microRNA biomarkers detected in this study. Known gene relationships depicted in the network are summarized in the SysBiomics repository (99).

Results

The median length of glioblastoma LE, OS, and PFS was 59 years, 13 months, and 6 months, respectively. The probabilities of survival at 12, 24, 36, and 48 months post-diagnostic were 0.55, 0.26, 0.16, and 0.11, respectively. Survival length indicators confirm previous reports that most TCGA samples correspond to primary glioblastoma with null or a very minor percentage of secondary glioblastoma cases (8, 124). MicroRNAs associated with the three glioblastoma survivals are listed in **Tables 2.2** to **2.4**, respectively. Hazard ratio estimates >1 indicate an increase in the hazard (decrease in survival probability) per unit increase in the level of microRNA expression, and hazard ratio estimates <1 denote the opposite trend, conditional on all other cohort and microRNA predictors in the model. Additional statistical significance indicators include the 95% confidence interval of the hazard ratio estimate and the statistical significant P-value.

Tables 2.2 to **2.4** also list previous studies that support the association between the microRNAs and glioblastoma identified in this study. Corroborating our findings, the majority of microRNAs associated with glioblastoma survival (25 out of 45 microRNAs) have also been associated with glioblastoma in independent studies, and the rest (20 microRNAs) have been associated with other types of cancer (**Tables 2.2** to **2.4**). MicroRNAs in two families (hsa-miR-181 and hsa-miR-34 family) and six other microRNAs were associated with multiple survival indicators, while 35 microRNAs were associated with one survival indicator. The same number of positive and negative associations (HR >1 or HR <1) between microRNA expression levels and the three glioblastoma hazards studied were revealed in this study (**Tables 2.2** to **2.4**). Twenty-six and 19 microRNAs had cohort-independent and -dependent relationships with glioblastoma survival, respectively. The survival plot in **Figure 2.1** depicts the lower post-diagnostic survival probability of females that have a low level of microRNA

ebv-miR-bhrf1-1 relative to males with a high expression level. Three microRNAs (hsa-miR-10b, hsa-miR-222, and hsa-miR-140) exhibited different hazard ratio trends across glioblastoma indicators, and the associated confidence interval allowed the identification of the trend best supported by the data.

Integration of Cox survival model, LOOCV, and discriminant analysis supported the correct classification of 98% and 93% of the patients into the high and low post-diagnostic survival or OS groups, respectively, and the area under the receiver operator characteristic (ROC) was 94%. Likewise, 100% and 91% of the individuals in the high and low PFS groups were correctly classified, and the area under the ROC was 97%. Finally, 86% and 75% of the patients in the high and low LE groups were correctly classified, and the area under the ROC was 85%. Another indicator of the reliability of the integrated approach is that all microRNAs detected in this study have been previously associated with cancer, and a majority were associated with glioblastoma in independent studies. An additional indicator supporting the microRNAs identified is that 239, 418, and 336 gene targets of the microRNAs were significantly associated with LE, OS, and PFS, respectively.

Several GO categories were enriched (FDR-adjusted *P*-value < 0.05) among the target genes significantly associated with multiple survival indicators. **Tables 2.5** to **2.7** summarize these findings, with the latter table including an FDR-adjusted *P*-value <0.01 and a minimum of six genes due to space limitations. Categories are sorted by GO theme, followed by level and *P*-value. The GO categories enriched across all three survival indicators included sensory perception (of chemical stimulus and smell), neurological process, olfactory receptor activity, rhodopsin-like receptor activity, and transmembrane receptor activity. All GO categories enriched in the post-diagnostic death or OS were also identified in either or both of the remainder indicators. **Figure 2.2** portrays the network including target genes (denoted in pink) of significant miRNAs that also themselves have a significant association with either glioblastoma OS or PFS, and have a minimum of one relationship and at most one indirect relationship with other target genes.

Discussion

All microRNAs associated with glioblastoma survival detected in this study have been confirmed to be associated with glioblastoma in previous independent studies (25 microRNAs) or have been associated with other cancer types (20 microRNAs). This extensive confirmation, the large number of target genes also significantly associated with glioblastoma survival, and the correct classification of patients into survival groups further supports the robustness of our findings. The equal number of positive and negative associations between microRNA expression levels and survival, and the fact that 17% of the microRNAs exhibited associations with multiple glioblastoma survival indicators confirm the paradigm that glioblastoma initiation and recurrence are impacted by microRNAs targeting a wide range of oncogenes, tumor suppressor genes, and pathways at different stages of tumor genesis and growth (125). Some of these genes and pathways are activated or silenced by microRNAs acting throughout the glioblastoma phases; meanwhile, others are phase or event dependent. Likewise, the equal number of positive and negative associations between the microRNA expression levels and hazard of the three glioblastoma events investigated are in agreement with previous reports on the complex and multifaceted regulation of cancer initiation and progression by microRNAs (125). Most microRNAs (64%) exhibited a broad, cohort-independent relationship with glioblastoma survival. This indicates that mainstream and general practices to treat glioblastoma on the basis of microRNA profiles alone are promising. The identification of gender-, race-, and therapy-dependent microRNA biomarkers indicates that general practices can be effectively complemented with personalized practices. The following discussion of the microRNA biomarkers focuses on novel and high impact discoveries, and relevant supporting references for all other microRNAs are listed in **Tables 2.2** to **2.4**.

Higher levels of Kaposi's sarcoma-associated herpes virus (kshv) miR-k12-1 were associated with all three glioblastoma survival indicators (**Tables 2.2** to **2.4**) in agreement with associations between this microRNA and two B-cell-derived cancer types (14). MicroRNAs ebv-miR-bhrf1-1, hsa-

miR-565, hsa-miR-137, and hsa-miR-512-3p had gender-, race-, and recurrence-dependent associations with OS and PFS (**Tables 2.3** and **2.4**). For these four microRNAs, cohort-independent trends in the same direction were reported respectively for Burkitt's lymphoma, ovarian cancer, chemoradiation-treated rectal cancer, and for both metastatic pancreatic ductal adenocarcinoma cell lines and hepatocellular carcinoma cells that have been linked to the inhibition of the tumorgenesis factor c-FLIP (126-128). Likewise, the gender-, therapy- and race-dependent associations between hsa-miR-93, hsa-miR-489, human cytomegalovirus (hcmv) miR-ul70-3p, hsa-miR-758, hsa-miR-143, and PFS (**Table 2.4**) have been confirmed at a cohort-independent level for T-cell leukemia, breast-cancer MCF-7 cells resistant to tamoxifen, tumors from various tissues (*e.g.* breast, colon, liver), a multidrug-resistant variant of a human gastric adenocarcinoma cell line, and for both B-cell chronic lymphocytic leukemia and colorectal cancer cell growth through inhibition of *KRAS* translation (129-134). The cohort-independent and gender-dependent association of hsa-miR-222 with OS and PFS (**Tables 2.3** and **2.4**, respectively) confirm the results of Ciafre *et al.* (135). The therapy-dependent association between glioblastoma and members of the hsa-miR-181 and hsa-miR-34 families (**Tables 2.2** to **2.4**) are consistent with previous reports (3, (17, 136). High levels of hsa-miR-140 were associated with higher LE and lower and therapy-dependent OS (**Tables 2.2** and **2.3**). The multiple modes of action of hsa-miR-140 are consistent with reports of up-regulation in most glioblastoma cases (136), inhibition of cell proliferation in osteosarcoma and colon cancer cell lines (137), and treatment-dependent action (137). Reanalysis of the association between glioblastoma survival and hsa-miR-140 alone (with and without cohort factors, results not shown) produced trends similar to that in the multi-microRNA models. Thus, our results suggest that the influence of hsa-miR-140 on glioblastoma survival may vary with the glioblastoma phase considered. A gender-dependent association between hsa-miR-26a and OS was uncovered (**Table 2.3**). The general trend is consistent with the proposed role of hsa-miR-26a promoting glioblastoma cell growth and formation (135, 138), and the gender-dependent model is in

agreement with the higher expression of hsa-miR-26a in women than in men diagnosed with hepatocellular carcinoma (139).

Additional analyses resolved the apparent inconsistencies in the trends between previous reports and our study for seven microRNAs hsa-miR-182 (140), hsa-miR-106b (141), ebv-miR-bart7 (142), hsa-miR-189 (143), hsa-miR-221 (144), hsa-miR-21 (145), and hsa-miR-10b (17, 140, 141). For hsa-miR-182, hsa-miR-106b, and hsa-miR-221, the individual microRNA analysis supported the multi-microRNA results. For ebv-miR-bart7, hsa-miR-189, and hsa-miR-10b, the individual analysis did not detect a significant trend. In one case, hsa-miR-21 was not detected when considered simultaneously with other microRNAs but was significant when considered alone, in agreement with Chan *et al.* (145). These results suggest that identification of biomarkers on an individual basis may result in spurious associations and also validates the approach used in this study to identify biomarkers that simultaneously considers multiple microRNAs.

The large number of gene targets of the detected microRNAs that also exhibited significant association with glioblastoma survival further substantiates our findings. Sensory perception, neurological process, olfactory receptor, and transmembrane receptor activity were among the processes and functions consistently over-represented among the target genes of microRNAs associated with all three glioblastoma survival indicators. The neurological and sensory perception processes are consistent with reports of glioblastoma candidates for single nucleotide polymorphisms of sensory perception genes and with reports that individuals with brain tumors lose sensory perception (146). Oncogenes act by mimicking the growth signals transmitted by transmembrane receptors (147). G Protein-coupled receptor (GPCR) activity (*e.g.* rhodopsin-like gene) regulates cellular motility, growth and differentiation, and gene transcription; three factors central to the biology of cancer (148). The network of gene targets that have significant association with glioblastoma survival displays known relationships, including many in the signaling pathways that involve GPCR, including MAPK, adipocytokine, chemokine, ErBB, FC epsilon RI, mTOR, neurotrophin, notch, p53,

phosphatidylinositol, RIG-I-like receptor, T-cell receptor, TGF-beta receptor, toll-like receptor, VEGF, and Wnt signaling pathways (**Figure 2.2**).

In summary, this study confirmed 25 microRNAs previously associated with glioblastoma survival and identified 20 other microRNAs that have been previously associated with other cancer types. This confirmation and the high correct classification of patients into survival groups suggests that the biomarkers revealed in this study are good leads for empirical confirmation, improved prognostic tools, and personalized treatments of glioblastoma multiforme. Six and 39 microRNAs were identified as biomarkers of multiple or single glioblastoma survival indicators, respectively, suggesting the multifactorial and multifaceted genomic basis of this cancer. Nineteen microRNAs exhibited gender-, race-, therapy-, or recurrence-dependent associations with glioblastoma survival, suggesting that personalized treatments that consider individual variation can improve the outcome for glioblastoma patients. Sensory perception and GPCR activities are among the processes of the microRNA target genes associated with survival.

Acknowledgements

The support of NCI (Grant Number: 1R03CA143975) to SRZ and KRD, and NIH/NIDA (Grant Number: R21DA027548 and P30DA018310) to BRS and SRZ and USDA NIFA (Number: ILLU-538-554) to NSV are greatly appreciated

Figures

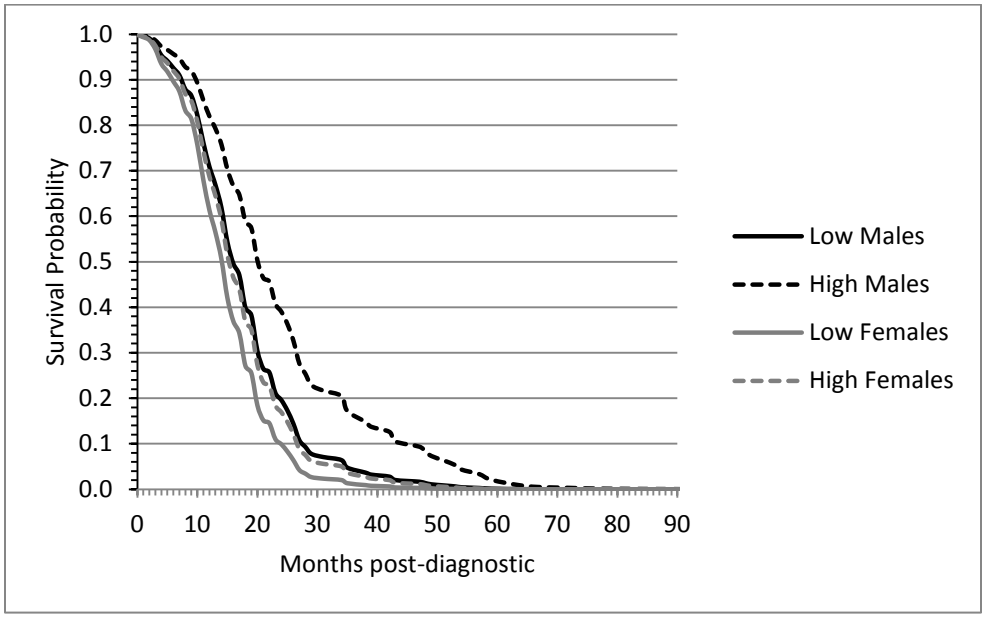


Figure 2.1. Overall survival plots for males (black lines) and females (gray lines) that have high (dashed lines) and low (solid line) levels of ebv-miR-bhrf1-1.

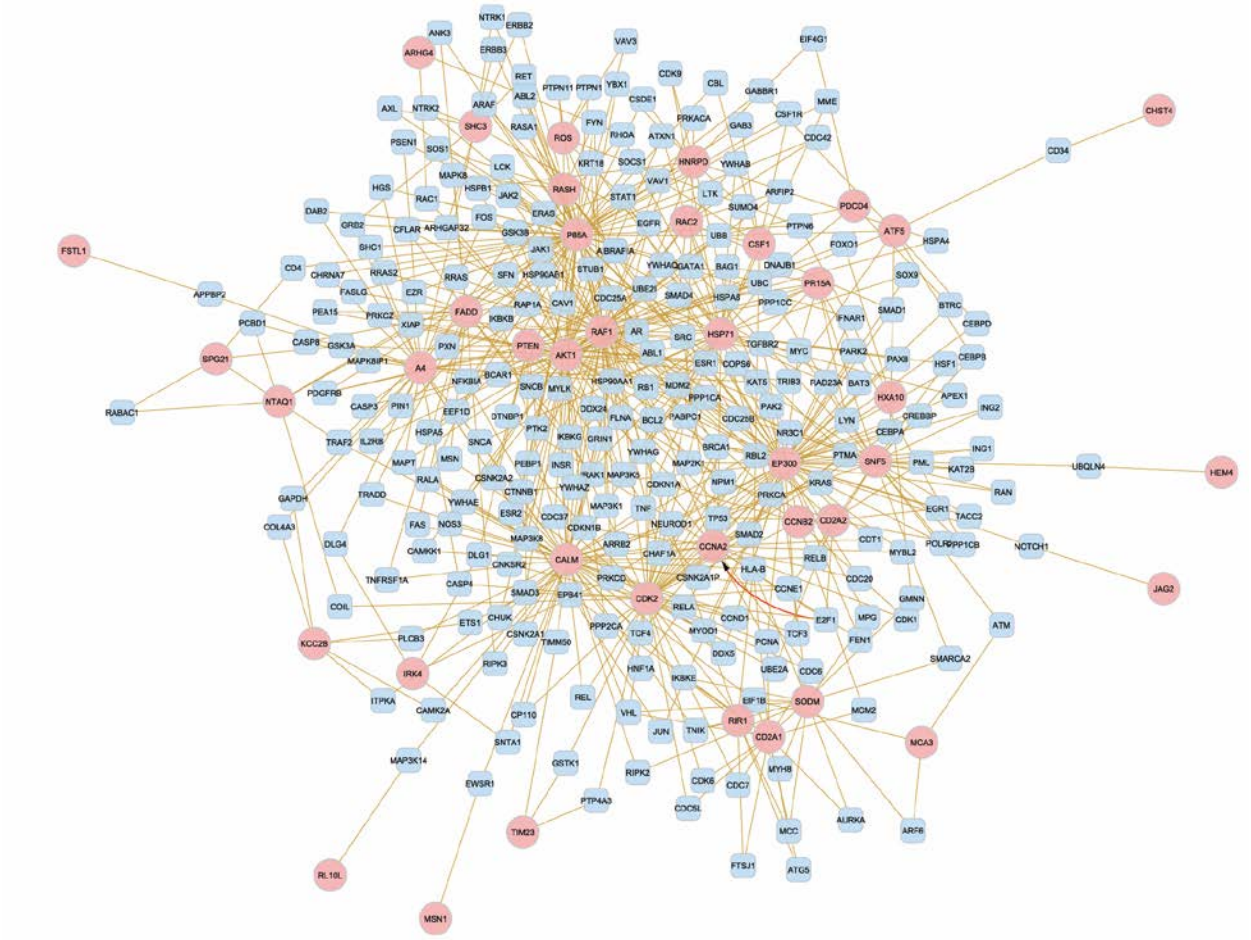


Figure 2.2 Network of target genes of glioblastoma microRNAs.
Circular pink nodes denote target genes of microRNAs associated with glioblastoma survival that also have a significant association with survival themselves. Square gray nodes denote a maximum of one intermediate gene between target genes. Edges denote known relationships between genes from several databases which are summarized in the SysBiomics repository.

Tables

Table 2.1 Number and distribution of individuals analyzed for overall and post-diagnostic hazard of glioblastoma death and post-diagnostic hazard of glioblastoma recurrence and levels of the cohort factors considered.

		Post-diagnostic survival (overall survival) [a]		Post-diagnostic recurrence (progression-free survival)	
		Number	Percentage	Number	Percentage
Total		253		192	
N Censored		26	10%	17	9%
Race	Caucasian	211	83%	161	84%
	Other	42	17%	31	16%
Gender	Females	91	36%	68	35%
	Males	162	64%	124	65%
Therapy	RX	38	15%	32	17%
	CRN	133	53%	113	59%
	CRT	31	12%	28	15%
	OTHER	51	20%	19	10%
Recurrence	Yes	192	76%	192	100%
	No	61	24%	0	0%

[a] The number of patients analyzed for post-diagnostic survival and life expectancy is the same.
N, Number of patients; RX, radiation therapy alone; CRN, chemotherapy plus radiation and no targeted therapy; CRT, chemotherapy plus radiation and targeted therapy; OTHER, all other therapies including no therapy.

Table 2.2 MicroRNAs associated with life expectancy on a cohort-independent or -dependent manner and supporting independent studies.

MicroRNA	*P*-value	Hazard ratio (95% C.I.)	Relevant literature references
hsa-miR-181a*	0.0537	RX=0; 0.33 (0.21 to 0.51)	(108, 109)[G]
		RX=1; 1.05 (0.33 to 3.38)	
hsa-miR-189	0.0204	0.20 (0.05 to 0.78)	(143)[O]
hsa-miR-19b	0.0049	1.46 (1.11 to 1.90)	(149)[G]
hsa-miR-222	0.0258	0.83 (0.70 to 0.98)	(17, 135, 136)[G]
hsa-miR-34a	0.05	RX=0; 0.69 (0.57 to 0.85)	(150, 151)[G]
		RX=1; 1.17 (0.72 to 1.89)	
hsa-miR-550	<0.0001	4.18 (2.31 to 7.56)	(152)[O]
hsa-miR-625	0.0119	2.48 (1.22 to 5.02)	(153)[O]
kshv-miR-k12-1	0.0023	2.08 (1.30 to 3.32)	(14)[O]
hsa-miR-10b	<0.0001	0.74 (0.64 to 0.85)	(17, 140, 141)[G]
hsa-miR-140	0.013	1.57 (1.10 to 2.24)	(136, 137)[G]
hsa-miR-149	0.0056	0.76 (0.63 to 0.92)	(154)[G]

C.I., Confidence Interval; RX=1 denotes radiation therapy alone, RX=0 denotes non-radiation therapy; [G] glioblastoma multiforme study; [O] study on any other type of cancer.

Table 2.3 MicroRNAs associated with overall survival on a cohort-independent or -dependent manner and supporting independent studies.

MicroRNA	*P*-value	Hazard ratio (95% C.I.)	Relevant literature references
hsa-miR-182	0.0245	RX=0: 0.67 (0.57 to 0.77)	(140)[G]
		RX=1: 1.00 (0.71 to 1.38)	
	0.0027	CRT=0: 0.66 (0.56 to 0.77)	
		CRT=1: 1.19 (0.83 to 1.69)	
hsa-miR-189	0.0316	0.12 (0.02 to 0.83)	(143)[O]
hsa-miR-196a	0.0168	1.39 (1.06 to 1.81)	(155)[G]
hsa-miR-221	0.0298	RX=0: 0.67 (0.43 to 1.04)	(17, 135, 136, 144)[G]
		RX=1: 0.41 (0.22 to 0.75)	
hsa-miR-222	<0.0001	2.14 (1.51 to 3.03)	(17, 135, 136)[G]
hsa-miR-23b	0.0135	1.61 (1.10 to 2.35)	(135)[G]
hsa-miR-26a	0.002	Male: 1.33 (1.02 to 1.71)	(135, 138, 139)[G]
		Female: 2.52 (1.78 to 3.58)	
hsa-miR-324-5p	<0.0001	2.73 (1.80 to 4.14)	(156)[G]
hsa-miR-34c	0.0106	0.62 (0.43 to 0.90)	(150, 151)[G]
ebv-miR-bhrf1-1	0.0009	Other: 0.09 (0.01 to 0.51)	(126)[O]
		Caucasian: 1.83 (1.16 to 2.88)	
	0.0008	Male: 0.65 (0.35 to 1.24)	
		Female: 2.77 (1.43 to 5.38)	
hsa-miR-512-3p	0.003	0.28 (0.12 to 0.65)	(153, 157)[O]
hsa-miR-565	0.0996	Other: 2.97 (1.71 to 5.16)	(127)[O]
		Caucasian: 1.80 (1.41 to 2.30)	
	0.0003	Pr/Re=0: 3.80 (2.40 to 6.02)	
		Pr/Re=1: 1.59 (1.27 to 2.00)	

Table 2.3 (con't)

MicroRNA	*P*-value	Hazard ratio (95% C.I.)	Relevant literature references
hsa-miR-572	0.0691	0.76 (0.57 to 1.02)	(158)[O]
hsa-miR-766	0.0052	1.57 (1.15 to 2.16)	(159)[O]
kshv-miR-k12-1	<0.0001	2.77 (1.78 to 4.31)	(14)[O]
kshv-miR-k12-6-3	0.0608	1.54 (0.98 to 2.43)	(160)[O]
hsa-miR-101	0.0065	1.63 (1.15 to 2.32)	(125)[G]
hsa-miR-10b	0.0146	RX=0: 1.16 (0.97 to 1.38)	(17, 140, 141)[G]
		RX=1: 0.74 (0.52 to 1.04)	
hsa-miR-134	0.0007	2.11 (1.37 to 3.25)	(141)[G]
hsa-miR-137	0.001	CRN=0: 2.11 (1.45 to 3.05)	(128)[O]
		CRN=1: 0.94 (0.67 to 1.32)	
hsa-miR-140	0.001	CRN=0: 0.21 (0.12 to 0.37)	(136, 137)[G]
		CRN=1: 0.65 (0.37 to 1.15)	
hsa-miR-148a	<0.0001	1.65 (1.35 to 2.02)	(161)[O]
hsa-miR-409-3p	0.0001	0.43 (0.28 to 0.66)	(162)[O]

C.I., Confidence Interval; RX=1 denotes radiation therapy alone, RX=0 denotes non-RX therapy; CRT=1 denotes chemotherapy plus radiation and targeted therapy, CRT=0 denotes non-CRT therapy; Pr/Re=1 denotes glioblastoma recurrence or progression report, Pr/Re=0 denotes no recurrence or progression report; CRN=1 denotes chemotherapy plus radiation and no targeted therapy, CRN=0 denotes non-CRN therapy; [G] glioblastoma multiforme study; [O] study on any other type of cancer.

Table 2.4 MicroRNAs associated with progression-free survival on a cohort-independent or –dependent manner and supporting independent studies

MicroRNA	*P*-value	Hazard ratio (95% C.I.)	Relevant literature references
hsa-miR-181c	0.0004	CRN=0: 0.27 (0.16 to 0.47)	(17, 136)[G]
		CRN=1: 0.82 (0.53 to 1.35)	
hsa-miR-188	<0.0001	2.30 (1.55 to 3.40)	(140)[G]
hsa-miR-222	0.0814	Male: 1.27 (1.02 to 1.58)	(17, 135, 136)[G]
		Female: 1.65 (1.29 to 2.12)	
hsa-miR-296	0.0247	RX=0: 1.56 (1.14 to 2.14)	(136, 163)[G]
		RX=1: 3.83 (1.82 to 8.07)	
	0.0633	CRT=0: 2.04 (1.51 to 2.76)	
		CRT=1: 0.90 (0.39 to 2.10)	
ebv-miR-bart7	<0.0001	0.05 (0.01 to 0.15)	(142)[O]
hsa-miR-486	0.0168	Other: 0.74 (0.44 to 1.25)	(140)[G]
		Caucasian: 1.53 (1.12 to 2.08)	
hsa-miR-489	0.0041	0.04 (0.00 to 0.36)	(130)[O]
hsa-miR-512-3p	0.0257	Other: 0.00 (0.00 to 0.04)	(153, 157)[O]
		Caucasian: 0.07 (0.02 to 0.28)	
hcmv-miR-ul70-3p	0.0004	Male: 0.43 (0.27 to 0.67)	(131)[O]
		Female: 1.13 (0.71 to 1.79)	
hsa-miR-552	0.0001	0.00 (0.00 to 0.01)	(125)[G]
hsa-miR-578	<0.0001	0.00 (0.00 to 0.00)	(164)[G]
hsa-miR-582	0.0003	5.49 (2.17 to 13.88)	(125)[G]
hsa-miR-584	0.0307	0.22 (0.05 to 0.87)	(125)[G]

Table 2.4 (con't)

MicroRNA	*P*-value	Hazard ratio (95% C.I.)	Relevant literature references
hsa-miR-758	0.0029	CRN=0: 0.77 (0.23 to 2.60)	(132)[O]
		CRN=1: 0.08 (0.03 to 0.21)	
hsa-miR-93	0.0006	2.63 (1.51 to 4.85)	(129)[O]
kshv-miR-k12-1	<0.0001	3.19 (1.93 to 5.29)	(14)[O]
kshv-miR-k12-6-5p	<0.0001	3.70 (1.93 to 7.10)	(165)[O]
hsa-miR-106b	0.0014	RX=0: 0.12 (0.06 to 0.22)	(141)[G]
		RX=1: 0.55 (0.22 to 1.40)	
hsa-miR-143	0.002	Other: 0.30 (0.16 to 0.54)	(133, 134)[O]
		Caucasian: 0.83 (0.61 to 1.12)	

C.I., Confidence Interval; CRN=1 denotes chemotherapy plus radiation and no targeted therapy, CRN=0 denotes non-CRN therapy; RX=1 denotes radiation therapy alone, RX=0 denotes non-RX therapy; CRT=1 denotes chemotherapy plus radiation and targeted therapy, CRT=0 denotes non-CRT therapy; [G] glioblastoma multiforme study; [O] study on any other type of cancer; n/a, no association with any type of cancer found in literature.

Table 2.5 Gene Ontology categories enriched (FDR-adjusted P-value <0.05) among the target genes of microRNAs associated with life expectancy

Gene Ontology	Level	Term	FDR *P*-value	No. of genes
Biological process	3	Neurological process (GO:0050877)	0.0248	219
Biological process	4	Sensory perception (GO:0007600)	0.0111	151
Biological process	5	Sensory perception of chemical stimulus (GO:0007606)	<0.0001	70
Biological process	6	Sensory perception of smell (GO:0007608)	<0.0001	63
Molecular function	4	Transmembrane receptor activity (GO:0004888)	0.0146	239
Molecular function	5	G Protein-coupled receptor activity (GO:0004930)	0.0039	160
Molecular function	6	Rhodopsin-like receptor activity (GO:0001584)	0.0023	134
Molecular function	7	Olfactory receptor activity (GO:0004984)	<0.0001	60

Table 2.6 Gene Ontology categories enriched (FDR-adjusted P-value <0.05) among the target genes of microRNAs associated with overall survival.

Gene Ontology	Level	Term	FDR *P*-value	No. of genes
Biological process	3	Neurological process (GO:0050877)	<0.0001	371
Biological process	3	Cell communication (GO:0007154)	0.0001	1546
Biological process	4	Sensory perception (GO:0007600)	<0.0001	255
Biological process	4	Signal transduction (GO:0007165)	0.0006	1400
Biological process	5	Sensory perception of chem. stimulus (GO:0007606)	<0.0001	129
Biological process	5	Cell surface receptor linked signal transduction (GO:0007166)	0.0145	684
Biological process	6	Sensory perception of smell (GO:0007608)	<0.0001	123
Biological process	6	G Protein-coupled receptor protein signaling pathway (GO:0007186)	0.0156	413
Molecular function	3	Receptor activity (GO:0004872)	0.004	711
Molecular function	3	Antigen binding (GO:0003823)	0.0422	16
Molecular function	4	Transmembrane receptor activity (GO:0004888)	0.0002	457
Molecular function	5	G Protein-coupled receptor activity (GO:0004930)	<0.0001	315
Molecular function	6	Rhodopsin-like receptor activity (GO:0001584)	0.0002	274
Molecular function	7	Olfactory receptor activity (GO:0004984)	<0.0001	120

Table 2.7 Gene Ontology categories enriched (FDR-adjusted P-value <0.01, number genes >6) among the target genes of microRNAs associated with progression-free survival.

Gene Ontology	Level	Term	FDR *P* -value	No. of genes
Biological process	3	Cell communication (GO:0007154)	<0.0001	975
Biological process	3	Multicellular development (GO:0007275)	<0.0001	507
Biological process	3	Neurological process (GO:0050877)	<0.0001	248
Biological process	3	Anatomical structure development (GO:0048856)	<0.0001	483
Biological process	3	Cellular organization & biogenesis (GO:0016043)	0.0008	639
Biological process	3	Cellular metabolic process (GO:0044237)	0.0014	2290
Biological process	3	Cellular developmental process (GO:0048869)	0.0066	551
Biological process	4	Signal transduction (GO:0007165)	<0.0001	889
Biological process	4	Sensory perception (GO:0007600)	<0.0001	172
Biological process	4	System development (GO:0048731)	<0.0001	386
Biological process	5	Sensory perception of chemical stimulus (GO:0007606)	<0.0001	86
Biological process	5	Cell surface receptor linked signal transduction (GO:0007166)	<0.0001	433
Biological process	5	Organ development (GO:0048513)	0.0085	285
Biological process	5	+ Regulation of metabolic process (GO:0009893)	0.0087	84
Biological process	5	Carboxylic acid metabolic process (GO:0019752)	0.0095	185
Biological process	5	+ Regulation of cellular process (GO:0048522)	0.0095	214
Biological process	6	Organ morphogenesis (GO:0009887)	<0.0001	57
Biological process	6	Sensory perception of smell (GO:0007608)	<0.0001	80
Biological process	6	G Protein-coupled receptor protein signaling pathway (GO:0007186)	0.0099	266
Molecular function	3	Protein binding (GO:0005515)	<0.0001	1601
Molecular function	4	Transmembrane receptor activity (GO:0004888)	0.0017	305
Molecular function	6	Rhodopsin-like receptor activity (GO:0001584)	0.0101	179
Molecular function	7	Olfactory receptor activity (GO:0004984)	0.0002	79

Chapter III

Transcription factor-microRNA-target gene networks associated with ovarian cancer survival and recurrence*

ABSTRACT

The identification of reliable transcriptome biomarkers requires the simultaneous consideration of regulatory and target elements including microRNAs (miRNAs), transcription factors (TFs), and target genes. A novel approach that integrates multivariate survival analysis, feature selection, and regulatory network visualization was used to identify reliable biomarkers of ovarian cancer survival and recurrence. Expression profiles of 799 miRNAs, 17,814 TFs and target genes and cohort clinical records on 272 patients diagnosed with ovarian cancer were simultaneously considered and results were validated on an independent group of 146 patients. Three miRNAs (hsa-miR-16, hsa-miR-22*, and ebv-miR-BHRF1-2*) were associated with both ovarian cancer survival and recurrence and 27 miRNAs were associated with either one hazard. Two miRNAs (hsa-miR-521 and hsa-miR-497) were cohort-dependent, while 28 were cohort-independent. This study confirmed 19 miRNAs previously associated with ovarian cancer and identified two miRNAs that have previously been associated with other cancer types. In total, the expression of 838 and 734 target genes and 12 and eight TFs were associated (FDR-adjusted P-value < 0.05) with ovarian cancer survival and recurrence, respectively. Functional analysis highlighted the association between cellular and nucleotide metabolic processes and ovarian cancer. The more direct connections and higher centrality of the miRNAs, TFs and target genes in the survival network studied suggest that network-based approaches to prognosticate or predict ovarian cancer survival may be more effective than those for ovarian cancer recurrence. This study demonstrated the feasibility to infer reliable miRNA-TF-target gene networks

*K.R. Delfino and S.L. Rodriguez-Zas. PLoS One. 2013;8(3):e58608.

associated with survival and recurrence of ovarian cancer based on the simultaneous analysis of co-expression profiles and consideration of the clinical characteristics of the patients.

Introduction

Ovarian cancer, the most malignant gynecologic neoplasm, is the fifth leading cause of cancer deaths among women. Approximately 45% of ovarian cancer patients survive more than five years after initial diagnosis and less than 20% surpass this milestone once the cancer has disseminated (39). Few gene expression profiles have been consistently related to ovarian cancer (166, 167). This may be due to the limited simultaneous consideration of the transcripts and transcript regulators associated with ovarian cancer.

MicroRNAs (miRNAs) are small, non-coding RNA molecules that bind to complementary sequences on target mRNA transcripts, and thus, regulate gene expression at the post-transcription stage. Transcription factors (TFs) are a different type of regulator. These proteins bind to specific DNA sequences in the promoter region, promoting or repressing transcription into mRNA, and thus, regulate genes at a pre-transcription stage (168). Transcription factorss and miRNAs can regulate each other and both can regulate the expression of target genes. Transcription factor-microRNA-target genes can function as onco or tumor suppressor networks, triggering global alterations of genetic programs implicated in cell proliferation, differentiation, apoptosis, and invasiveness in cancer.

Few associations between ovarian cancer and miRNAs or TF have been validated in independent studies (166, 167). Several reasons may be behind the limited understanding of the regulatory networks associated with ovarian cancer. First, most studies associate ovarian cancer to genes (miRNAs or TFs) on an individual basis instead of considering multiple profiles simultaneously. Second, even when studies analyze multiple genome profiles simultaneously, the relationship between target genes and regulatory miRNAs and TFs are not used. Third, most studies do not consider clinical or cohort-dependent factors when characterizing associations

between expression profiles and ovarian cancer. Lastly, most studies consider the binary qualitative trait presence or absence of cancer, and more quantitative measurements such as survival and recurrence are not evaluated.

The main objectives of this study were a) to develop a model to identify and characterize miRNAs, TFs, and target genes associated with ovarian cancer survival, and b) use this information to identify TF-miRNA-target gene networks associated with survival in ovarian cancer. Our overarching hypothesis was that reliable gene expression biomarkers of cancer can be obtained from the consideration of all components in a network simultaneously. A systems biology approach was used to investigate the simultaneous association between multiple miRNAs, TFs, and target genes and cancer survival or recurrence, accounting for non-genetic patient-to-patient sources of variation, and the corresponding networks were analyzed. Results were validated in an independent data set. The study also identified enriched functional categories and pathways of genes associated with cancer survival and recurrence. Understanding the molecular basis of ovarian cancer is key to developing improved prognostic indicators and effective therapies. Given the heterogeneity of this disease, improvements in long-term survival might be achieved by translating recent insights at the molecular and clinical levels into personalized individual treatment strategies.

Materials and Methods

Training data set

Clinical information

Survival, recurrence, cohort, and genomic expression information from 272 patients diagnosed with ovarian cancer was obtained from The Cancer Genome Atlas (http://cancergenome.nih.gov/; accessed September 2009) repository (112). Cohort factors analyzed include treatment received (only chemotherapy, 93%; chemotherapy plus another treatment, 5%; and any treatment other than chemotherapy, 2%); preadjuvant therapy (yes, 8% or no, 92%); additional treatment (only chemotherapy, 41%; chemotherapy plus another treatment, 14%; and any treatment other than chemotherapy, 45%); tumor stage (stage I or II, 4%; stage III, 88%; stage IV, 8%); tumor grade (grade I or II, 4%; any grades other than I or II, 96%); tumor residual disease (no macroscopic disease, 26%; 1-20mm, 61%; greater than 20 mm, 13%); recurrence (yes, 58% or no, 42%), and age at diagnosis (in years). Preadjvant therapy refers to any treatment that the patient received prior to surgery and sample collection. Tumor stage refers to the pathological stage of the tumor in AJCC format (Primary Tumor: T; Stage 1: 1A; 1B; 1C; Stage II: IIA; IIB; IIC; Stage III: IIIA; IIIB; IIIC; Stage IV: IV). Tumor grade is the numeric value used to express the degree of abnormality of cancer cells and is a measure of differentiation and aggressiveness. Tumor residual disease is the measure of the largest remaining nodule. Age refers to the age in years of the individual at the time of diagnosis of ovarian cancer. These cohort factors were accounted for in the analysis because of their known association with survival (169).

Expression profiling

The expression levels of 799 miRNAs were measured using the Agilent 8 × 15K Human microRNA platform (Agilent Technologies, http://www.genomics.agilent.com/). The expression levels of 17,814 TFs and target genes were measured using the Agilent Custom Gene Expression G4502A_07 human gene platform. The transcriptome data is available at (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm). The expression measurements were quantile normalized (probe level), collapsed within the miRNA, TF or gene, and $\log_2$ transformed following the procedures available in the Beehive (http://stagbeetle.animal.uiuc.edu/Beehive) system (77) and previously described in (170-172).

Model and profile selection

Two ovarian cancer response variables were studied: 1) survival time from diagnosis to death (months from diagnosis to death); and 2) recurrence time from diagnosis to recurrence (months from diagnosis to recurrence). Information on comorbidities or cause of death was unavailable, thus the first variable describes the time-dependent likelihood of death, conditional on a prior ovarian cancer diagnostic, irrespective of cause of death or comorbidity. An ovarian cancer predictive model that simultaneously considered all miRNAs and cohort information was used to identify general (or cohort-independent) and personalized (or cohort-dependent) biomarkers. This model overcame limitations of prior studies which ignored the simultaneous association by only analyzing one miRNA at a time or ignoring possible cohort relationships. A biomarker identification pipeline was implemented based on the multivariate Cox survival analysis and complementary feature selection strategies (170, 171, 173). The Cox proportional hazard model assumes a parametric model to test the association between the covariates and the hazard ratio (HR) of the event. After transformation, the hazard (instant probability) of event

(death or recurrence) was modeled with a linear combination of a baseline hazard and explanatory covariates including all the cohort variables, the expression profiles of all genome variables (miRNAs, TFs, or gene targets), and the interaction between them (81). Stepwise and forward selection strategies were used to identify the expression profiles associated with survival or recurrence because of the complementary advantages of these strategies. Profiles remained in the hazard predictive model after consideration of other biomarkers at P-value < 0.1. The significant profiles from the previous stepwise and forward model were included in a model that was subjected to a stepwise selection. This step allowed the identification of broad or general associations between profiles and ovarian cancer hazards that can be used as population prognostic biomarkers. The relaxed P-value threshold allowed detection of profiles that may have weak associations among large sets of profiles and stronger associations as the set was streamlined. In the second step, the interaction between the selected profiles and cohort indicators were evaluated using the stepwise approach. This step allowed the identification of cohort-dependent associations between profiles and the hazard of ovarian cancer death or recurrence that can be used as individualized predictive biomarkers. In the third step, all selected profiles and interactions were combined and further streamlined using the stepwise method. The association between the ovarian cancer hazards and the cohort factors and expression profiles was visualized by plotting the probability of survival predicted from the Cox model against time. The test of no association between the miRNA, TF, gene or cohort prognostic markers and the HR between cohort groups and the 95% confidence interval limits follow a Chi-square distribution. Hazard ratio estimates > 1 (< 1) indicate an increase in the hazard (decrease in the hazard) or decrease in survival probability (increase in survival probability) per unit increase in the level of gene expression. A False Discovery Rate (FDR)–adjusted P < 0.05 and

|HR/expression unit| > 1.15 thresholds were used to identify molecular factors associated with ovarian cancer survival or recurrence. The analysis was implemented using PROC PHREG in SAS (174).

The Cox model assumes proportional hazards across the period studied. This assumption can be expressed as parallel survival functions across the levels of expression profiles or cohort variables in the model. This assumption was tested for the two hazards considered, and there was no indication of significant departure from the assumption. Furthermore, visualization of the survival and residuals did not suggest departure from the model assumptions. There was no indication of significant departure from the proportional hazards assumption, also confirmed by the overlap on miRNAs, TFs, and genes between survival indicators. Biomarkers identified in this study were searched against the ovarian cancer and cancer literature based on independent data sets.

Functional enrichment and miRNA-TF-target gene networks

The known and predicted relationships between miRNAs, TFs, and target genes were obtained from the MIR@NT@N resource (http://maia.uni.lu/mironton.php, (175)). Only the relationship between transcription factors, miRNAs, and target gene supported by a mapping score > 0.85 that correspond to a median P-value < 1×10^{-3} and 90% of the relationships with P-values < 1×10^{-2} were considered. The enrichment of Gene Ontology (GO, http://www.geneontology.org/) (176) molecular functions and biological processes and KEGG (http://www.genome.jp/kegg/) (93, 177) pathways was studied among the target genes associated with ovarian survival and recurrence. Two functional analyses were evaluated. The first functional analysis consisted on Fisher's exact (two-tailed) test implemented in DAVID

(http://david.abcc.ncifcrf.gov/) version 6.7(92) was used to identify the functional categories enriched among all target genes associated (FDR-adjusted P-value < 0.05) with survival or recurrence (170, 171). Categories that had at least 5 genes and were significant at FDR-adjusted P-value < 0.1 were considered enriched. This analysis offered a baseline understanding of the categories associated with ovarian cancer.

The second functional analysis consisted on a set enrichment analysis (87) of all target genes regardless of the significance level of the association with ovarian cancer survival or recurrence. This analysis considered the association between survival or recurrence and gene expression through the sorting of the target genes by the magnitude, sign, and standard error of the estimate in the underlying scale or $\log_e$(HR). Positive estimates correspond to HRs > 1 and thus lower survival or higher risk of recurrence. Conversely, negative estimates correspond to HRs < 1 and thus higher survival or lower risk of recurrence. The set enrichment analysis implemented in Babelomics (http://babelomics.bioinfo.cipf.es/) version 4.3 (87) was used to apply a segmentation test that identifies for asymmetrical distributions of functional categories between the genes ranked from negative to positive $\log_e$ (HR) estimates for ovarian cancer survival or recurrence. Categories significant at FDR-adjusted P-value < 0.05 and having at least 75 genes were considered enriched. The less stringent threshold used for the Fisher's enrichment analysis relative to the set enrichment analysis was motivated by the higher number of target genes analyzed in the second analysis relative to the first analysis. The genes associated with ovarian cancer hazard were also searched against the Dragon database (http://apps.sanbi.ac.za/ddoc/) of ovarian cancer genes (178). The networks of TFs, miRNAs, and target genes significantly associated with ovarian cancer survival or recurrence (P-value < 0.01) were depicted using Cytoscape (http://www.cytoscape.org/) software (95), an open source

software platform for visualizing networks and including attributes. The distribution and connectivity of the TFs, miRNAs, and target genes within sub-networks and the overall network were characterized

Validation data set

The associations between expression profiles and ovarian cancer survival or recurrence identified based on P-values and characterized based on HR estimates were validated on an independent data set of 146 patients obtained from the TCGA repository. Two indicators of the reliability of the predictive profiles in the independent validation were considered. First, mean square error (MSE) was used as measure of the lack of adequacy of the cohort-independent and -dependent expression profiles to accurately predict the time to death or recurrence in the training and validation data sets. Second, additional validation of the detected profile association was gained from the study of the correlation of the estimates ($\log_e(HR)$) corresponding to each profile between training and validation data sets. The Pearson and Spearman correlations of the profile associations with death and recurrence between the training and validation data sets were computed.

Results and Discussion

Table 3.1 summarizes the number and distribution of individuals studied across levels of the cohort covariates considered in the training and validation data sets. The median age at diagnosis was 60.2 years and 59.6 years for the training and validation data sets, respectively. These were consistent with the National Cancer Institute reports that the median age at diagnosis for cancer of the ovary (from 2004-2008) was 63 years of age and the median age at death was 71 years of age (38). The range of age at diagnosis was 57 years (ages from 27 to 84 years) and 52 years (ages from 37 to 89 years) for the training and validation data sets, respectively. The median time for survival and recurrence for the training set was 2.4 years and 47.4 months and for the validation set were 3.3 years and 58. 7 months, respectively. The Pearson and Spearman correlation coefficients between both events (age at death and at recurrence) were 0.72 and 0.77 (P-value < 0.0001), respectively in the training data set and 0.69 and 0.68 (P-value < 0.0001), respectively in the validation data set. These statistics were in agreement with previously documented survival rates of ovarian cancer: 1 year: 77.5%, 2 year: 64%, 3 year: 54.4%, 5 year: 43.9%, 8 year: 37.8%, 10 year: 36.4% (38). Median survival for patients was 25.7 months for early treatment patients and 27.1 months for those patients in a delayed treatment group (38). The distribution of observations per cohort variable level in the training and validation sets was consistent (**Table 3.1**). The representation of treatment, preadjuvant therapy, additional treatment, tumor stage, tumor grade, tumor residual disease, and recurrence was comparable between data sets. None of the 15 sample source sites dominated the representation in either training or validation set.

The correlations between the observed and predicted time-to-death and time-to-recurrence were approximately 0.60. Higher correlations (0.8 on average) were observed when

only the lower times-to-event were considered because more observations were available and more precise predictions could be obtained. Prediction of longer time-to-event intervals were associated with higher uncertainty due to fewer observations within cohort variable levels, and thus lower correlations between training and validation data sets. The moderate correlation between the two time-to-event analyses suggests the differences in the magnitude and direction of genomic and environmental effects on ovarian cancer survival and recurrence.

miRNA biomarkers of ovarian cancer survival and recurrence

Tables 3.2 and **3.3** list the 16 and 14 miRNAs simultaneously associated with ovarian cancer survival and recurrence detected by the three-step feature selection approach and supporting literature references. The vast majority of the miRNAs associated with survival detected in this study have been reported by other studies. This level of validation reaffirms the validity of the approach undertaken and of the results presented. Of the 16 miRNAs associated with survival, 12 miRNAs have been previously associated with ovarian cancer (hsa-miR-144, hsa-miR-16, hsa-miR-182*, hsa-miR-521, hsa-miR-18b*, hsa-miR-19a*, hsa-miR-22*, hsa-miR-381, hsa-miR-485-3p, hsa-miR-509-3-5p, hsa-miR-148a, and hsa-miR-106b) and one miRNA has been associated with cervical cancer (hsa-miR-329; (179). The literature review supporting these results was summarized in **Table 3.2**.

Of the miRNAs previously associated with ovarian cancer, the trends of all 12 miRNA were consistent with those reported in previous studies. Hsa-miR-144 (HR=1.30), hsa-miR-16 (HR=2.07), hsa-miR-182* (HR=2.35), hsa-miR-18b* (HR=1.95), hsa-miR-19a* (HR=1.75), and hsa-miR-106b (HR=1.57) were over-expressed in all 3 ovarian tumor histologic subtypes relative to normal primary human ovarian surface epithelium cultures (162). The consistency between

the detected and previously reported trends further supports the biomarker detection strategy presented.

Hsa-miR-182 was also up-regulated in ovarian carcinoma in Stage III/IV epithelial ovarian carcinoma versus normal tissue (13) and has been associated with higher death hazard in glioblastoma multiforme patients receiving chemotherapy plus radiation and targeted treatment (170). The region containing hsa-miR-182 was amplified in 28.9% of the epithelial ovarian cancer, implying an oncogene-type function, and possibly targets genes forkhead box O1, forkhead box O3 (*FOXO1*;*FOXO3*) which are involved in promoting differentiation and grown inhibition (tumor suppressors,(13)). Hsa-miR-18b* and hsa-miR-16 were found to robustly distinguish ovarian cancer tumors from normal tissue and were significantly up-regulated in ovarian cancer (180). Hsa-miR-16 (HR=2.07) has been shown to be up-regulated in serous ovarian carcinoma versus normal ovarian tissues, as well as up-regulated in stage III/IV ovarian cancer versus normal ovarian tissue (13, 41). Hsa-miR-22 (HR=0.25) was under-expressed in 3 ovarian tumor histologic subtypes relative to normal primary human ovarian surface epithelium cultures (162). Hsa-miR-22 was also down -regulated in Stage III/IV epithelial ovarian carcinoma versus normal and up-regulated in primary versus recurrent serous papillary ovarian carcinomas (13). Hsa-miR-148a (HR=0.78) was down-regulated in ovarian cancer cell lines and may be involved in the carcinogenesis of ovarian cancer through deregulation of cell proliferation (181). Hsa-miR-509-3-5p (HR=0.69) was over-expressed in stage I ovarian cancer relative to stage III ovarian cancer with a p-value=0.017 and fold-change=4.01(182). Both, hsa-miR-521 and hsa-miR-381 were over-expressed in platinum resistant versus platinum sensitive ovarian cancer (183, 184).

The evaluation of clinical factor dependent associations between miRNAs and ovarian cancer survival offer insights into general and condition-specific biomarkers. Of the 16 miRNAs associated with ovarian cancer survival, 15 exhibited general (clinically independent) associations with survival, meanwhile hsa-miR-521 had a tumor grade-dependent association with survival. The hazard of ovarian cancer death increased 2.10 per unit increase in hsa-miR-521 level in patients that have grade I or II tumors and decreased 0.55 per unit increase in the miRNA in patients that have higher level tumors. The survival plot in **Figure 3.1** depicts the association between the probability of ovarian cancer survival and the interaction between miRNA expression and tumor grade. Lower expression of hsa-miR-521 was associated with the lowest and highest probability of survival in the presence of high (Rest) and low (I and II) grade tumors, respectively.

Similarly to the findings for survival, the majority of the miRNAs associated with recurrence have been previously associated with ovarian cancer, thus reaffirming the reliability of the feature selection approach implemented. Among the 14 miRNAs associated with recurrence in ovarian cancer (**Table 3.3**), 9 have been previously linked to ovarian cancer (hsa-miR-146a, hsa-miR-15b, hsa-miR-16, hsa-miR-206, hsa-miR-214*, hsa-miR-22*, hsa-miR-223, hsa-miR-497, and hsa-miR-96), and one had a previous association with paired lung primary tumors (hsa-miR-369-3p). **Table 3.3** summarizes the literature review supporting the detected associations.

The trends of the 9 miRNA previously linked to ovarian cancer and also found in this study were consistent with previously reported. Hsa-miR-146a (HR=0.62) was under-expressed in ovarian tumor histologic subtypes relative to normal primary human ovarian surface epithelium cultures (162). Hsa-miR-206 (HR=0.59) was down-regulated in ovarian cancer cell

lines versus normal (185). Hsa-miR-22 (HR=0.24) was over-expressed in recurrent ovarian cancer versus primary ovarian cancer (186). This miRNA also was down-regulated in ovarian carcinoma in early stage versus late stage; down-regulated in Stage III/IV epithelial ovarian carcinoma versus normal; and up-regulated in primary versus recurrent serous papillary ovarian carcinomas (13). Hsa-miR-497 (in this study HR Chemo=0.84; Chemo_Other=0.53; Other=0.17) was down-regulated in ovarian cancer cell line versus normal ovarian cell lines (187, 188). Hsa-miR-16 (HR=2.76) up-regulated in serous ovarian carcinoma versus normal ovarian tissues, as well as up-regulated in stage III/IV ovarian cancer versus normal ovarian tissue as well (13, 41, 180) . Hsa-miR-214 (HR=2.03) was over-expressed ovarian tumor histologic subtypes relative to normal primary human ovarian surface epithelium cultures (162) . In a study of epithelial ovarian cancer, hsa-miR-214 was differentially expressed in those with recurrence compared with those without recurrence in both a training and validation set. Tumor tissue samples from those with recurrence were up-regulated compared with those without recurrence in epithelial ovarian cancer (189). Hsa-miR-214 expression was associated with high grade and late stage tumors, was up-regulated in ovarian cancer tumor tissues, and has a potential role in recurrence (13). Hsa-miR-214 was also found to play a role in ovarian cancer by targeting PTEN (166).

Hsa-miR-223 (HR=1.69) was over-expressed in all 3 ovarian tumor histologic subtypes relative to normal primary human ovarian surface epithelium cultures (162). Hsa-miR-223 was over-expressed in recurrent ovarian cancer versus primary ovarian cancer (186). Hsa-miR-223 was up-regulated in tumor tissue sample from those with recurrence compared with those without recurrence in epithelial ovarian cancer (189). Hsa-miR-96 (HR=1.22) was over-expressed in ovarian cancer cell lines versus normal ovarian cell lines (162, 180, 187, 190). Hsa-miR-369-3p

(HR=1.52), associated with ovarian cancer recurrence in this study, was similarly up-regulated in paired lung primary tumors (191).

The study of interactions between miRNA expression and cohort factors supported the identification of individualized biomarkers. General associations between miRNAs and recurrence irrespective of cohort factors were identified for 13 miRNA. A treatment-dependent association between risk or hazard of recurrence and hsa-miR-497 was identified. The hazard for ovarian cancer recurrence decreased with increasing miRNA level in patients for all three treatments (Chemo, Chemo_Other, Other), and the hazard was lowest (0.17) for individuals receiving the Other treatment. **Figure 3.2** depicts the association between the probability of non-recurrence interaction between level of hsa-miR-497 and treatment. The probability of non-recurrence was distinctively lower in patients with low miRNA levels receiving Chem treatment, however was not different between patients receiving Chemo or Chemo and Other treatments when the levels of miRNA were high.

Transcription factors and target genes associated with survival and recurrence

In total, the expression of 838 and 734 target genes and 12 and eight TFs were associated (FDR-adjusted P-value < 0.05) with ovarian cancer survival and recurrence, respectively. The TFs associated with ovarian cancer survival and recurrence and supporting literature review are listed in **Tables 3.4** and **3.5**, respectively.

The four TFs significantly associated with both ovarian cancer survival and recurrence (early growth response 1 (*EGR1*), early growth response (*EGR2*), FBJ murine osteosarcoma viral oncogene homolog (*FOS*), and transforming growth factor beta 1(*TGFB1*), exhibited trends consistent with previous studies. *EGR1* (HR=1.15 for survival and recurrence) has a key role in

carcinogenesis and cancer recurrence, and exhibits increased expression in gastric cancer tissues relative to normal mucosa (192). The positive association between *EGR2* and hazard uncovered in this study (HR=1.17 for survival and recurrence) was confirmed with reports that this TF plays a key role in the PTEN-induced apoptotic pathway. Furthermore, studies suggest that this TF may be a promising target molecule for gene therapy to treat a variety of cancers (193). *FOS* expression (HR=1.15; 1.13 for death and recurrence in this study, respectively) has been associated with ovarian cancer, and is a molecular predictor of recurrence and survival in epithelial ovarian carcinomas (194). *TGFB1* (HR=0.46; 0.56 for death and recurrence in this study, respectively) has been linked to ovarian cancer (195-197) , and may play an important role in ovarian cancer biology with potential effects on tumor growth and angiogenesis (198).

Transcription factors associated with survival

Eight TFs were solely associated with the hazard of ovarian cancer death: circadian locomotor output cycles kaput (*CLOCK*), estrogen receptor 2 (*ESR2*), v-ets erythroblastosis virus E26 oncogene homolog 2 (*ETS2*), histone deacetylase 3 (*HDAC3*), homeobox A1 (*HOXA1*), v-myc myelocytomatosis viral oncogene homolog (*MYC*), nuclear receptor subfamily 5, group A, Member 1 (*NR5A1*), and POU class 2 homeobox 2 (*POU2F2*), and their trends were in agreement with previous studies. *MYC* (HR=1.27) contributes independently to ovarian and breast pathogenesis when over-expressed (199), and was more frequently detected in malignant ovarian tumors when compared with benign ovarian tumors (104, 200). *ESR2* (HR=0.66) was significantly lower in ovarian cancer cell lines and tissues than in their corresponding normal counterparts (201). *ESR2* has been associated with malignant ovarian epithelial cells (202) and may be a susceptibility marker for epithelial ovarian cancer (203).

The opposite association between *POU2F2* expression and ovarian cancer hazard (HR=0.64) detected in this study was consistent with reports that this TF, a member of the POU homeodomain family of transcriptional regulators critical for normal embryonic development, was associated with down-regulation of B-cell CLL/lymphoma 2 (*BCL-2*) that results in apoptosis (204) . Over-expression of *ETS2* has previously been shown in human esophageal squamous cell carcinoma and breast cancer (202, 205) . This TF also plays a role in regulation of the production of TF *MYC*, also significantly associated with increased hazard (HR=1.27) in ovarian cancer in this study (206). These findings were in agreement with the positive association between *ETS2* and ovarian cancer death hazard detected in this study (HR=1.32). Defects in *NR5A1* (HR=0.53 in this study) can result in arrest of ovarian function (207). The relationship between *CLOCK* and ovarian cancer survival detected in the present study (HR=0.81) agrees with the report that variations in the epigenetics of *CLOCK* may lead to increased risk of breast cancer (208), and that in women with breast cancer, there was significantly less methylation of the *CLOCK* promoter region (209). Similar to this study, *HDAC* (HR=1.63) was over-expressed in 80% of cases of ovarian cancer, with no significant difference in the expression profiles between histological subtypes (210) . Suppression of *HOXA1* (HR=0.71) has been linked to an increase of invasive cancer cells in human pancreatic cancer (211).

Transcription factors associated with recurrence

Four TFs were only associated with the hazard of ovarian cancer recurrence, and their trends were all consistent with previous work: CCCTC-binding factor (*CTCF*), Myogenic Differentiation 1(*MYOD1*), SRY (sex determining region Y)-Box 18 (*SOX18*), and TATA Box Binding Protein (*TBP*). *CTCF* (HR=1.71) plays a role in breast cancer (212, 213), and *MYOD1*

*(*HR=0.77) has previously been associated with cervical cancer (214, 215). *SOX18* (HR=0.77), a member of the SOX family of transcription factors involved in the determination of the cell fate, has been proposed as a useful target for human cancer treatment (216, 217). Consistent with our findings on ovarian cancer, *TBP* (HR=1.63), which is highly expressed in the ovary, has elevated expression in human colon carcinomas (218).

Target gene biomarkers of ovarian cancer survival and recurrence

Among the target genes associated with ovarian cancer survival, 16 were identified in the Dragon database of ovarian cancer genes: acetylcholinesterase (*ACHE*); BCL2-antagonist/killer 1 (*BAK1*); B-cell CLL/lymphoma 2 (*BCL2*); CD44 molecule (Indian blood group) (*CD44*); CD63 molecule (*CD63*); cadherin 13, H-cadherin (heart) (*CDH13*); cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4) (*CDKN2B*); colony stimulating factor 1 receptor (*CSF1R*); cathepsin D (*CTSD*); discoidin domain receptor tyrosine kinase 1 (*DDR1*); galactose-1-phosphate uridylyltransferase (*GALT*); kallikrein-related peptidase 9 (*KLK9*); mitogen-activated protein kinase kinase 1 (*MAP2K1*); mitogen-activated protein kinase kinase 3 (*MAP3K3*); *MYC*, and platelet-derived growth factor receptor, alpha polypeptide (*PDGFRA*). Likewise, among the genes associated with ovarian cancer recurrence, 9 were identified in the Dragon database: *ACHE*; *BAK1*; breast cancer 1, early onset (*BRCA1***)**; *CD44*; *CTSD*; *DDR1*; *KLK9*; antigen identified by monoclonal antibody Ki-67 (*MKI67*), and topoisomerase (DNA) II alpha 170kDa (*TOP2A*).

Validation

Two indicators of the reliability of the predictive profiles in the independent validation were considered. The relative increment in MSE of the model including the cohort-independent and dependent-expression profiles between the training and validation data set were 13.4% and

15.4% for survival and recurrence, respectively. As expected, the predictive equation offered a better description of the data used to develop the equation (i.e. the training data set), and a small difference between training and validating data was expected due to sampling effects such as between-patient variation. The small increase in MSE between the training and validating data set was a first, global indicator of the similar profile-hazard relationship identified in both independent data sets and of the replicability of our findings. Second, the Pearson (and Spearman) correlations of the profile associations with death and recurrence between the training (e.g. **Tables 3.2** to **3.5**) and validation data sets were 89.7% (84.5%) and 87.3% (82.4%), respectively. The cross-validation results and the agreement between the literature review and the present findings further suggest that the detected profiles associated with ovarian cancer are likely to be replicable. Experimental confirmation of the findings is needed.

Functional gene groups associated with ovarian cancer survival and recurrence

The functional analyses of the target genes associated with ovarian cancer uncovered enriched pathways and processes, many of which were previously associated with ovarian cancer. Analysis of the significant target genes associated with ovarian cancer survival using a Fisher exact test uncovered enrichment of biological processes including ribonucleotide biosynthetic process (P-value < 0.002, 14 genes) and immune response (P-value < 0.0004, 49 genes) and the KEGG pathways lysosome (P-value < 0.001, genes15) and epithelial cell signaling (P-value < 0.001, genes 11). Likewise, analysis of the significant target genes associated with ovarian cancer recurrence using a Fisher exact test uncovered enrichment of biological processes including the NAD metabolic process (P-value < 0.001, 6 genes), M phase

(P-value < 0.002, 24 genes), and pyrimidine-and nicotineamide-nuclotide metabolic processing (P-value < 0.02, 6 genes).

The set enrichment analysis of all target genes segmented by their positive or negative association with survival or recurrence offered additional insights into the functional categories differentially represented among gene groups. **Table 3.6** lists the GO biological processes differentially (FDR-adjusted P-value < 0.05, > 75 genes) represented between the genes that have negative or positive associations with death and recurrence hazard. Two GO biological processes had significant differential enrichment between the genes segmented by low and high hazard of ovarian cancer death. Likewise, 12 GO biological processes had significant differential enrichment between the genes associated with low and high hazard of ovarian cancer recurrence. **Table 3.6** includes the corresponding characterization of the differential enrichment ($\log_e$(odds ratio)), and the statistical significance level. A $\log_e$(odds ratio) > 0 (< 0) indicates that the category was more (less) enriched among the genes with lower hazard relative to the genes with higher hazard of death or recurrence. Among the significant categories, all were characterized by $\log_e$(odds ratio) > 0, indicating that there were more genes pertaining to the category in the low hazard group relative to the high hazard group. **Supplementary Figures 3.1** and **3.2** depict the relation between the GO biological processes associated with the hazards of ovarian cancer death and recurrence inferred from the set enrichment analysis, respectively. Processes associated with general metabolism were differentially enriched among the genes associated with ovarian cancer death. Processes associated with nucleotide metabolism and transcription were enriched among the genes associated with ovarian cancer survival recurrence. The processes identified by the set enrichment analyses were consistent with the results from the significant gene list enrichment analyses. The results from our functional analyses of target genes associated with ovarian cancer

survival and recurrence were in agreement with previous studies. Transcriptome analysis has shown that suppression of NOTCH signaling in ovarian cancer cells led to down-regulation of genes in pathways involved in cell-cycle regulation and nucleotide metabolism (219). The inhibition of cell proliferation in an ovarian cancer cell line in response to a differentiation-inducing agent was related to a shift in the direction of the purine metabolism from anabolism to catabolism (220). Inhibition of cell metabolism has been proposed as an effective treatment against human epithelial ovarian carcinomas (221).

microRNA-transcription factor-target gene networks of survival and recurrence

In gene regulatory networks, TFs and miRNAs regulate each other and the expression of target genes (57). The binding sites of TFs and genes can be the target of miRNAs and other TFs. Transcription factors regulate genes at the DNA level, while miRNAs regulate gene expression post-transcriptionally (175). Applying previously advocated approaches, this study combined TF and miRNA target prediction together with context-linked (cohort) information and experimental genome-wide co-expression data to identify biologically meaningful molecular interactions (100, 175). The networks of TFs, miRNAs, and target genes significantly associated with survival or recurrence (P-value < 0.01), were reconstructed using Cytoscape. The reconstructed molecular networks that integrate the pattern of association between TFs, miRNAs, target genes and survival or recurrence aid in the identification of robust network biomarkers of ovarian cancer.

Figures 3.3 and **3.4** depict a global network of the miRNAs, TFs and target genes (irrespective of significance) for ovarian cancer survival and recurrence, respectively. These general networks include six and four TFs, 15 and 13 miRNAs and 167 and 89 target genes

associated with survival and recurrence, respectively. **Supplementary Figures 3.3** and **3.4** depict local sub-networks of significant miRNAs, TFs, and targets all significantly associated with ovarian cancer survival and recurrence, respectively. These targeted networks include six and three TFs, 14 and 13 miRNAs and 71 and 56 target genes associated with survival and recurrence, respectively.

The difference in topology between global networks offers insights into the most effective therapies to ameliorate both phenotypes. The networks of survival and recurrence differ in interconnectivity and relation between driver and passenger genomic units. The survival network includes a larger number of driver TFs and miRNAs and a much larger number of target genes that translates into a more driver-centric connectivity than in the recurrence network. Up-regulated miRNAs (red nodes indicate higher hazard of death) appear to dominate as hubs in the survival network, meanwhile a similar number of up and down-regulated miRNAs were hubs in the recurrence network. There were more up than down-regulated TFs in the survival network and no down-regulated TFs in the recurrence network.

For the general and targeted survival networks, four edges was the most frequent shortest path length and was near double the number of paths of length two or three. For the general recurrence networks, four edges was the most frequent shortest path length and was near triple the number of paths of length two or three, meanwhile the distribution of path length was fairly uniform from two to six edges. This pathway comparison indicates that the connections between miRNAs, TFs and target genes were more direct for survival than for recurrence. This result was consistent with the distribution of shared neighbors and average neighborhood connectivity. This distribution was dominated by one shared neighbor in both survival and recurrence networks. However, two and three shared neighbors were more common in the recurrence network. The

median average neighborhood connectivity was 15 and seven for the general survival and recurrence networks, respectively. Betweeness and closeness centrality measurements confirmed these trends. Also the centralization of the survival network was double that of the recurrence network meanwhile the density of the networks follows an approximately inverse relationship. The more direct connections and higher centrality of the survival network suggest that network-based approaches to prognosticate or predict ovarian cancer survival may be more effective than those for ovarian cancer recurrence.

Conclusion

This study demonstrated the feasibility to infer reliable miRNA-TF-target gene networks associated with survival and recurrence of ovarian cancer based on the simultaneous analysis of co-expression profiles and consideration of the clinical characteristics of the patients. The expression of three miRNAs (hsa-miR-16, hsa-miR-22*, and ebv-miR-BHRF1-2*), four TFs (*FOS, EGR2, EGR1,* and *TGFB1*) and 308 genes were associated with the hazard of ovarian cancer survival and recurrence. Both hsa-miR-16 and hsa-miR-22* were previously linked to ovarian cancer and exhibited trends in this study similar to those in independent studies. The expression of TFs *FOS*, *EGR1*, and *EGR2* was positively associated with ovarian cancer hazard, meanwhile the expression of *TGFB1* was negatively associated with the hazard. These overlapping results suggest the importance of these biomarkers in the recurrence of ovarian cancer and are a strong lead for further experimental validation. This study confirmed 19 miRNAs previously associated with ovarian cancer and identified two miRNAs that have previously been associated with other cancer types. Three miRNAs were associated with both ovarian cancer survival and recurrence and 27 miRNAs were associated with only one hazard. Two miRNAs (hsa-miR-521 and hsa-miR-497) were cohort-dependent, while 28 were cohort-independent. Empirical confirmation of these general and cohort-dependent findings could lead to improved prognostic and predictive tools. In total, the expression of 838 and 734 target genes and 12 and eight TFs were associated (FDR-adjusted P-value < 0.05) with ovarian cancer survival and recurrence, respectively. Functional analysis highlighted the association between cellular and nucleotide metabolic processes and ovarian cancer. The more direct connections and higher centrality of the miRNAs, TFs and target genes in the survival network suggest that network-based approaches to prognosticate or predict ovarian cancer survival may be more

effective than those for ovarian cancer recurrence. The understanding the biology and molecular pathogenesis of ovarian cancer is key to developing improved prognostic indicators and effective therapies.

Figures

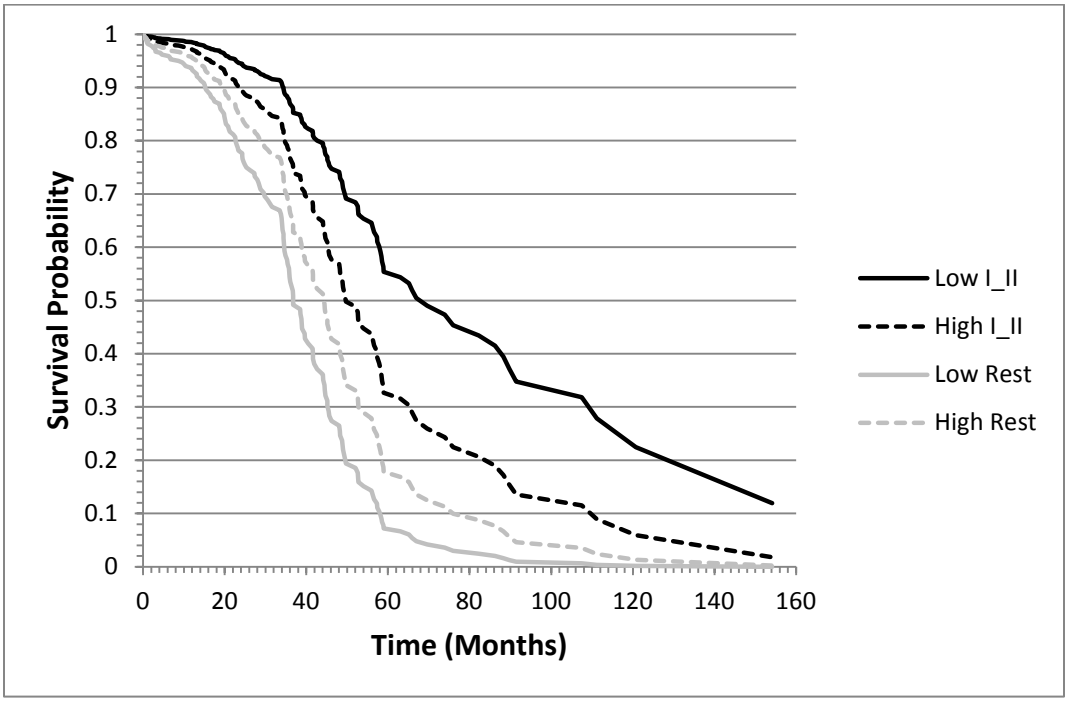


Figure 3.1 Probability of ovarian cancer survival for patients that have lower grade (I and II) tumors (black lines) or higher (Rest) grade tumors (gray lines) and high (dashed lines) or low (solid line) levels of hsa-miR-521.

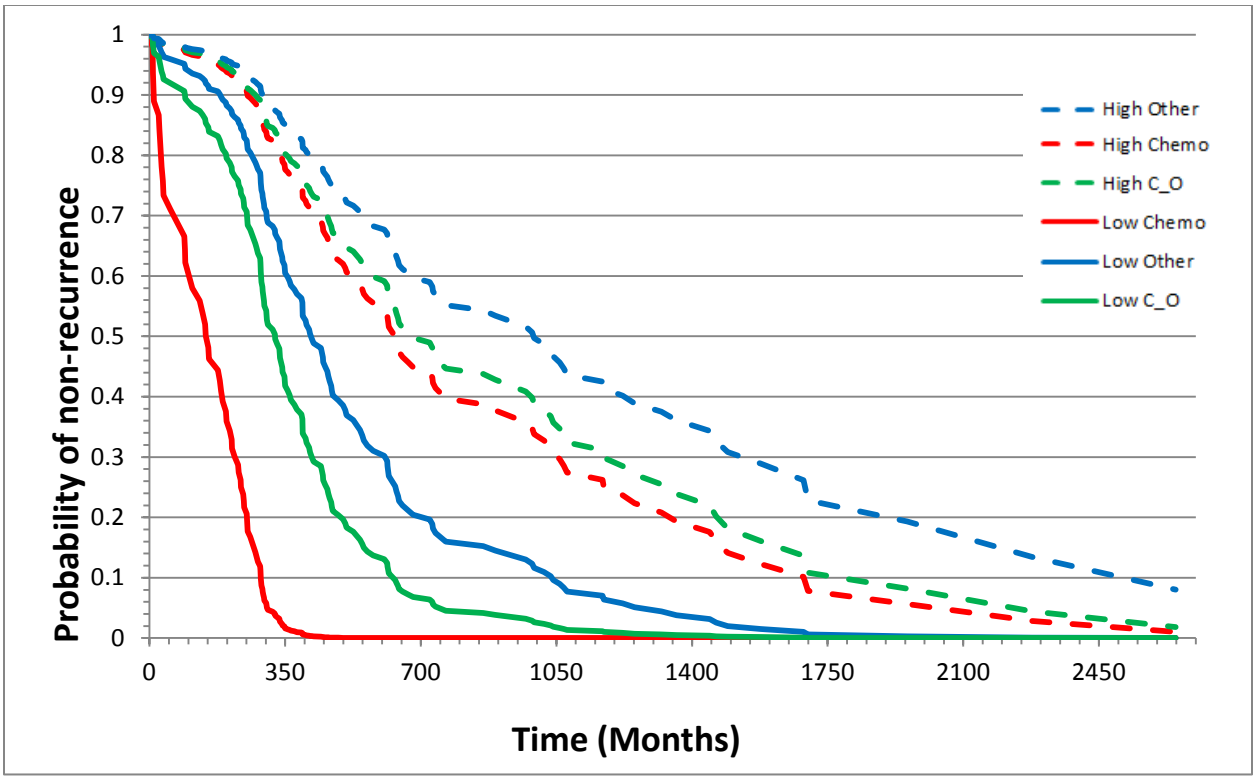


Figure 3.2 Probability of ovarian cancer non-recurrence for patients receiving the treatment chemotherapy only (Chemo; Red), chemotherapy along with another treatment (C_O; Blue) or some other treatment or combination of treatments except chemotherapy (Other; Green) that have high (dashed lines) or low (solid line) levels of hsa-miR-497.

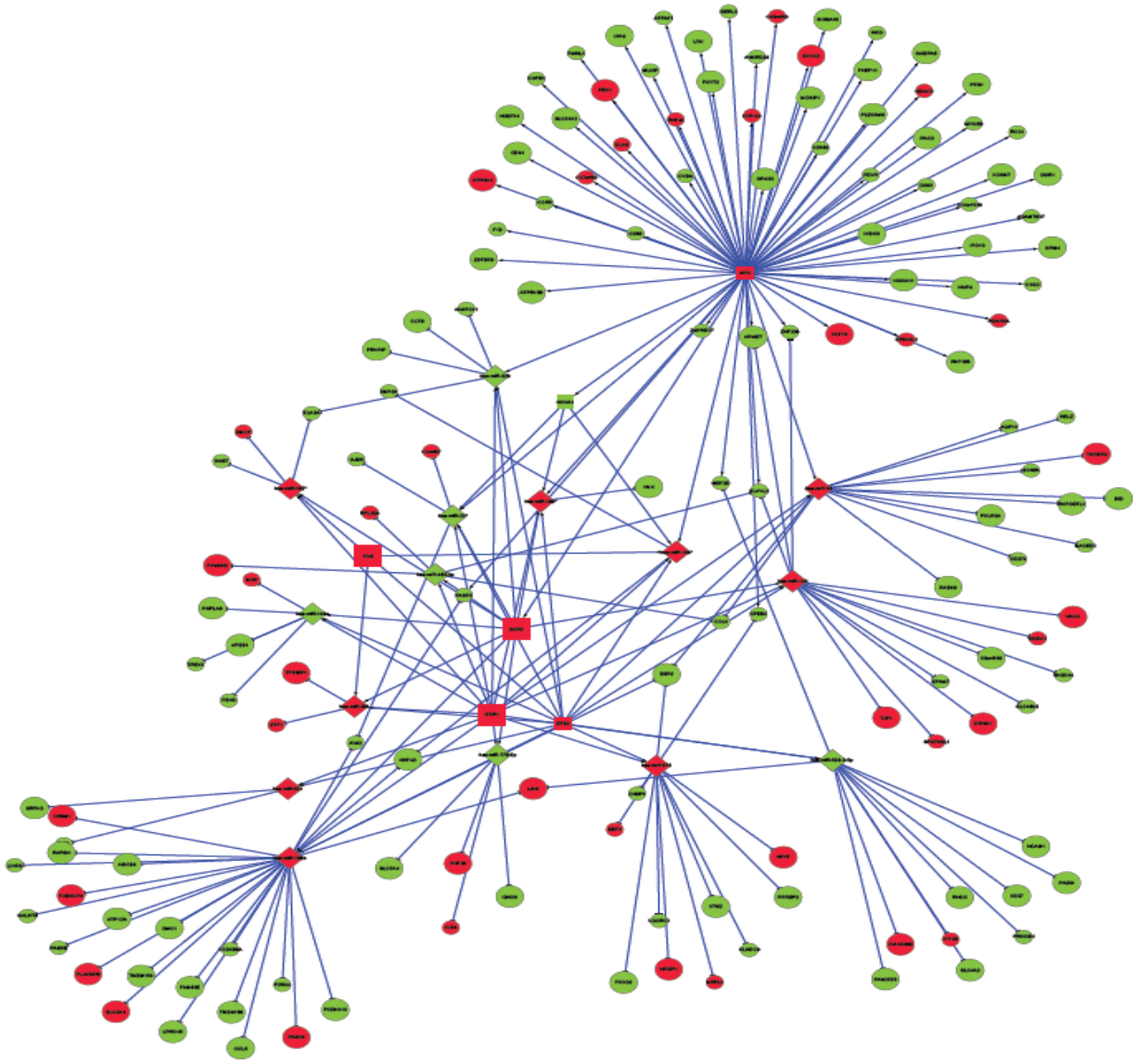


Figure 3.3 Network of microRNAs, transcription factors, and target genes associated with survival in ovarian cancer.
Node Shape: microRNA=diamond, target gene=circle, transcription factor=square; Node Color: Red indicates increased hazard with high expression, Green indicates decreased hazard with high expression; Node Size: larger indicates a more extreme association (P-value < 0.006), smaller indicates a less extreme association

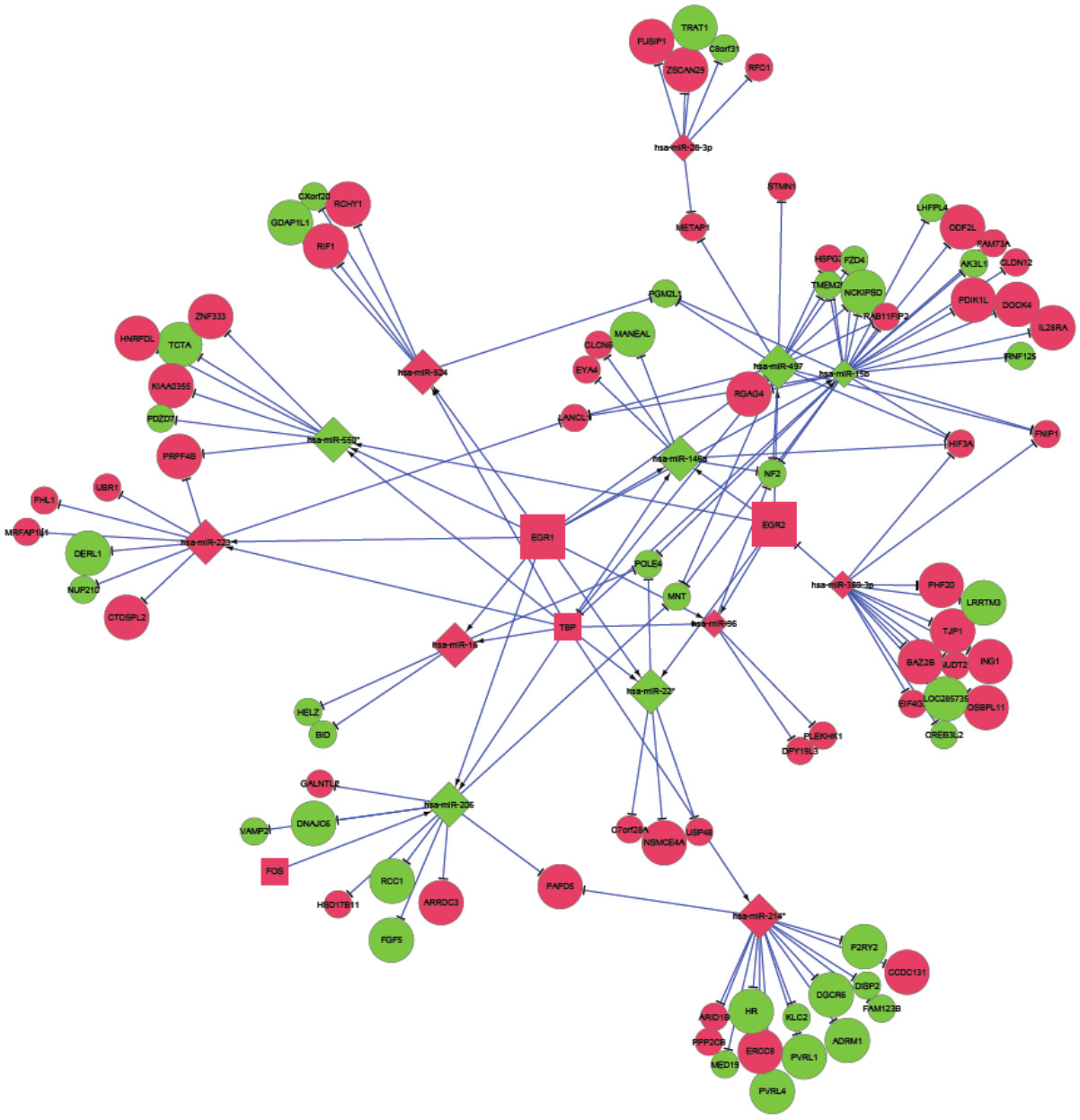


Figure 3.4 Network of microRNA, transcription factors, and target genes associated with ovarian cancer recurrence.
Node Shape: microRNA=diamond, target gene=circle, transcription factor=square; Node Color: Red indicates increased hazard with high expression, Green indicates decreased hazard with high expression; Node Size: larger indicates a more extreme association (P-value < 0.006), smaller indicates a less extreme association

Tables

Table 3.1 Number and distribution of individuals analyzed for post-diagnostic survival and post diagnostic recurrence and levels of the cohort factors considered.

		Survival				Recurrence			
		Training Set		Validation Set		Training Set		Validation Set	
		Number	Percent	Number	Percent	Number	Percent	Number	Percent
Total		272		146		157		92	
N[1] Censored		107	39%	75	51%	31	20%	38	41%
Treatment[2]	*Chemo*[3]	253	93%	101	69%	150	96%	67	72%
	Chemo_Other[4]	14	5%	25	17%	5	3%	18	20%
	Other[5]	5	2%	20	14%	2	1%	7	8%
Preadjuvant Therapy[6]	*Yes*	21	8%	36	25%	7	4%	25	27%
	No	251	92%	110	75%	150	96%	67	73%
Additional Treatment[7]	*Chemo*[3]	113	41%	49	34%	106	67%	45	49%
	Chemo_Other[4]	37	14%	35	24%	34	22%	35	38%
	Other[5]	122	45%	62	42%	17	11%	12	13%
Tumor Stage[8]	*I_II*[9]	12	4%	17	12%	5	3%	9	10%
	III[10]	239	88%	89	61%	143	91%	59	64%
	IV[11]	21	8%	40	27%	9	6%	24	26%
Tumor Grade[12]	*I or II*[13]	12	4%	45	31%	6	4%	35	38%
	Rest[14]	260	96%	101	69%	151	96%	57	62%
Tumor Residual Disease[15]	*0*[16]	71	26%	52	36%	34	22%	30	33%
	1_20[17]	167	61%	60	41%	102	65%	40	43%
	>20[18]	34	13%	34	23%	21	13%	22	24%
Recurrence[19]	*Yes*	157	58%	92	63%	157	100%	92	100%
	No	115	42%	54	37%	0	0%	0	0%

[1]N: Number of patients;[2]Treatment: Type of treatment received;[3]Chemo: Only chemotherapy;[4]Chemo_Other: Chemotherapy plus another treatment;[5]Other: Any treatment other than chemotherapy;[6]Preadjuvant Therapy: Any treatment that the patient received prior to surgery and sample collection;[7]Additional Treatment: Treatment given after initial first round treatment;[8]Tumor Stage: pathological stage of the tumor in AJCC format (Primary Tumor: T; Stage I: IA; IB; IC; Stage II: IIA; IIB; IIC; Stage III: IIIA; IIIB; IIIC; Stage IV: IV);[9]I_II: Stage I or II ovarian cancer;[10]III: Stage III ovarian cancer;[11]IV: Stage IV ovarian cancer;[12]Tumor Grade: Numeric value used to express the degree of abnormality of cancer cells;[13]I or II: Grade I or II tumor;[14]Rest: Any tumor grades other than I or II;[15]Tumor Residual Disease: Measure of the largest remaining nodule;[16]0: No macroscopic disease;[17]1_20: 1-20mm;[18]>20: Greater than 20 mm;[19]Recurrence: Return of cancer

Table 3.2 MicroRNAs associated with post-diagnostic survival and supporting independent studies.

MicroRNA	*P*-Value	Estimate	Hazard Ratio (95% C.I.[1])	Relevant Literature References
hsa-miR-22*	<.0001	-1.4007	0.25 (0.14 to 0.44)	(13, 162, 183, 186, 189)[O]
hsa-miR-770-5p	<.0001	-1.2946	0.27 (0.16 to 0.47)	NA
hsa-miR-485-3p	<.0001	-0.8158	0.44 (0.30 to 0.66)	(181)[O]
hsa-miR-16	<.0001	0.7249	2.07 (1.53 to 2.79)	(13, 41, 162)[O]
hsa-miR-144	<.0001	0.2644	1.3 (1.14 to 1.49)	(162)[O]
ebv-miR-BHRF1-2*	0.0001	1.4787	4.39 (2.06 to 9.33)	NA[2]
hsa-miR-182*	0.0001	0.8547	2.35 (1.51 to 3.65)	(13, 162, 222, 223)[O]
hsa-miR-381	0.0001	0.6801	1.97 (1.40 to 2.79)	(184)[O]
hsa-miR-509-3-5p	0.0001	-0.3725	0.69 (0.57 to 0.83)	(224)[O]
hsa-miR-19a*	0.0002	0.5574	1.75 (1.31 to 2.33)	(162)[O]
hsa-miR-573	0.0007	0.6298	1.88 (1.31 to 2.70)	NA
hsa-miR-329	0.0031	-1.4082	0.25 (0.10 to 0.62)	(179)[Z]
hsa-miR-106b	0.0024	0.4525	1.57 (1.17 to 2.11)	(13, 162)[O]
hsa-miR-18b*	0.0042	0.6678	1.95 (1.24 to 3.08)	(162, 180)[O]
hsa-miR-521	0.0051	1.3416	I_II[3] = 2.10 (0.89 to 4.97) Rest[4] = 0.55 (0.40 to 0.76)	(183)[O]
hsa-miR-148a	0.0063	-0.2493	0.78 (0.65 to 0.93)	(181)[O]

[1]C.I.: Confidence Interval; [2]NA: No information found; [O]: Associated with Ovarian Cancer; [Z]: Associated with other cancer type; [3]I_II: Grade I or II tumor; [4]Rest: Any tumor grades other than I or II

Table 3.3 MicroRNAs associated with post-diagnostic recurrence on a cohort-independent or -dependent manner and supporting independent studies.

MicroRNA	*P* -Value	Estimate	Hazard Ratio (95% C.I.[1])	Relevant Literature References
hsa-miR-550*	<.0001	-2.1165	0.12 (0.05 to 0.29)	NA
hsa-miR-22*	<.0001	-1.4397	0.24 (0.12 to 0.46)	(13, 186, 189, 225)[O]
hsa-miR-223	<.0001	0.5267	1.69 (1.36 to 2.12)	(186, 189, 225)[O]
hsa-miR-146a	<.0001	-0.4869	0.62 (0.49 to 0.77)	(225)[O]
hsa-miR-497	0.0001	1.5869	Chemo[3] = 0.84 (0.69 to 1.03)	(187, 188)[O]
		1.125	C_O[4] = 0.53 (0.20 to 1.41)	
			Other[5] = 0.17 (0.08 to 0.35)	
hsa-miR-214*	0.0001	0.7059	2.03 (1.41 to 2.91)	(13, 166, 189, 225)[O]
ebv-miR-BHRF1-2*	0.0028	1.092	2.98 (1.46 to 6.10)	NA
hsa-miR-96	0.0065	0.1984	1.22 (1.06 to 1.41)	(180, 187, 190, 225)[O]
hsa-miR-924	0.0102	1.3019	3.68 (1.36 to 9.92)	NA
hsa-miR-28-3p	0.0109	1.1811	3.26 (1.31 to 8.09)	NA[2]
hsa-miR-369-3p	0.013	0.4208	1.52 (1.09 to 2.12)	(191)[Z]

[1]C.I.: Confidence Interval; [2]NA: No information found; [O]: Associated with Ovarian Cancer; [Z]: Associated with other cancer type; [3]Chemo: Only chemotherapy; [4]C_O: Chemotherapy plus another therapy; [5]Other: Any therapy other than chemotherapy

Table 3.4 Transcription factors associated with ovarian cancer survival.

Transcription Factor	Estimate	Hazard Ratio (95% C.I.[1])	Relevant Literature References
CLOCK	0.0097	0.81 (0.58 to 1.11)	(208, 209)[Z]
EGR1	0.0065	1.15 (1.03 to 1.28)	(192)[Z]
EGR2	0.0038	1.17 (1.06 to 1.30)	(193, 226)[Z]
ESR2	0.0065	0.66 (0.49 to 0.88)	(201-203)[O]
ETS2	0.0098	1.32 (0.99 to 1.76)	(202, 205)[Z]
FOS	0.0056	1.15 (1.03 to 1.28)	(227)[O]
HDAC3	0.0093	1.63 (1.13 to 2.37)	(210)[O]
HOXA1	0.0096	0.71 (0.53 to 0.97)	(211)[Z]
MYC	0.009	1.27 (1.05 to 1.54)	(199, 200, 228)[O]
NR5A1	0.0086	0.53 (0.34 to 0.82)	(207)[O]
POU2F2	0.008	0.64 (0.45 to 0.93)	(204)[Z]
TGFB1	0.0054	0.46 (0.32 to 0.66)	(195-198)[O]

[1]C.I.: Confidence Interval; [2]NA: No information found; [O]: Associated with Ovarian Cancer; [Z]: Associated with other cancer types

Table 3.5 Transcription factors associated with ovarian cancer recurrence.

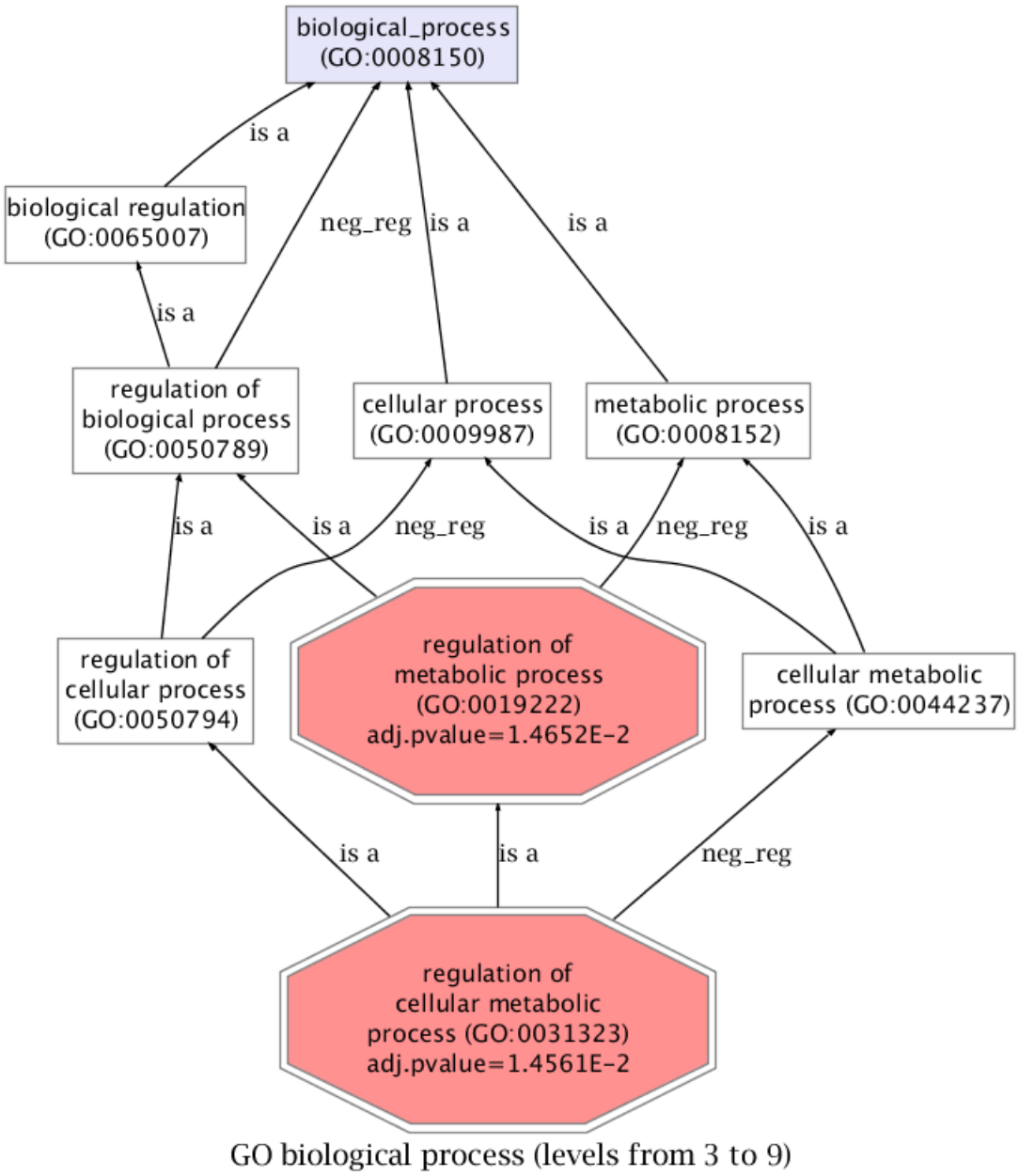
Transcription Factor	**Estimate**	**Hazard Ratio (95% C.I.[1])**	**Relevant Literature References**
CTCF	0.0063	1.71 (1.11 to 2.64)	(212, 213)[Z]
EGR1	0.0076	1.15 (1.03 to 1.28)	(192)[Z]
EGR2	0.0054	1.17 (1.05 to 1.31)	(193, 226)[Z]
FOS	0.0092	1.13 (1.01 to 1.27)	(194)[O]
MYOD1	0.0075	0.77 (0.63 to 0.95)	[(214, 215)[Z]
SOX18	0.0082	0.77 (0.62 to 0.95)	(216, 217)[Z]
TBP	0.0088	1.63 (1.19 to 2.24)	(218)[Z]
TGFB1	0.0088	0.56 (0.39 to 0.80)	(195-198)[O]

[1]C.I.: Confidence Interval ; [O]: Associated with Ovarian Cancer; [Z]: Associated with other cancer type

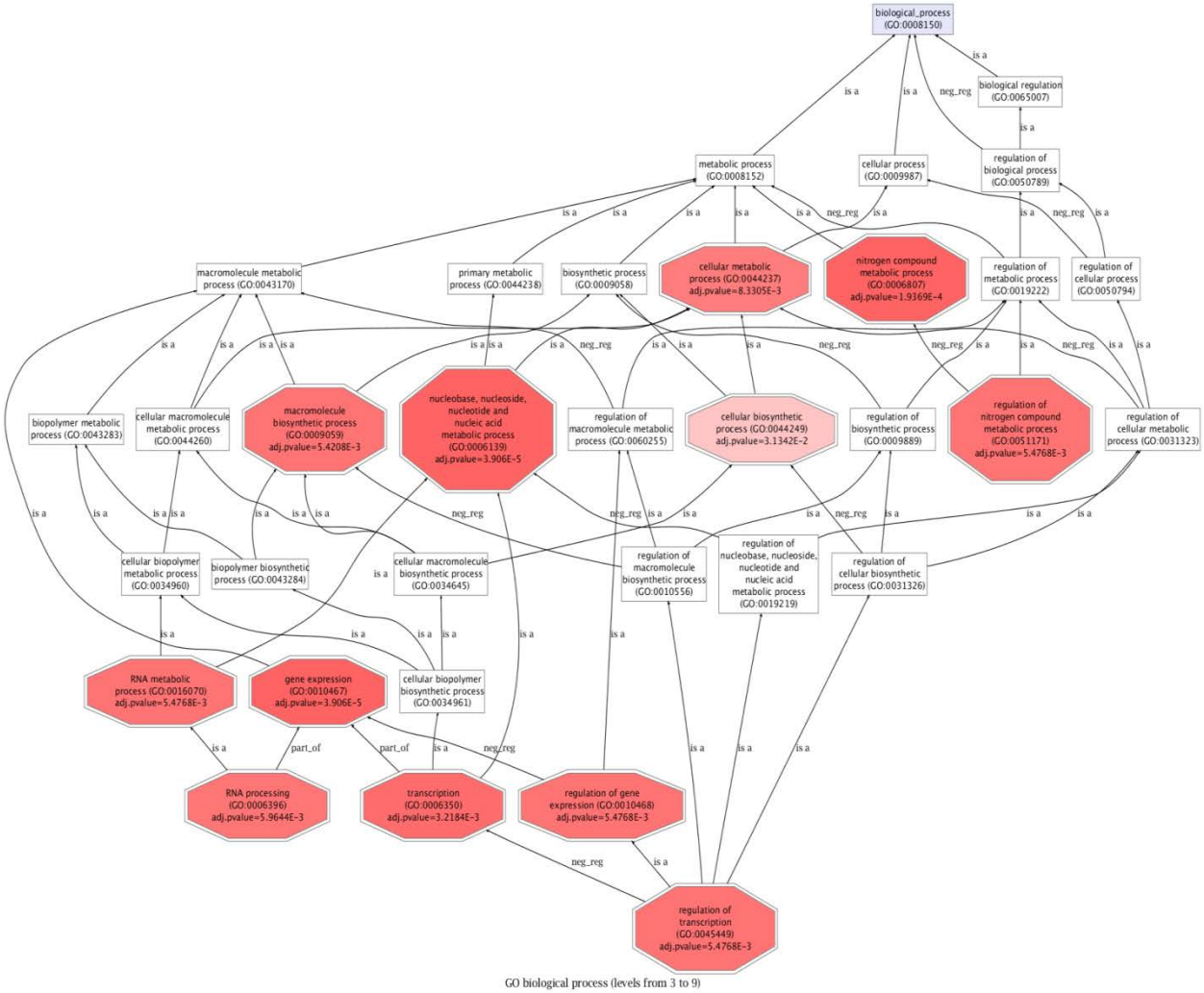
Table 3.6 Differentially enriched Gene Ontology biological processes among all target genes segmented by low and high hazard of ovarian cancer death or recurrence identified by set enrichment analyses.

Trait and GO Category	GO identifier	- hazard genes[1] In GO	- hazard genes[1] Not in GO	+ hazard genes[2] In GO	+ hazard genes[2] Not in GO	Log_e[3] (odds ratio)	FDR-P-value[4]
Survival							
regulation of cellular metabolic process	GO:0031323	218	502	198	766	0.52	1.46E-02
regulation of metabolic process	GO:0019222	229	491	214	750	0.49	1.47E-02
Recurrence							
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GO:0006139	117	175	321	1034	0.77	3.91E-05
RNA processing	GO:0006396	29	263	49	1306	1.08	5.96E-03
nitrogen compound metabolic process	GO:0006807	122	170	354	1001	0.71	1.94E-04
regulation of gene expression	GO:0010468	84	208	237	1118	0.64	5.48E-03
gene expression	GO:0010467	109	183	291	1064	0.78	3.91E-05
regulation of transcription	GO:0045449	78	214	213	1142	0.67	5.48E-03
cellular metabolic process	GO:0044237	264	223	492	668	0.47	8.33E-03
cellular biosynthetic process	GO:0044249	106	186	342	1013	0.52	3.13E-02
RNA metabolic process	GO:0016070	80	212	219	1136	0.67	5.48E-03
transcription	GO:0006350	82	210	219	1136	0.71	3.22E-03
regulation of nitrogen compound metabolic process	GO:0051171	84	208	237	1118	0.64	5.48E-03
macromolecule biosynthetic process	GO:0009059	97	195	283	1072	0.63	5.42E-03

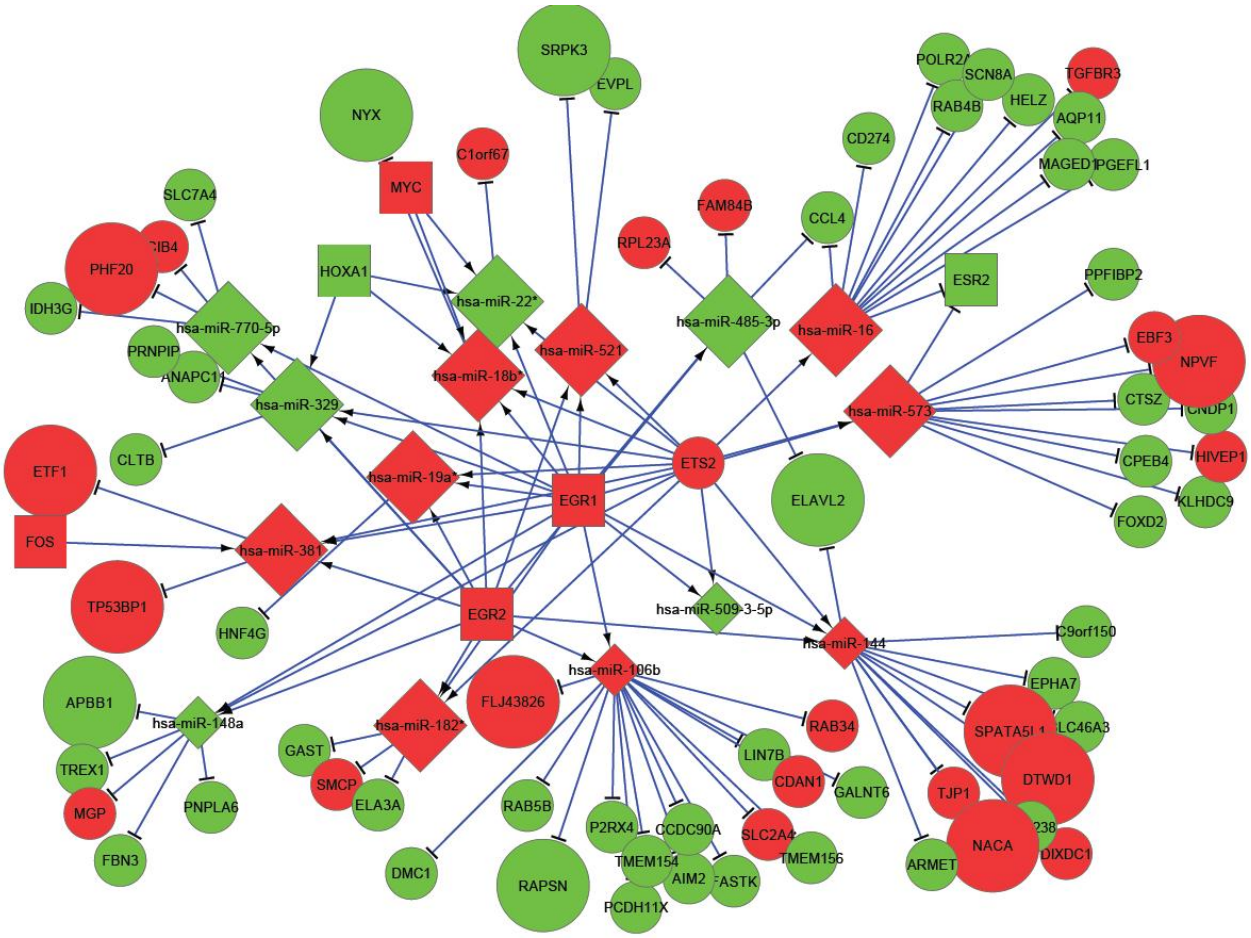
[1]- hazard genes: number of genes that have a negative association between the hazard of ovarian cancer death (higher survival) or recurrence and expression. [2]+ hazard genes: number of genes that have a positive association between the hazard of ovarian cancer death (lower survival) or recurrence and expression. [3]Log_e(Odds Ratio): values > 1 indicate that the category was more enriched among the genes that have a negative association with hazard than among the genes that have a positive association with hazard; values < 1 indicate that the category was more enriched among the genes that have a positive association with hazard than among the genes that have a negative association with hazard; Extreme values indicate higher difference in the enrichment percentages between the negative and positive association groups. Values close to zero indicate similar enrichment percentages between positive and negative association groups. [4]FDR-adjusted P-value: False discovery rate adjusted P-value of the log odds ratio test. Enrichment at FDR-adjusted Pvalue < 0.05) and ≥ 75 genes in the category.



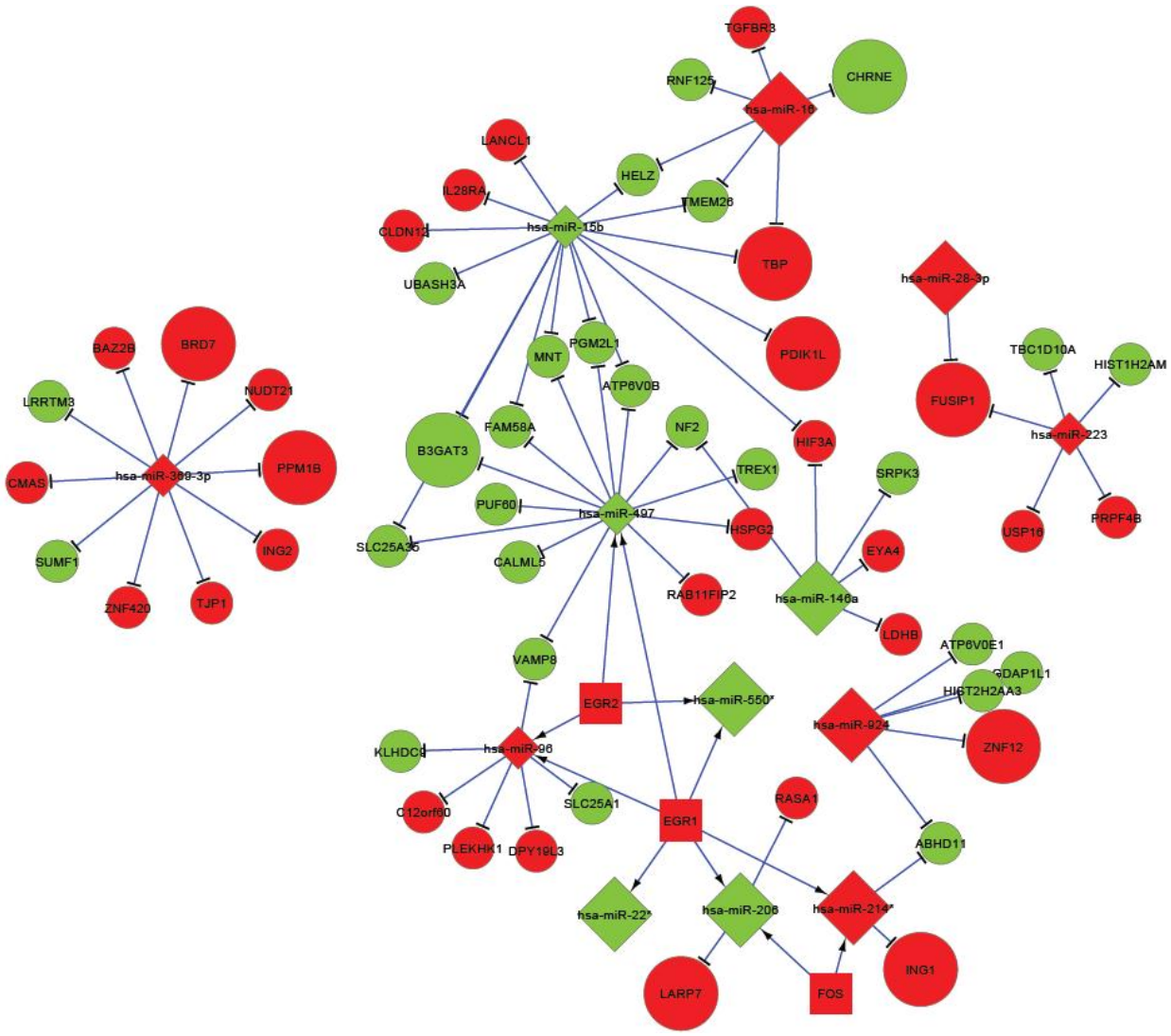
Supplementary Figure 3.1 Relation between the Gene Ontology biological processes associated with ovarian cancer death inferred from the set enrichment analysis.



Supplementary Figure 3.2 Relation between the Gene Ontology biological processes associated with ovarian cancer recurrence inferred from the set enrichment analysis.



Supplementary Figure 3.3 Targeted sub-network of microRNAs, transcription factors, and target genes associated with ovarian cancer survival.
Node Shape: microRNA=diamond, target gene=circle, transcription factor=square; Node Color: Red indicates increased hazard with high expression, Green indicates decreased hazard with high expression; Node Size: larger indicates a more extreme association (HR ≥ |1.6|), smaller indicates a less extreme association



Supplementary Figure 3.4 Targeted sub-network of microRNAs, transcription factors, and target genes associated with post-diagnostic recurrence in ovarian cancer.
Node Shape: microRNA=diamond, target gene=circle, transcription factor=square; Node Color: Red indicates increased hazard with high expression, Green indicates decreased hazard with high expression; Node Size: larger indicates a more extreme association (HR ≥ |1.6|), smaller indicates a less extreme association

Chapter IV

Progression in the dysregulation of transcriptome pathways from acute to chronic illicit drug exposure*

ABSTRACT

Exposure to illicit drugs elicit changes in the brain transcriptome that result in the dysregulation of pathways. Pathway changes as the exposure progresses from acute to chronic are only partially understood. Microarray gene expression studies are typically small in size and often consider specific challenges. This situation hinders the elucidation of pathway dysregulation within and across levels of drug exposure persistency. To address these limitations, a meta-analysis of six individual microarray experiments measuring gene expression in the brain of mice under acute and chronic drug exposure was undertaken. Validation on an independent data set, subsequent functional analysis, and network visualization offered insights into the network changes across drug exposure levels. Meta-analyses uncovered 263 and 2,641 genes differentially expressed (FDR-adjusted P-value < 0.1) between control and acute and chronic exposure, respectively. These results confirm that the more extensive the exposure, the more extensive the impact on the transcriptome profile. The MAPK signaling pathway and the molecular functions of protein dimerization and leucine zipper transcription factor were enriched in response to acute exposure. These processes give way to the enrichment of the molecular functions of ubiquitin conjugation, nucleotide binding, RNA splicing, and associated ribosomal pathways which were enriched in response to chronic exposure. Individual genes in these processes have been previously associated with drug exposure and reward-dependent behaviors. The meta-analyses allowed the uncovering of consistent profiles that support novel functional

*K.R. Delfino, B.R. Southey, and S.L. Rodriguez-Zas. Submitted to PLoS One

understanding. This is the first study able to detect the progression of drug exposure pathways using meta, functional, and network-analyses.

Introduction

Microarray gene expression experiments have helped in the characterization of hundreds of genes associated with drug addiction behaviors (58, 63). The mice striatum provides a well-established system to study gene expression changes associated with illicit drug exposure (229). The detection of differentially expressed genes between control and drug-treated striatum samples offers insights into the pathways affected by drug abuse as well as potential prognostic tools and diagnostic therapies (63, 230).

The identification and understanding of genes and gene pathways dysregulated by drug exposure has been hindered by the limited size and focus of most experiments on a particular drug (e.g. cocaine, morphine, or methamphetamine) or single exposure (66, 67). Applying accepted definitions of acute (single dose exposure) and chronic (multiple dose exposure), few experiments have investigated the impact of acute and chronic drug exposure on the striatum transcriptome profile (61-63). However, no reports have investigated the dysregulation of pathways that are shared by different drugs or the progression of processes and pathways from acute to chronic drug exposure. The goals of this study were to survey the genes, pathways, sub-networks and molecular functions associated with illicit drug exposure, and to investigate the progress of the gene and pathway dysregulation from acute to chronic exposure. To accomplish this, five mice microarray experiments that evaluated different illicit drugs and extent of exposure were meta-analyzed. Results were validated on an independent data set. Insights into the meta-analysis results were gained from functional analysis and network visualization.

Materials and Methods

Data sets

Five mice microarray experiments that compared gene expression between saline (control) and acute or chronic exposure to illicit drugs were integrated. The experiments are available in the Gene Expression Omnibus (GEO) repository ((231), http://ncbi.nlm.nih.gov/geo). The GEO identifiers of the experiments and associated reference are: GSE10869 (66), GSE10870 (67), GSE13386 (61), GSE7762 (62), and GSE8948 (232). Cocaine was used in experiments GSE10869, GSE13386, GSE8948, and GSE10870, and morphine was used in experiment GSE7762.

All experiments evaluated a single drug dose (defined as acute exposure) and saline treatment, and striatum samples were obtained one hour after drug exposure (61, 62). In addition, experiments GSE7762 and GSE13386 included multiple drug exposures (defined as a chronic exposure), and samples were obtained four hours after the final exposure (61, 62).
In experiments GSE10869, GSE10870, and GSE8948, exposure constituted intraperitoneal (i.p.) injection of either 25 mg/kg cocaine or saline, and the mice were sacrificed after one hour by cervical dislocation. For GSE13386, in order to habituate the mice prior to handling, all mice were first injected with 100 µl saline (vehicle) once daily for eight days. For the acute cocaine exposure, mice were injected with a dose of 20 mg/kg cocaine or 100 µl saline on day nine. The striata were harvested four hours after the final injection. For chronic cocaine studies, mice were injected with 20 mg/kg cocaine or 100 µl saline once daily for fifteen days, and the striata were harvested four hours after the last injection. For GSE7762, acute (20 mg/kg, subcutaneous) and chronic morphine administration (10-40 mg/kg, three times daily for five days) were compared. On the fifth day of the experiment, acute treated mice were injected with a single dose of

morphine (20 mg/kg) and killed by decapitation after four hours. Chronically treated mice were injected with increasing doses of morphine for five days. Mice received morphine thrice daily (09:00 hours, 13:00 hours and 17:00 hours) for four days using a dosing schedule of 10, 20, 40 and 40 mg/kg of morphine on days one, two, three and four, respectively. On the last day, a final morphine dose of 40 mg/kg was administered and four hours after the last injection, the mice were sacrificed. Mice in control groups were killed four hours after the last injection of saline. The number of saline: acute: chronic dose samples per experiment were: GSE10869 6:12:0; GSE10870 6:6:0; GSE8948 6:10:0; GSE13386 12:6:6; and GSE7762 12:12:12.

GSE10869 used mice with ablation of the *Camk4* gene in neurons expressing the dopamine receptor 1 (*Drd1a*) and were 5- to 6-week-old $Creb1^{Camkcre4}$, $Crem^{-/-}$ mice (before onset of neurodegeneration) and 5- to 10-week-old $Creb1^{Camkcre4}$, $Crem^{+/-}$ mice with littermate controls ($Creb1^{loxP/loxP}$, $Crem^{+/-}$ and $Creb1^{loxP/loxP}$, $Crem^{-/-}$) (66). In GSE10870, behavioral experiments were performed on 10- to 15-week-old transgenic mice using Cre-negative, $Srf^{loxP/loxP}$ or $Srf^{loxP/wt}$ littermates as controls (67). In GSE8948, transgenic 5- to 6-week-old littermate male and female mice with targeted mutation of the *Creb1* gene were used (232). In GSE13386 lines CP73 and CP101 transgenic mice that expressed enhanced green fluorescent protein (EGFP)-tagged ribosomal protein L10a were used (61). In GSE7762 adult male (8- to 10-week-old) 129P3/J (000690), DBA/2J (000671), C57BL/6J (000664), and SWR/J (000689) mice were used (62).

The Affymetrix GeneChip Genome Mouse 430 2.0 Array platform was used in experiments GSE13386 and GSE7762, and the Affymetrix GeneChip Mouse 430A 2.0 array platform was used in experiments GSE10869, GSE8948, and GSE10870 ((233), http://www.affymetrix.com). All probe sets represented on the GeneChip Mouse Genome 430

2.0 array are included on the GeneChip Mouse Genome 430A 2.0 array and represent approximately 39,000 transcripts from 14,000 genes. Probes common to both platforms were considered resulting in 45,036 probes analyzed.

Probe expression intensities were normalized using the GCRMA affy R package following standard protocols (234-238) and centered by microarray corresponding to the first stage of a two-stage analysis (83, 84, 90, 239-241). Data processing and normalization were implemented in Beehive ((77, 114); http://stagbeetle.animal.uiuc.edu/Beehive1.0).

Meta-analyses

Individual-experiment and meta-analyses of all experiments were implemented (83, 84). In the individual-experiment analyses, the normalized expression of each gene was described with a linear mixed-effects model that included the fixed effects of drug exposure (saline, acute, or chronic) and the random effect of mice sample nested within exposure. Lists of differentially expressed genes from each experiment were obtained, and the genes that overlap two lists were identified. In the sample-level meta-analysis, the normalized gene expression across all experiments was described with a linear mixed-effects model that included the fixed effects exposure and the random effects of sample and experiment. The random effects of sample and experiment allowed adjustment for differences between experiments such as strain and sex differences. Experiments were assumed to be normally distributed with 0 mean and unstructured variance to accommodate heterogeneity of variance among experiments.

Results from the individual-experiment analyses were denoted with the experimental GEO identifier (e.g. GSE7762) as the prefix, and the results from the meta-analysis were denoted with the Meta prefix. Of interest was the contrast between acute (single) drug exposure and

saline treatment (denoted with suffix S) and between chronic (multiple) drug exposure and saline treatments (denoted with suffix M) within experiment. To protect for multiple-testing, false-discovery rate (FDR) adjustment of the P-values was used to identify differentially expressed genes across drug exposure levels (78). Probes were considered differentially expressed when the FDR-adjusted P-value was less than 0.1 and fold change greater than |1.25|. All analyses were performed using the mixed model procedure in SAS version 9.3 ((242), SAS Institute, Cary, NC, USA).

.

Functional enrichment and gene networks

The enrichment of Gene Ontology (GO) ((176), http://www.geneontology.org/) molecular functions and biological processes, and KEGG ((177), http://www.genome.jp/kegg/) pathways was studied among the genes differentially expressed that were identified by the meta-analyses. The functional analysis consisted of Fisher's exact (two-tailed) test implemented in DAVID v6.7 ((92), http://david.abcc.ncifcrf.gov/). DAVID was used to identify the functional categories enriched among all genes differentially expressed (FDR-adjusted P-value < 0.1) between saline and acute exposure (MetaS) and saline and chronic exposure (MetaM). The DAVID Functional Annotation Clustering was used to detect enriched categories while accommodating for the relationships among the annotation terms and minimizing redundant and heterogeneous annotation contents ((243), http://david.abcc.ncifcrf.gov/manuscripts/fuzzy_cluster/).

The networks of genes differentially expressed (FDR-adjusted P-value < 0.01) in the acute and chronic drug exposure relative to saline resulting from the MetaS and MetaM analyses were depicted using Cytoscape ((95), http://www.cytoscape.org/), an open source software platform for visualizing networks and including attributes. The distribution and connectivity of

the genes within sub-networks and the overall networks were characterized. Network pathways were obtained from WikiPathways ((97, 98), http://www.wikipathways.org) and implemented in Cytoscape with the GPML plugin ((96), http://apps.cytoscape.org/apps/gpmlplugin). The GPML plugin for Cytoscape is a converter between Cytoscape networks and the GPML (GenMAPP Pathway Markup Language) pathway format.

Validation

The findings from the meta-analysis were validated on the independent experiment GSE15774 (244). Experiment GSE15774 compared the effects of various substances of abuse on gene expression profiles in the mouse striatum. Samples were obtained one, two, four, or eight hours after a single acute morphine (20 mg/kg), heroin (10 mg/kg), ethanol (2 g/kg), nicotine (1 mg/kg), methamphetamine (2 mg/kg) or cocaine (25 mg/kg) i.p. injection, with respective saline and naïve control groups. Samples from two mice were pooled for each microarray. Three biological replicates of the microarrays were prepared per experimental group. Illumina MouseWG-6 v1.1 and 84 Illumina MouseWG-6 v2 array platforms that included probes representing approximately 48,000 transcripts were used. Data processing, normalization, and analysis followed the same procedures that were applied to the five individual experiments analyzed. Experiment GSE15774 was well-suited for independent validation because the drug exposures were comparable to the meta-analyzed experiments. Also, the array platform used in GSE15774 was vastly distinct to that used in the meta-analyzed experiments, and samples were pooled and used in triplicate instead of individually studied.

Results and Discussion

Comparison of single experiment and meta-analyses of acute and chronic drug exposure

Tables 4.1 and **4.2** list the number of significant (FDR-adjusted P-value < 0.1) probes per analysis and the overlap of probes detected between pairs of analyses contrasting acute (single dose) exposure relative to saline (denoted with the suffix S) and chronic (multiple dose) exposure relative to saline (denoted with the suffix M). Results from the individual experiment analyses are labeled with the corresponding GSE identifier as prefix and results from the meta-analysis across experiments are labeled with the Meta prefix. The number of significant (FDR-adjusted P-value < 0.1) probes was 57, 191, 1, 1243, 3, 0, 2531, 263, and 2641 for the analyses GSE10869, GSE10879, GSE13386S, GSE7762S, GSE8948, GSE13386M, GSE7762M, MetaS, and MetaM, respectively. The common FDR-adjusted P-value < 0.1 threshold corresponded to the unadjusted P-values < 0.00025, 0.00085, 0.00001, 0.00275, 0.00002, 0.000001,0.0056, 0.00060, 0.0059, for the analyses GSE10869, GSE10879, GSE13386S, GSE7762S, GSE8948, GSE13386M, GSE7762M, MetaS, and MetaM, respectively

Table 4.1 demonstrates the enhanced precision of the meta-analysis to detect differentially expressed transcripts compared to the analysis of individual experiments separately. From the 57 transcripts detected in GSE10869, at most 40 transcripts were confirmed in one other experiment, yet 44 transcripts were confirmed by the meta-analysis. Likewise, from the 191 transcripts detected in the analysis of GSE10870, at most 40 transcripts were confirmed in one other experiment, yet 84 were confirmed by the meta-analysis.

The validation of the findings from the meta-analyses using the independent experiment GSE15774 supports our findings. The correlation between the estimates from the meta-analysis and GSE15774 was over 50%. This correlation is notable considering the experimental design of

the validation data set that included two sets of pooled samples in triplicate. The experimental design, though comprehensive on treatments evaluated, had limited precision and statistical power. Consistent with our findings, the reported analysis of GSE15774 identified 42 drug-responsive genes (244).

Confirmation of the benefits of meta-analysis over simple overlap between individual experiment analyses is that the average overlap between an individual experiment and meta-analysis was 52%, meanwhile the average overlap between any two individual experiment analyses was 42% (excluding experiments with 0 transcripts detected). **Table 4.1** also demonstrates that the results from the meta-analysis were not biased by any one experiment. From the 1,243 differentially expressed transcripts detected in the analysis of experiment GSE7762, at most 19 transcripts were confirmed by one other study and 55 transcripts were confirmed by the meta-analysis. The majority of overlapping transcripts between individual analyses expressed the same trends, and thus equal sign of the expression contrast. For example all 40 differentially abundant transcripts in common between GSE10869 and GSE 10870 had the same estimate sign.

Table 4.2 offers insights into the transition between acute and chronic drug exposure. Comparison of the differentially expressed transcripts detected by the meta-analysis of acute dose (MetaS) and chronic dose (MetaM) identified the more consistently differentially expressed transcripts within or across drug exposure. One-third (34%) of the transcripts differentially expressed in MetaS were also differentially expressed in MetaM. There were more differentially expressed transcripts in the contrast between chronic exposure and saline than between acute exposure and saline treatment, both in the individual experiment analysis and meta-analysis.

Comparison of the number of differentially expressed transcripts identified by the single experiment analysis (GSE7762) and meta-analysis (MetaM) contrasting chronic drug exposure and saline treatment confirmed the advantages of meta-analysis noted for the acute drug exposure versus saline treatment (**Table 4.2**). Meta-analysis was able to detect more differentially expressed transcripts than the single experiment analysis. This trend is the result of meta-analysis gathering more precision through the consideration of multiple analysis and combination of estimates that may not reach statistical significance in the single-experiment analysis. Also, meta-analysis is not biased by one experiment because although the single experiment analysis of GSE7762M detected all the differentially expressed transcripts identified by experiments, including chronic exposure, only 16% of the transcripts detected by the single-experiment analysis were detected by the meta-analysis MetaM (**Table 4.2**).

The higher number of differentially expressed transcripts under chronic exposure relative to acute exposure could be due to a more marked effect of the drug on the expression of the same transcripts or related transcripts in the same pathways or to the effect of the chronic exposure on additional transcripts and pathways. The additional number of transcripts associated with the chronic exposure is unlikely to be due to the additional time required by the chronic exposure because samples from the acute and chronic drug exposures were collected at the same time in the experiments.

Transcripts associated with acute and chronic drug exposure

Tables 4.3 and **4.4** list the highly differentially expressed transcripts identified by the MetaS and MetaM analyses associated with addiction and behavioral disorders. A comprehensive list of all differentially expressed transcripts (FDR-adjusted P-value < 0.01)

identified by the MetaS and MetaM are summarized in **Supplementary Materials Tables 4.1** and **4.2**. Numerous genes previously associated with addiction, reward-related behaviors, and neurological disorders were identified by the MetaS and MetaM analyses with the supporting references listed in **Tables 4.3** and **4.4**, respectively. FBJ osteosarcoma oncogene B (*Fosb*), activity regulated cytoskeletal-associated protein (*Arc*), TCDD-inducible poly(ADP-ribose) polymerase (*Tiparp*), Jun oncogene (*Jun*), calcium/calmodulin-dependent protein kinase I gamma (*Camk1g*), FBJ osteosarcoma oncogene (*Fos*), G-protein coupled receptor 3 (*Gpr3*), cyclin-dependent kinase inhibitor 1A (*Cdkn1a*), Rho family GTPase 3 (*Rnd3*), selenoprotein W, muscle 1 (*Sepw1*), dynamin 1 (*Dnm1*), and regulator of G protein signaling 7 (*Rgs7*) have been reported to be linked to illicit drug dependency.

The AP-1 family, comprised of Fos and Jun family proteins, has been shown to play a significant role in mediating adaptations to drugs of abuse (245, 246). *Fosb* is one of the key proteins implicated in the gene expression changes in the nucleus accumbens caused by drugs of abuse which contributes to the complex circuit adaptations underlying addiction-related behaviors (245). *Fos* has been shown to be regulated in response to morphine exposure in adult mice (247). Expression of *Jun* in brain regions of adult mice decrease development of cocaine-induced conditioned place preference, suggesting reduced sensitivity to rewarding effects of cocaine (246). *Arc* expression is altered when interacted with stress and cocaine in mice and may be a potential molecular target modulated by stress to alter cellular sensitivity to cocaine (248). Modulatory activity of *Gpr3* in the brain has been related to the control of emotional behaviors, and altered signaling pathways of *Gpr3* have been associated with the neurobiological substrate involved in developing addiction to cocaine in humans (249). Expression levels of *Rnd3* are up-regulated in the hippocampus and prefrontal cortex of mice when injected with

cocaine (250). *Sepw1*, detected in the brain and shown to be associated with alcohol and alcohol drinking behavior, was differentially expressed between operant ethanol and water injections in rodents (251, 252). Regulation of *Dmn1* expression by nicotine, via a microRNA pathway, indicates that *Dmn1* may play an important role in neural plasticity and the underlying mechanism of nicotine addiction (115). *Rgs7* plays an essential role in controlling cocaine sensitization and is associated with higher sensitivity to locomotor stimulating effects of cocaine in rodents (253).

Among the genes known to be associated with neurological disorders that were detected by the meta-analyses are: growth arrest and DNA-damage-inducible, beta (*Gadd45b*), dual specificity phosphatase 1 (*Dusp1*), TAR DNA binding protein (*Tardbp*), survival motor neuron 1 (*Smn1*), Family with sequence similarity 123A also known as APC membrane recruitment 2 (*Fam123a*; also known as *Amer2*), Rho GTPase activating protein 6 (*Arhgap6*), hyaluronan and proteoglycan link protein 1 (*Hapln1*), SERTA domain containing 1 (*Sertad1*), N-myc downstream regulated gene 4 (*Ndrg4*), hepatic leukemia factor (*Hlf*), hairy and enhancer of split 5 (*Hes5*), phosphodiesterase 4A, cAMP specific (*Pde4a*), and nemo like kinase (*Nlk*).

Stress response gene *Gadd45b* is induced by neuronal activity and has been implicated in the promotion of adult neurogenesis (254, 255). *Tiparp*, which previous studies have found to have a strong association with alcohol dependence, is located in a region that influences antisocial behavior and substance dependence vulnerability (256). *Sertad1* expression is essential for developmental neuronal death in the cerebral cortex and may be a suitable target for investigation on Alzheimer's disease (257). *Dusp1* expression is up-regulated in the sensory input neurons of the thalamus and thalamic-recipient layer IV and VI neurons of the mouse cortex suggesting it has specialized regulation to sensory input neurons of the thalamus and

telencephalon (258). *Tardbp* is correlated with behavioral and cognitive changes and has been found in multiple neurological disorders, including sporadic motor neurone disease, amyotrophic lateral sclerosis, frontotemporal dementia, Parkinson's disease, and Alzheimer's disease (259). *Smn1* is involved in motor neuron pathology and behavioral alterations in the mouse model (260). *Fam123a* is strongly expressed in the central as well as the peripheral nervous system of the mouse, and plays an important role during neurogenesis (261). *Arhgap6*, a candidate for cell-based therapy of neurodegenerative diseases, is involved in neurite outgrowth, early neuronal cell development, and neuropeptide signaling and synthesis (262). *Hes5* expression enhances fear retention (263). *Hapln1* plays a role in the control of the central nervous system plasticity and triggers the formation of perineuronal nets in mice (264). Associated with spatial learning and memory, *Ndrg4* is expressed in various neurons of the brain and is necessary for the preservation of spatial learning and the resistance to neuronal cell death caused by ischemic stress (265). Expressed in the central nervous system, *Hlf* plays a role in the function of differentiated neurons in the adult nervous system (266). *Pde4a* has been implicated in the control of cognitive function and plays a role in cognitive deficit in schizophrenia (267). Up-regulated in the brain during Huntington's disease, *Nlk* supports growth of sensory neurons and is involved in neuronal growth, glucose metabolism, cell motility, and differentiation (268).

Processes and pathways associated with acute and chronic drug exposure

Tables 4.5 and **4.6** list the top (FDR-adjusted P-value < 0.1) enriched GO molecular functions, biological processes, KEGG pathways, and other DAVID categories expressed by the transcripts identified by the MetaS and MetaM analyses. **Supplementary Tables 4.3** and **4.4** include the gene identifiers that map to these and other significantly (raw P-value < 0.005) enriched categories. The comparison of the molecular functions and biological processes

enriched among the transcripts differentially expressed under acute and chronic drug exposure aids in the understanding of the progression in the mechanisms altered by acute and chronic drug exposure.

Consistent with the number of genes detected by either meta-analysis, the number of categories significantly associated with drug exposure and the number of genes within enriched category was lower in MetaS relative to MetaM. The top enriched categories among the genes identified by the MetaS analysis include protein dimerization and leucine zipper transcription factor that have been previously associated with illicit drug usage and other reward dependence behaviors. Little is known about the association between genes in the protein dimerization category and drug exposure. Mu-opioid receptor agonists promote dimerization of *wntless* and mu-opioid receptors, preventing *wntless* from mediating Wnt protein secretion that is critical for neuronal development (269). Adaptations in this pathway may be a novel pharmacological target in the treatment of opiate addiction and pain (269).

Expression of genes in the leucine zipper transcription factor category, such as *c-fos* and *jun B*, can be regulated by stimuli that affect the dopaminergic nigrostriatal system (270). Studies using adult rat striatum and cocaine showed activation of members of leucine zipper transcription factor categoryin response to catecholaminergic stimulation (270). Cocaine induces coordinate expression of *c-fos* and *jun B* mRNAs in neurons of a rat's striatum, and may contribute to response specificity of striatal neurons to stimulation by monoamines including dopamine (271). Stimulation of dopamine receptors plays a huge role in the effects on addictive behavior, moment control, and working memory (272).

Enriched categories such as ubiquitin conjugation, nucleotide binding, and RNA splicing detected by the MetaM analysis are in agreement with previous reports. In ubiquitin cojugation,

the dysregulation of Ubl-substrate modification and mutations in the Ubl-conjugation machinery are involved in the etiology and progression of a number of neurodegenerative disorders (273). Inhibition of the ubiquitin system occurs by aggregated proteins (274). Although still poorly understood, the aggregation of brain proteins into defined lesions is emerging as a common theme in sporadic and hereditary neurodegenerative disorders (274).

Processes remaining enriched from acute to chronic drug exposure

Table 4.7 lists the top (P-value < 0.005) enriched biological processes and molecular functions among the differentially expressed transcripts both in the acute and chronic drug exposure contrasts against saline treatment identified by the MetaS and MetaM meta-analyses. **Supplementary Table 4.5** includes the gene identifiers that map to these enriched categories. Working with the 83 transcripts differentially expressed both in the MetaS and MetaM analysis, a less extreme significant threshold was used to contain the potential false negative rate of enrichment on a low number of transcripts. The identification of biological categories shared by the acute and chronic drug exposure provides leads on the processes that are permanently dysregulated from the initiation during the progression of drug exposure. The enduring enrichment of these categories suggests critical processes that may be at the center of the permanent harmful effects of drug addiction. These processes may be at the center of recovery therapies that have a lasting impact.

Categories enriched among the genes detected both by the MetaS and MetaM analyses are in agreement with previous reports. Spliceosome and RNA degradation have been associated with neuropsychiatric disorders and are candidates for their therapeutic mechanism (275). Abiotic stimulus has been associated with differences in locomotor activity. Depression,

Parkinson's disease, Huntington's disease, and activity disorders are all associated with deficits in locomotion (276).

Network visualization

Investigation of the relationship between the genes in categories associated with acute and chronic drug exposure offered insights into the unique gene relationships. The networks for the MAPK signaling pathway (enriched in the acute versus saline contrast) and for the ErbB pathway (enriched in the chronic versus saline contrast) are depicted in **Figures 4.1** and **4.2.** Acute drug exposure had a negative effect on the MAPK signaling pathway, resulting on under-expression of multiple key genes (e.g. *Jun, Fos*). **Figure 4.1** highlights the main trends. Many of these genes code elements on the terminal end of the pathway (*Atf4, Nr4a1, Jun, Hspa8, Fos, Atf4, Dusp1*) with direct effect on cell proliferation and differentiation.

Chronic drug exposure had mostly positive effects on various subnetworks of the ErbB pathway (**Figure 4.2**). Chronic drug exposure resulted in the over-expression of genes associated with protein synthesis and the mTOR signaling pathway and metabolism. Noteworthy is the under-expression of *Hbegf* and over-expression of *Erbb4* as result of chronic drug exposure identified in this study. The KEGG pathway reports a positive association between these two components. Chronic exposure also resulted in over-expression of genes associated with proliferation, adhesion, and migration. Chronic drug exposure was associated with under-expression of genes associated with angiogenesis and adhesion through the MAPK signaling pathway, however, more terminal genes were not differentially expressed.

Figures

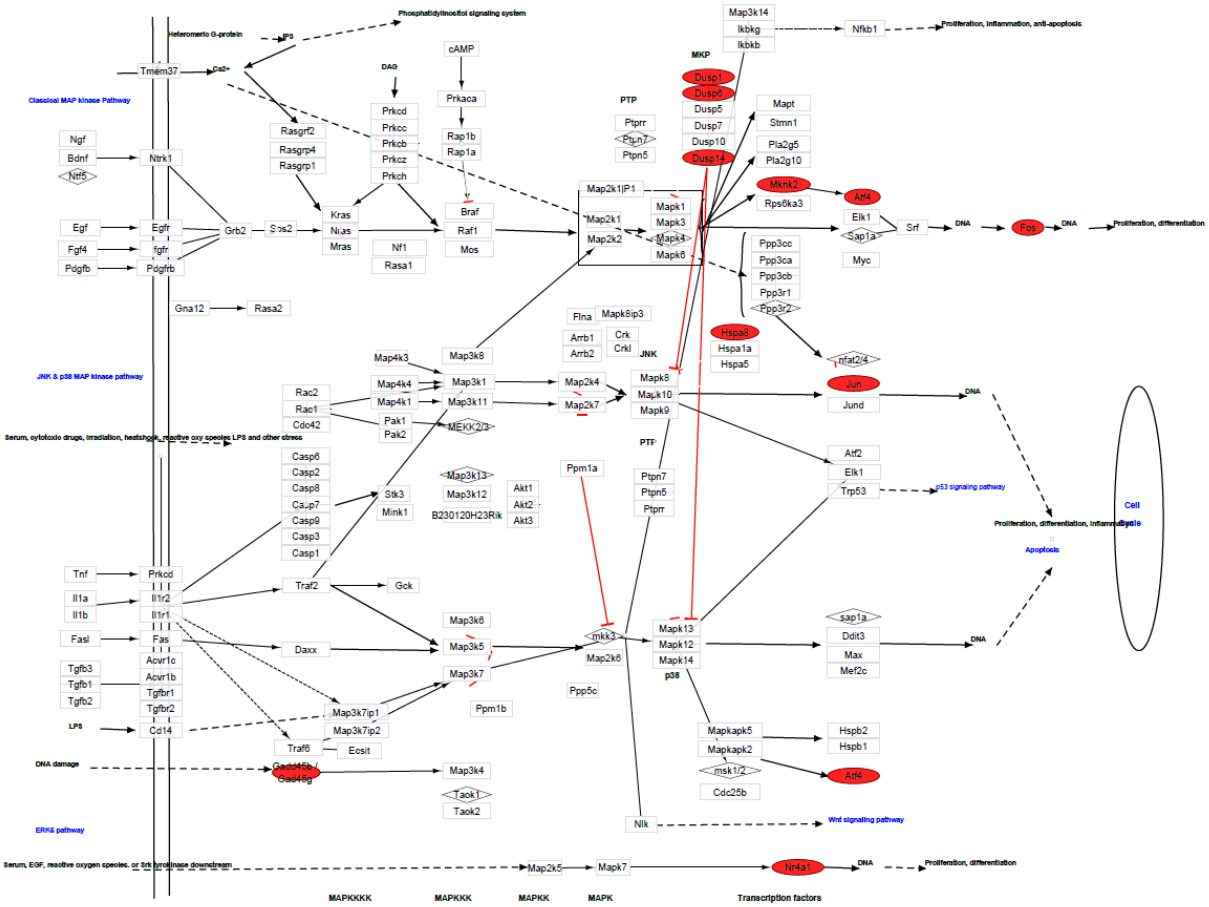


Figure 4.1 MAPK Signaling Pathway
Node Shape: Oval: Significantly expressed gene in platform, Square: Gene in platform but not significantly expressed, Diamond: Not in platform; Node Color: Red=negative gene estimate; Green=positive gene estimate; Line Edge: Arrow=promotion, T-shape=inhibition

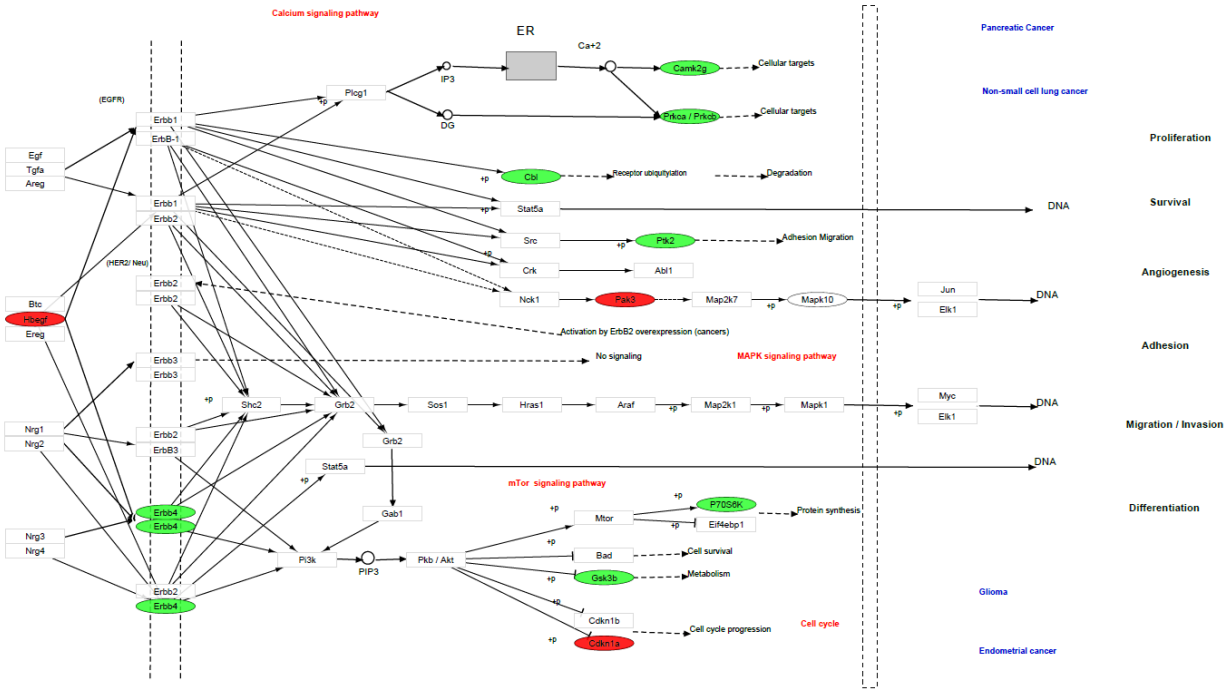


Figure 4.2 ErbB Pathway
Node Shape: Oval: Significantly expressed gene in platform, Square: Gene in platform but not significantly expressed, Diamond: Not in platform; Node Color: Red=negative gene estimate, Green=positive gene estimate; Line Edge: Arrow=promotion, T-shape=inhibition

Tables

Table 4.1 Number (and relative percentage) of transcripts differentially expressed (FDR-adjusted P-value < 0.1) between acute dose drug exposure and saline treatment resulting from the individual experiment and meta-analyses

	GSE10869S[a]	GSE10870S	GSE13386S	GSE7762S	GSE8948S	MetaS
GSE10869S	57[b]	40 (40)[c]	0 (0)	13 (12)	0 (0)	44 (0)
GSE10870S	70%	191	0 (0)	19 (15)	2 (2)	84 (0)
GSE13386S	0%	0%	1	0 (0)	0 (0)	0 (0)
GSE7762S	23%	10%	0%	1243	0 (0)	55 (1)
GSE8948S	0%	67%	0%	0%	3	2 (0)
MetaS	77%	44%	0%	21%	67%	263

[a]Prefix GSE denotes the individual experiment using the Gene Expression Omnibus identifier and prefix Meta denotes meta-analysis. Suffix S denotes that the contrast between a acute dose drug exposure and saline treatment was considered; [b]Diagonals correspond to the detected number of transcripts within analysis, upper off-diagonals correspond to the number of transcripts detected by two analyses, and lower off-diagonals correspond to the percentage of transcripts detected by two analyses; [c]Numbers in parenthesis indicate the number of overlaps that also have consistent estimate signs

Table 4.2 Number (and relative percentage) of transcripts differentially expressed (FDR-adjusted P-value < 0.1) between acute (S) and chronic (M) dose drug exposure and saline treatment resulting from the individual experiment and meta-analyses

	GSE13386S[a]	GSE13386M	GSE7762S	GSE7762M	MetaS	MetaM
GSE13386S	1[b]	0 (0)[c]	0 (0)	0 (0)	0 (0)	0 (0)
GSE13386M	0%	0	0 (0)	0 (0)	0 (0)	0 (0)
GSE7762S	0%	0%	1243	291 (282)	55 (1)	340 (338)
GSE7762M	0%	0%	23%	2531	160 (0)	399 (9)
MetaS	0%	0%	21%	61%	263	89 (88)
MetaM	0%	0%	27%	16%	34%	2641

[a]Prefix GSE denotes the individual experiment using the Gene Expression Omnibus identifier and prefix Meta denotes meta-analysis. Suffix S denotes that the contrast between an acute dose drug exposure and saline treatment was considered. Suffix M denotes that the contrast between a chronic dose drug exposure and saline treatment was considered ; [b]Diagonals correspond to the detected number of transcripts within analysis, upper off-diagonals correspond to the number of transcripts detected by two analyses and, lower off-diagonals correspond to the percentage of transcripts detected by two analyses; [c]Numbers in parenthesis indicate the number of overlaps that also have consistent estimate signs

Table 4.3 Most significant differentially expressed transcripts between the acute drug exposure and saline treatment identified by the MetaS meta-analysis and associated with addiction, behavioral or neurological disorders.

Probe ID	Gene Symbol	Estimate[1]	Standard Error	FDR P Value	Reference
1422134_at	*Fosb*	-1.31	0.17	2.79E-07	(245)
1449773_s_at	*Gadd45b*	-0.93	0.12	2.79E-07	(254, 255)
1418687_at	*Arc*	-1.46	0.21	4.19E-06	(248)
1426721_s_at	*Tiparp*	-1.14	0.16	4.19E-06	(188)
1417406_at	*Sertad1*	-0.71	0.11	6.23E-06	(257)
1448830_at	*Dusp1*	-0.94	0.14	6.32E-06	(258)
1417409_at	*Jun*	-0.78	0.12	1.67E-05	(246)
1424633_at	*Camk1g*	-0.61	0.09	1.67E-05	(62)
1423100_at	*Fos*	-1.77	0.28	1.97E-05	(247)
1460275_at	*Gpr3*	-0.67	0.11	1.06E-04	(249)
1424638_at	*Cdkn1a*	-0.66	0.11	2.37E-04	(277)
1416700_at	*Rnd3*	-0.48	0.09	3.79E-04	(250)

[1]Estimate is the log2(fold change)

Table 4.4 Most significant differentially expressed transcripts between the chronic drug exposure and saline treatment identified by the MetaM meta-analysis and associated with addiction, behavioral or neurological disorders.

Probe ID	Gene Symbol	Estimate[1]	Standard Error	FDR P Value	Reference
1423723_s_at	*Tardbp*	-0.41	0.07	2.72E-04	(259, 278)
1426596_a_at	*Smn1*	0.72	0.12	2.72E-04	(260)
1454051_at	*Fam123a*	0.85	0.13	3.12E-04	(261)
1460561_x_at	*Sepw1*	0.28	0.05	3.34E-04	(251, 252)
1417704_a_at	*Arhgap6*	-0.72	0.13	3.57E-04	(262)
1423146_at	*Hes5*	0.62	0.11	3.57E-04	(263)
1426294_at	*Hapln1*	-0.51	0.09	3.57E-04	(264)
1426615_s_at	*Ndrg4*	0.25	0.04	3.57E-04	(265)
1434736_at	*Hlf*	0.54	0.1	3.57E-04	(266)
1421535_a_at	*Pde4a*	1.01	0.18	3.82E-04	(267)
1419112_at	*Nlk*	0.55	0.1	3.82E-04	(268)
1460365_a_at	*Dnm1*	0.33	0.06	4.93E-04	(279)
1450659_at	*Rgs7*	0.39	0.07	5.45E-04	(253)

[1]Estimate is the log2(fold change)

Table 4.5 Most significant (FDR-adjusted P-value < 0.1) enriched biological processes and molecular functions among the differentially expressed transcripts between acute drug exposure and saline treatment identified by the meta-analysis MetaS.

Category[a]	Term	Count[b]	%[c]	FDR P-value
GOTERM_MF_FAT[d]	GO:0046983~protein dimerization activity	17	7.59	2.35E-03
UP_SEQ_FEATURE[e]	domain:Leucine-zipper	13	5.8	1.07E-05
INTERPRO[f]	IPR004827:Basic-leucine zipper transcription factor	9	4.02	2.03E-04
SMART[g]	SM00338:BRLZ	9	4.02	4.92E-04
GOTERM_MF_FAT	GO:0046983~protein dimerization activity	17	7.59	2.35E-03
INTERPRO	IPR011616:bZIP transcription factor, bZIP-1	6	2.68	2.68E-02
UP_SEQ_FEATURE	DNA-binding region:Basic motif	11	4.91	2.97E-02
INTERPRO	IPR011616:bZIP transcription factor, bZIP-1	6	2.68	2.68E-02

[a]category: original database/resource repository where term orients; [b]count: the number of genes involved in the term; [c]%: percentage of the involved genes/total genes in list; [d]GOTERM_MF_FAT: GO is Gene Ontology; MF is molecular function; FAT is a David created category; [e]UP_SEQ_FEATURE: UP SEQ is a Uniprot sequence feature [f]INTERPRO: database accessible at http://www.ebi.ac.uk/interpro/; [g]SMART: database accessible at http://smart.embl-heidelberg.de/

Table 4.6 Most significant (FDR-adjusted P-value < 0.1) enriched biological processes and molecular functions among the differentially expressed transcripts between chronic drug exposure and saline treatment identified by the meta-analysis MetaM.

Category[a]	Term	Count[b]	%[c]	FDR P-value
SP_PIR_KEYWORDS[d]	ubl conjugation	103	4.7	2.23E-07
SP_PIR_KEYWORDS	isopeptide bond	61	2.79	2.68E-05
GOTERM_MF_FAT[e]	GO:0000166~nucleotide binding	327	14.93	2.02E-10
SP_PIR_KEYWORDS	nucleotide-binding	250	11.42	5.46E-08
GOTERM_MF_FAT	GO:0032553~ribonucleotide binding	262	11.96	2.15E-06
GOTERM_MF_FAT	GO:0032555~purine ribonucleotide binding	262	11.96	2.15E-06
SP_PIR_KEYWORDS	atp-binding	199	9.09	3.84E-06
GOTERM_MF_FAT	GO:0032559~adenyl ribonucleotide binding	217	9.91	1.80E-05
GOTERM_MF_FAT	GO:0005524~ATP binding	214	9.77	2.86E-05
GOTERM_MF_FAT	GO:0017076~purine nucleotide binding	265	12.1	3.11E-05
GOTERM_MF_FAT	GO:0001883~purine nucleoside binding	223	10.18	1.49E-04
GOTERM_MF_FAT	GO:0001882~nucleoside binding	224	10.23	1.62E-04
GOTERM_MF_FAT	GO:0030554~adenyl nucleotide binding	220	10.05	2.83E-04
UP_SEQ_FEATURE[f]	nucleotide phosphate-binding region:ATP	147	6.71	8.16E-03
SP_PIR_KEYWORDS	Kinase	111	5.07	9.47E-03
SP_PIR_KEYWORDS	rna-binding	92	4.2	1.48E-05
INTERPRO[g]	IPR000504:RNA recognition motif, RNP-1	50	2.28	2.85E-05
INTERPRO	IPR012677:Nucleotide-binding, alpha-beta plait	49	2.24	9.24E-05
SMART[h]	SM00360:RRM	50	2.28	4.98E-04
GOTERM_BP_FAT	GO:0000377~RNA splicing, via transesterification	17	0.78	3.84E-04
GOTERM_BP_FAT	GO:0000375~RNA splicing, via transesterification	17	0.78	3.84E-04
GOTERM_BP_FAT	GO:0000398~nuclear mRNA splicing, spliceosome	17	0.78	3.84E-04
GOTERM_BP_FAT	GO:0010629~negative regulation of gene expression	77	3.52	3.25E-04
GOTERM_BP_FAT	GO:0010605~negative regulation of macromolecule	88	4.02	1.29E-03
GOTERM_BP_FAT	GO:0031327~negative regulation of cell biosynthesis	75	3.42	8.54E-03
GOTERM_BP_FAT	GO:0009890~negative regulation of biosynthesis	75	3.42	1.21E-02
GOTERM_BP_FAT	GO:0016071~mRNA metabolic process	63	2.88	1.07E-04
GOTERM_BP_FAT	GO:0006397~mRNA processing	54	2.47	1.58E-03
GOTERM_MF_FAT	GO:0019904~protein domain specific binding	40	1.83	4.48E-02

Table 4.6 (con't)

Category[a]	Term	Count[b]	%[c]	FDR P-value
KEGG_PATHWAY	mmu03010:Ribosome	29	1.32	6.93E-05
GOTERM_BP_FAT	GO:0046907~intracellular transport	76	3.47	4.86E-03
GOTERM_BP_FAT	GO:0008104~protein localization	117	5.34	6.12E-03
GOTERM_BP_FAT	GO:0030182~neuron differentiation	69	3.15	2.88E-02
GOTERM_BP_FAT	GO:0030030~cell projection organization	57	2.6	6.82E-02

[a]category: original database/resource repository where term orients; [b]count: the number of genes involved in the term; [c]%: percentage of the involved genes/total genes in list; [d]SP_PIR_KEYWORDS: the SwissProt Protein Information repository database; [e]GOTERM_MF_FAT: GO is Gene Ontology; MF is molecular function; FAT is a David created category; [f]UP_SEQ_FEATURE: UP SEQ is a Uniprot sequence feature; [g]INTERPRO: database accessible at http://www.ebi.ac.uk/interpro/; [h]SMART: database accessible at http://smart.embl-heidelberg.de/

Table 4.7 Most significant (P-value < 0.05) enriched biological processes and molecular functions among the differentially expressed transcripts common to the chronic and acute drug exposure contrasts relative to saline treatment identified by the MetaS and MetaM meta-analyses.

Category[a]	Term	Count[b]	%[c]	P Value
SP_PIR_KEYWORDS[d]	Spliceosome	5	5.81	0.0013
COG_ONTOLOGY[e]	RNA processing and modification	3	3.49	0.0039
GOTERM_BP_FAT[f]	GO:0009628~response to abiotic stimulus	6	6.98	0.0057

[a]category: original database/resource repository where term orients; [b]count: the number of genes involved in the term; [c]%: percentage of the involved genes/total genes in list; [d]SP_PIR_KEYWORDS: the SwissProt Protein Information repository database; [e]COG_ONTOLOGY: Clusters of Orthologous Groups; [f]GOTERM_BP_FAT: GO is Gene Ontology; BP is; FAT is a David created category

Chapter V: Conclusion

My study of the dysregulation of the transcriptome in diseases such as cancer and drug abuse can be the basis for preventative and therapeutic remedies, along with being a starting point for future work on biomarkers and applied research. The goal of my first project was to integrate statistical and bioinformatics tools to identify miRNA and gene expression profiles that were associated with the survival of patients diagnosed with brain cancer and characterize profiles that are common to all cancer-diagnosed patients and those that are cohort-dependent, where cohort can encompass gender, race, and therapy. This information is valuable in the development of effective personalized therapies. In my second project, I further used the interaction between miRNA, transcription factors and target genes to identify accurate combined biomarkers of ovarian cancer survival. In my last project, I continued the progression towards inferring gene networks by applying systems biology approaches to further mine the transcriptome results, facilitate interpretation, and further augment the understanding of the molecular mechanisms and changes underlying drug abuse.

In my studies, the identification of miRNA and gene profiles associated with cancer survival and drug abuse has three possible benefits. First, the identification of individual transcripts associated in a reliable manner to survival can help in the development of disease biomarkers. Second, functional analysis of differentially expressed transcripts can provide insights into biological processes, molecular functions, and pathways that have a substantial impact on survival. Third, gene network reconstruction based on the correlated expression profiles can augment the understanding of the pathway components and how they are related to different survival outcomes.

In the past, few studies have looked at the simultaneous consideration of transcripts and transcript regulators. It is important to look at all multiple profiles as opposed to only those on an individual basis in order to be able to draw conclusions and determine correctly affected biomarkers and pathways. The relationship between target genes and regulatory microRNAs and transcription factors is important, and future studies should focus on digging deeper into these relationships and how they work together to affect the progression of diseases. The compilation of the information on biomarkers and their relationships on a searchable online database available to the public would be beneficial to the scientific community and public in general.. This would help in the translation of recent insights at the molecular level into treatment strategies, as well as in the possible early detection of diseases.

Depicting known relationships between microRNA, transcription factors, and target genes can help a researcher visualize how these factors all work together to affect the various pathways of the transcriptome. Improvements in the program Cytoscape, such as easier depiction of relationships, would aid in the sharing of vital information to other researchers. Systems biology, which focuses on the interactions within biological systems, is an emerging field of study that can lead to vast improvements in the understanding of disease initiation and progression. Improvements in the sharing of data such as frequently updated online repositories will save time and improve efficiency by allowing researchers to focus more time on actual experiments and interpretation, rather than the time consuming actions of having to perform various internet searchers. Systems biology allows for the integration and analysis of complex data sets and multiple experiments. Using technology to expand upon this will lead to quicker and more accurate studies thus vastly improving our understanding of disease biomarkers and the progression of dysregulation of pathways.

Chapter VI: References

1. Larson DR, Singer RH, Zenklusen D. A single molecule view of gene expression. Trends Cell Biol. 2009 Nov;19(11):630-7.

2. Krebs JE, Goldstein ES, Kilpatrick ST. Lewin's genes X: 10th ed. Sudbury, MA: Jones & Bartlett Publishers; 2009.

3. Montgomery SB, Dermitzakis ET. The resolution of the genetics of gene expression. Hum Mol Genet. 2009 Oct 15;18(R2):R211-5.

4. White RJ, Sharrocks AD. Coordinated control of the gene expression machinery. Trends Genet. 2010 May;26(5):214-20.

5. Hobert O. Gene regulation by transcription factors and microRNAs. Science. 2008 Mar 28;319(5871):1785-6.

6. Yusuf D, Butland SL, Swanson MI, Bolotin E, Ticoll A, Cheung WA, et al. The transcription factor encyclopedia. Genome Biol. 2012;13(3):R24,2012-13-3-r24.

7. Phillips T. Regulation of transcription and gene expression in eukaryotes. Nature Education. 2008;1(1).

8. Novakova J, Slaby O, Vyzula R, Michalek J. MicroRNA involvement in glioblastoma pathogenesis. Biochem Biophys Res Commun. 2009 Aug 14;386(1):1-5.

9. Lee YS, Dutta A. MicroRNAs in cancer. Annu Rev Pathol Mech Dis. 2009;4:199-227.

10. Wang Z. MicroRNA: A matter of life or death. World J Biol Chem. 2010 Apr 26;1(4):41-54.

11. Garzon R, Calin GA, Croce CM. MicroRNAs in cancer. Annu Rev Med. 2009;60:167-79.

12. Lee Y, Jeon K, Lee JT, Kim S, Kim VN. MicroRNA maturation: Stepwise processing and subcellular localization. EMBO J. 2002 Sep 2;21(17):4663-70.

13. Dahiya N, Morin PJ. MicroRNAs in ovarian carcinomas. Endocr Relat Cancer. 2010 Jan 29;17(1):F77-89.

14. Bandyopadhyay S, Mitra R, Maulik U, Zhang MQ. Development of the human cancer microRNA network. Silence. 2010 Feb 2;1(1):6.

15. Trang P, Weidhaas JB, Slack FJ. MicroRNAs as potential cancer therapeutics. Oncogene. 2008 Dec;27 Suppl 2:S52-7.

16. Zhang W, Dahlberg JE, Tam W. MicroRNAs in tumorigenesis: A primer. Am J Pathol. 2007 Sep;171(3):728-38.

17. Zhou X, Kang C, Pu P. MicroRNA and brain tumors. J Clin Oncol. 2007;4:355-9.

18. Zhang H, Li Y, Lai M. The microRNA network and tumor metastasis. Oncogene. 2010 Feb 18;29(7):937-48.

19. Catto JW, Alcaraz A, Bjartell AS, De Vere White R, Evans CP, Fussel S, et al. MicroRNA in prostate, bladder, and kidney cancer: A systematic review. Eur Urol. 2011 May;59(5):671-81.

20. MicroCosm [Internet]. Available from: http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5.

21. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: Tools for microRNA genomics. Nucleic Acids Res. 2008 Jan;36(Database issue):D154-8.

22. Friard O, Re A, Taverna D, De Bortoli M, Cora D. CircuitsDB: A database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse. BMC Bioinformatics. 2010 Aug 23;11:435.

23. Wu W, Sun M, Zou GM, Chen J. MicroRNA and cancer: Current status and prospective. Int J Cancer. 2007 Mar 1;120(5):953-60.

24. Cree IA. Cancer biology. Methods Mol Biol. 2011;731:1-11.

25. Croce CM. Oncogenes and cancer. N Engl J Med. 2008 Jan 31;358(5):502-11.

26. Kopnin BP. Targets of oncogenes and tumor suppressors: Key for understanding basic mechanisms of carcinogenesis. Biochemistry (Mosc). 2000 Jan;65(1):2-27.

27. Sherr CJ. Principles of tumor suppression. Cell. 2004 Jan 23;116(2):235-46.

28. Ozols RF. American cancer society atlas of clinical oncology: Ovarian cancer. Hamilton, London: BC Decker Inc; 2003.

29. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, et al. The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathol. 2007 Aug;114(2):97-109.

30. Furnari FB, Fenton T, Bachoo RM, Mukasa A, Stommel JM, Stegh A, et al. Malignant astrocytic glioma: Genetics, biology, and paths to treatment. Genes Dev. 2007 Nov 1;21(21):2683-710.

31. Kaku T, Ogawa S, Kawano Y, Ohishi Y, Kobayashi H, Hirakawa T, et al. Histological classification of ovarian cancer. Med Electron Microsc. 2003 Mar;36(1):9-17.

32. Khalil I, Brewer MA, Neyarapally T, Runowicz CD. The potential of biologic network models in understanding the etiopathogenesis of ovarian cancer. Gynecol Oncol. 2010 Feb;116(2):282-5.

33. Holland EC. Glioblastoma multiforme: The terminator. Proc Natl Acad Sci U S A. 2000 Jun 6;97(12):6242-4.

34. Krex D, Klink B, Hartmann C, von Deimling A, Pietsch T, Simon M, et al. Long-term survival with glioblastoma multiforme. Brain. 2007 Oct;130(Pt 10):2596-606.

35. Glioblastoma multiforme [Internet].; 2009. Available from: http://emedicine.medscape.com/article/283252-overview.

36. Rao SK, Edwards J, Joshi AD, Siu IM, Riggins GJ. A survey of glioblastoma genomic amplifications and deletions. J Neurooncol. 2010 Jan;96(2):169-79.

37. Lotfi M, Afsharnezhad S, Raziee HR, Ghaffarzadegan K, Sharif S, Shamsara J, et al. Immunohistochemical assessment of MGMT expression and p53 mutation in glioblastoma multiforme. Tumori. 2011 Jan-Feb;97(1):104-8.

38. [Internet]. Available from: http://www.cancer.org/.

39. Hu X, Macdonald DM, Huettner PC, Feng Z, El Naqa IM, Schwarz JK, et al. A miR-200 microRNA cluster as prognostic marker in advanced ovarian cancer. Gynecol Oncol. 2009 Sep;114(3):457-64.

40. Pan Y, Jiao J, Zhou C, Cheng Q, Hu Y, Chen H. Nanog is highly expressed in ovarian serous cystadenocarcinoma and correlated with clinical stage and pathological grade. Pathobiology. 2010;77(6):283-8.

41. Nam EJ, Yoon H, Kim SW, Kim H, Kim YT, Kim JH, et al. MicroRNA expression profiles in serous ovarian carcinoma. Clin Cancer Res. 2008 May 1;14(9):2690-5.

42. Nowak RA. Identification of new therapies for leiomyomas: What in vitro studies can tell us. Clin Obstet Gynecol. 2001 Jun;44(2):327-34.

43. Husseinzadeh N. Status of tumor markers in epithelial ovarian cancer has there been any progress? A review. Gynecol Oncol. 2011 Jan;120(1):152-7.

44. Nicholas MK, Lukas RV, Chmura S, Yamini B, Lesniak M, Pytel P. Molecular heterogeneity in glioblastoma: Therapeutic opportunities and challenges. Semin Oncol. 2011 Apr;38(2):243-53.

45. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PLoS One. 2010 Feb 12;5(2):e8918.

46. Zhou X, Ren Y, Moore L, Mei M, You Y, Xu P, et al. Downregulation of miR-21 inhibits EGFR pathway and suppresses the growth of human glioblastoma cells independent of PTEN status. Lab Invest. 2010 Feb;90(2):144-55.

47. Pedeboscq S, Gravier D, Casadebaig F, Hou G, Gissot A, De Giorgi F, et al. Synthesis and study of antiproliferative activity of novel thienopyrimidines on glioblastoma cells. Eur J Med Chem. 2010 Jun;45(6):2473-9.

48. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061-8.

49. Chow LM, Endersby R, Zhu X, Rankin S, Qu C, Zhang J, et al. Cooperativity within and among pten, p53, and rb pathways induces high-grade astrocytoma in adult brain. Cancer Cell. 2011 Mar 8;19(3):305-16.

50. [Internet]. Available from: http://www.ncbi.nlm.nih.gov/.

51. Ioana M, Angelescu C, Burada F, Mixich F, Riza A, Dumitrescu T, et al. MMR gene expression pattern in sporadic colorectal cancer. J Gastrointestin Liver Dis. 2010 Jun;19(2):155-9.

52. Pothuri B, Leitao MM, Levine DA, Viale A, Olshen AB, Arroyo C, et al. Genetic analysis of the early natural history of epithelial ovarian carcinoma. PLoS One. 2010 Apr 26;5(4):e10358.

53. Pliarchopoulou K, Pectasides D. Epithelial ovarian cancer: Focus on targeted therapy. Crit Rev Oncol Hematol. 2011 Jul;79(1):17-23.

54. Pengetnze Y, Steed M, Roby KF, Terranova PF, Taylor CC. Src tyrosine kinase promotes survival and resistance to chemotherapeutics in a mouse ovarian cancer cell line. Biochem Biophys Res Commun. 2003 Sep 19;309(2):377-83.

55. Choudhry H, Catto JW. Epigenetic regulation of microRNA expression in cancer. Methods Mol Biol. 2011;676:165-84.

56. Meng F, Wehbe-Janek H, Henson R, Smith H, Patel T. Epigenetic regulation of microRNA-370 by interleukin-6 in malignant human cholangiocytes. Oncogene. 2008 Jan 10;27(3):378-86.

57. Drakaki A, Iliopoulos D. MicroRNA gene networks in oncogenesis. Curr Genomics. 2009 Mar;10(1):35-41.

58. Li CY, Mao X, Wei L. Genes and (common) pathways underlying drug addiction. PLoS Comput Biol. 2008 Jan;4(1):e2.

59. Jonkman S, Kenny PJ. Molecular, cellular, and structural mechanisms of cocaine addiction: A key role for microRNAs. Neuropsychopharmacology. 2013 Jan;38(1):198-211.

60. Koob G, Kreek MJ. Stress, dysregulation of drug reward pathways, and the transition to drug dependence. Am J Psychiatry. 2007 Aug;164(8):1149-59.

61. Heiman M, Schaefer A, Gong S, Peterson JD, Day M, Ramsey KE, et al. A translational profiling approach for the molecular characterization of CNS cell types. Cell. 2008 Nov 14;135(4):738-48.

62. Korostynski M, Piechota M, Kaminska D, Solecki W, Przewlocki R. Morphine effects on striatal transcriptome in mice. Genome Biol. 2007;8(6):R128.

63. Piechota M, Korostynski M, Sikora M, Golda S, Dzbek J, Przewlocki R. Common transcriptional effects in the mouse striatum following chronic treatment with heroin and methamphetamine. Genes Brain Behav. 2012 Jun;11(4):404-14.

64. [Internet]. Available from: http://www.unodc.org/documents/data-and-analysis/WDR2012/WDR_2012_web_small.pdf.

65. Wang JC, Kapoor M, Goate AM. The genetics of substance dependence. Annu Rev Genomics Hum Genet. 2012;13:241-61.

66. Bilbao A, Parkitna JR, Engblom D, Perreau-Lenz S, Sanchis-Segura C, Schneider M, et al. Loss of the Ca2+/calmodulin-dependent protein kinase type IV in dopaminoceptive neurons enhances behavioral effects of cocaine. Proc Natl Acad Sci U S A. 2008 Nov 11;105(45):17549-54.

67. Parkitna JR, Bilbao A, Rieker C, Engblom D, Piechota M, Nordheim A, et al. Loss of the serum response factor in the dopamine system leads to hyperactivity. FASEB J. 2010 Jul;24(7):2427-35.

68. Stekel D. Microarray bioinformatics. 2004;93(5):615-6.

69. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG. The affymetrix GeneChip platform: An overview. Methods Enzymol. 2006;410:3-28.

70. Jiang N, Leach LJ, Hu X, Potokina E, Jia T, Druka A, et al. Methods for evaluating gene expression from affymetrix microarray datasets. BMC Bioinformatics. 2008 Jun 17;9:284.

71. Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP. A benchmark for affymetrix GeneChip expression measures. Bioinformatics. 2004 Feb 12;20(3):323-31.

72. Auer H, Newsom DL, Kornacker K. Expression profiling using affymetrix GeneChip microarrays. Methods Mol Biol. 2009;509:35-46.

73. Dufva M. Introduction to microarray technology. Methods Mol Biol. 2009;529:1-22.

74. Moorcroft MJ, Meuleman WR, Latham SG, Nicholls TJ, Egeland RD, Southern EM. In situ oligonucleotide synthesis on poly(dimethylsiloxane): A flexible substrate for microarray fabrication. Nucleic Acids Res. 2005 May 3;33(8):e75.

75. Speed TP, McPeek MS, Evans SN. Robustness of the no-interference model for ordering genetic markers. Proc Natl Acad Sci U S A. 1992 Apr 1;89(7):3103-6.

76. Wu Z, Irizarry R, Gentlemen R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. J Am Stat Assoc. 2004;99(468):909-17.

77. Smith BJ, Ko Y, Southey BR, Rodriguez-Zas SL. BEEHIVE - A suite of tools to manage, analyze and interpret honey bee microarray experiments. May 6-8, 2007; Cold Spring Harbor, NY. ; 2007.

78. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995;Series B (methodological), 57(1):289-300.

79. Kutner MH, Nachtshiem CJ, Neter J. Applied linear regression models. NY: McGraw-Hill Irwin; 2004.

80. Kleinbaum DG. Survival analysis: A self-leaning text. 1st ed. New York: Springer-Verlag; 1996.

81. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: Basic concepts and first analyses. Br J Cancer. 2003 Jul 21;89(2):232-8.

82. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: Multivariate data analysis--an introduction to concepts and methods. Br J Cancer. 2003 Aug 4;89(3):431-6.

83. Adams HA, Southey BR, Robinson GE, Rodriguez-Zas SL. Meta-analysis of genome-wide expression patterns associated with behavioral maturation in honey bees. BMC Genomics. 2008 Oct 24;9:503-18.

84. Rodriguez-Zas SL, Ko Y, Adams HA, Southey BR. Advancing the understanding of the embryo transcriptome co-regulation using meta-, functional, and gene network analysis tools. Reproduction. 2008 Feb;135(2):213-24.

85. Dreyfuss JM, Johnson MD, Park PJ. Meta-analysis of glioblastoma multiforme versus anaplastic astrocytoma identifies robust gene markers. Mol Cancer. 2009 Sep 4;8:71.

86. Miguez FE, Villamil MB, Long SP, Bollero GA. Meta-analysis of the effects of management factors on *Miscanthus*×*giganteus* growth and biomass production. Agri For Meter. 2008;148(8-9):1280-92.

87. Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, et al. Babelomics: An integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W210-3.

88. Piantoni P, Bionaz M, Graugnard DE, Daniels KM, Everts RE, Rodriguez-Zas SL, et al. Functional and gene network analyses of transcriptional signatures characterizing pre-weaned bovine mammary parenchyma or fat pad uncovered novel inter-tissue signaling networks during development. BMC Genomics. 2010 May 26;11:331.

89. Al-Shahrour F, Minguez P, Tarraga J, Montaner D, Alloza E, Vaquerizas JM, et al. BABELOMICS: A systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W472-6.

90. Rodriguez-Zas SL. Functional or pathway analysis and network visualization. Adapted from class notes of ANSC 545 - Statistical Genomics. 2011.

91. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50.

92. Sherman BT, Huang da W, Tan Q, Guo Y, Bour S, Liu D, et al. DAVID knowledgebase: A gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics. 2007 Nov 2;8:426.

93. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27-30.

94. Latendresse M, Karp PD. Web-based metabolic network visualization with a zooming user interface. BMC Bioinformatics. 2011 May 19;12:176.

95. Killcoyne S, Carter GW, Smith J, Boyle J. Cytoscape: A community-based framework for network modeling. Methods Mol Biol. 2009;563:219-39.

96. GPMP plugin [Internet]. Available from: http://apps.cytoscape.org/apps/gpmlplugin.

97. Kelder T, Pico AR, Hanspers K, van Iersel MP, Evelo C, Conklin BR. Mining biological pathways using WikiPathways web services. PLoS One. 2009 Jul 30;4(7):e6447.

98. WikiPathways [Internet]. Available from: http://www.wikipathways.org.

99. BisoGenet [Internet]. Available from: http://bio.cigb.edu.cu/bisogenet-cytoscape/.

100. Wu M, Chan C. Learning transcriptional regulation on a genome scale: A theoretical analysis based on gene expression data. Brief Bioinform. 2011 May 26.

101. Marbach D, Mattiussi C, Floreano D. Combining multiple results of a reverse-engineering algorithm: Application to the DREAM five-gene network challenge. Ann N Y Acad Sci. 2009 Mar;1158:102-13.

102. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006 Mar 20;7 Suppl 1:S7.

103. Brameier M, Wiuf C. Ab initio identification of human microRNAs based on structure motifs. BMC Bioinformatics. 2007 Dec 18;8:478.

104. Chen CZ. MicroRNAs as oncogenes and tumor suppressors. N Engl J Med. 2005 Oct 27;353(17):1768-71.

105. Shalgi R, Brosh R, Oren M, Pilpel Y, Rotter V. Coupling transcriptional and post-transcriptional miRNA regulation in the control of cell fate. Aging (Albany NY). 2009 Sep 8;1(9):762-70.

106. Re A, Cora D, Taverna D, Caselle M. Genome-wide survey of microRNA-transcription factor feed-forward regulatory circuits in human. Mol Biosyst. 2009 Aug;5(8):854-67.

107. Wrensch M, Minn Y, Chew T, Bondy M, Berger MS. Epidemiology of primary brain tumors: Current concepts and review of the literature. Neuro Oncol. 2002 Oct;4(4):278-99.

108. Shi L, Cheng Z, Zhang J, Li R, Zhao P, Fu Z, et al. Hsa-mir-181a and hsa-mir-181b function as tumor suppressors in human glioma cells. Brain Res. 2008 Oct 21;1236:185-93.

109. Chen G, Zhu W, Shi D, Lv L, Zhang C, Liu P, et al. MicroRNA-181a sensitizes human malignant glioma U87MG cells to radiation by targeting bcl-2. Oncol Rep. 2010 Apr;23(4):997-1003.

110. Gaire RK, Bailey J, Bearfoot J, Campbell IG, Stuckey PJ, Haviv I. MIRAGAA--a methodology for finding coordinated effects of microRNA expression changes and genome aberrations in cancer. Bioinformatics. 2010 Jan 15;26(2):161-7.

111. Cui JG, Zhao Y, Sethi P, Li YY, Mahta A, Culicchia F, et al. Micro-RNA-128 (miRNA-128) down-regulation in glioblastoma targets ARP5 (ANGPTL6), bmi-1 and E2F-3a, key regulators of brain cell proliferation. J Neurooncol. 2010 Jul;98(3):297-304.

112. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061-8.

113. TCGA [Internet]. Available from: http://cancergenome.nih.gov/dataportal/.

114. Beehive [Internet]. Available from: http://stagbeetle.animal.uiuc.edu/Beehive1.0.

115. Chuang LY, Yang CS, Li JC, Yang CH. Chaotic genetic algorithm for gene selection and classification problems. OMICS. 2009 Oct;13(5):407-20.

116. Chuang LY, Yang CH, Wu KC, Yang CH. A hybrid feature selection method for DNA microarray data. Comput Biol Med. 2011 Mar 2;41(4):228-37.

117. Dumur CI, Ladd AC, Wright HV, Penberthy LT, Wilkinson DS, Powers CN, et al. Genes involved in radiation therapy response in head and neck cancers. Laryngoscope. 2009 Jan;119(1):91-101.

118. Li J, Tang X. A new classification model with simple decision rule for discovering optimal feature gene pairs. Comput Biol Med. 2007 Nov;37(11):1637-46.

119. Petrausch U, Martus P, Tonnies H, Bechrakis NE, Lenze D, Wansel S, et al. Significance of gene expression analysis in uveal melanoma in comparison to standard risk factors for risk assessment of subsequent metastases. Eye (Lond). 2008 Aug;22(8):997-1007.

120. Oyan AM, Bo TH, Jonassen I, Gjertsen BT, Bruserud O, Kalland KH. cDNA microarray analysis of non-selected cases of acute myeloid leukemia demonstrates distinct clustering independent of cytogenetic aberrations and consistent with morphological signs of differentiation. Int J Oncol. 2006 May;28(5):1065-80.

121. Grutzmann R, Boriss H, Ammerpohl O, Luttges J, Kalthoff H, Schackert HK, et al. Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. Oncogene. 2005 Jul 28;24(32):5079-88.

122. Serão NVL, Delfino KR, Southey BR, Rodriguez-Zas SL. Development of a transcriptomic-based index to prognosticate cancer. In the Abstract Book of the 6th International Symposium on Bioinformatics Research and Applications 2010. 2010.

123. Gene Ontology Consortium. The gene ontology in 2010: Extensions and refinements. Nucleic Acids Res. 2010 Jan;38(Database issue):D331-5.

124. Brennan C, Momota H, Hambardzumyan D, Ozawa T, Tandon A, Pedraza A, et al. Glioblastoma subclasses can be defined by activity among signal transduction pathways and associated genomic alterations. PLoS One. 2009 Nov 13;4(11):e7752.

125. Wu C, Lin J, Hong M, Choudhury Y, Balani P, Leung D, et al. Combinatorial control of suicide gene expression by tissue-specific promoter and microRNA regulation for cancer therapy. Mol Ther. 2009 Dec;17(12):2058-66.

126. Pfeffer S, Zavolan M, Grasser FA, Chien M, Russo JJ, Ju J, et al. Identification of virus-encoded microRNAs. Science. 2004 Apr 30;304(5671):734-6.

127. Taylor DD, Gercel-Taylor C. MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer. Gynecol Oncol. 2008 Jul;110(1):13-21.

128. Svoboda M, Izakovicova Holla L, Sefr R, Vrtkova I, Kocakova I, Tichy B, et al. Micro-RNAs miR125b and miR137 are frequently upregulated in response to capecitabine chemoradiotherapy of rectal cancer. Int J Oncol. 2008 Sep;33(3):541-7.

129. Yeung ML, Yasunaga J, Bennasser Y, Dusetti N, Harris D, Ahmad N, et al. Roles for microRNAs, miR-93 and miR-130b, and tumor protein 53-induced nuclear protein 1 tumor suppressor in cell growth dysregulation by human T-cell lymphotrophic virus 1. Cancer Res. 2008 Nov 1;68(21):8976-85.

130. Miller TE, Ghoshal K, Ramaswamy B, Roy S, Datta J, Shapiro CL, et al. MicroRNA-221/222 confers tamoxifen resistance in breast cancer by targeting p27Kip1. J Biol Chem. 2008 Oct 31;283(44):29897-903.

131. Navon R, Wang H, Steinfeld I, Tsalenko A, Ben-Dor A, Yakhini Z. Novel rank-based statistical methods reveal microRNAs with differential expression in multiple cancer types. PLoS One. 2009 Nov 25;4(11):e8003.

132. Zhu ZL, Zhao ZR, Zhang Y, Yang YH, Wang ZM, Cui DS, et al. Expression and significance of FXYD-3 protein in gastric adenocarcinoma. Dis Markers. 2010;28(2):63-9.

133. Akao Y, Nakagawa Y, Kitade Y, Kinoshita T, Naoe T. Downregulation of microRNAs-143 and -145 in B-cell malignancies. Cancer Sci. 2007 Dec;98(12):1914-20.

134. Chen X, Guo X, Zhang H, Xiang Y, Chen J, Yin Y, et al. Role of miR-143 targeting KRAS in colorectal tumorigenesis. Oncogene. 2009 Mar 12;28(10):1385-92.

135. Ciafre SA, Galardi S, Mangiola A, Ferracin M, Liu CG, Sabatino G, et al. Extensive modulation of a set of microRNAs in primary glioblastoma. Biochem Biophys Res Commun. 2005 Sep 9;334(4):1351-8.

136. Malzkorn B, Wolter M, Liesenberg F, Grzendowski M, Stuhler K, Meyer HE, et al. Identification and functional characterization of microRNAs involved in the malignant progression of gliomas. Brain Pathol. 2010 May;20(3):539-50.

137. Song B, Wang Y, Xi Y, Kudo K, Bruheim S, Botchkina GI, et al. Mechanism of chemoresistance mediated by miR-140 in human osteosarcoma and colon cancer cells. Oncogene. 2009 Nov 19;28(46):4065-74.

138. Huse JT, Brennan C, Hambardzumyan D, Wee B, Pena J, Rouhanifard SH, et al. The PTEN-regulating microRNA miR-26a is amplified in high-grade glioma and facilitates gliomagenesis in vivo. Genes Dev. 2009 Jun 1;23(11):1327-37.

139. Ji J, Shi J, Budhu A, Yu Z, Forgues M, Roessler S, et al. MicroRNA expression, survival, and response to interferon in liver cancer. N Engl J Med. 2009 Oct 8;361(15):1437-47.

140. Jiang L, Mao P, Song L, Wu J, Huang J, Lin C, et al. miR-182 as a prognostic marker for glioma progression and patient survival. Am J Pathol. 2010 Jul;177(1):29-38.

141. Sasayama T, Nishihara M, Kondoh T, Hosoda K, Kohmura E. MicroRNA-10b is overexpressed in malignant glioma and associated with tumor invasive factors, uPAR and RhoC. Int J Cancer. 2009 Sep 15;125(6):1407-13.

142. Zhu JY, Pfuhl T, Motsch N, Barth S, Nicholls J, Grasser F, et al. Identification of novel epstein-barr virus microRNA genes from nasopharyngeal carcinomas. J Virol. 2009 Apr;83(7):3333-41.

143. George GP, Mittal RD. MicroRNAs: Potential biomarkers in cancer. Indian Journal of Clinical Biochemistry. 2010;25:4-14.

144. Zhang C, Wang G, Kang C, Du Y, Pu P. Up-regulation of p27(kip1) by miR-221/222 antisense oligonucleotides enhances the radiosensitivity of U251 glioblastoma. Zhonghua Yi Xue Yi Chuan Xue Za Zhi. 2009 Dec;26(6):634-8.

145. Chan JA, Krichevsky AM, Kosik KS. MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. Cancer Res. 2005 Jul 15;65(14):6029-33.

146. Mukand JA, Blackinton DD, Crincoli MG, Lee JJ, Santos BB. Incidence of neurologic deficits and rehabilitation of patients with brain tumors. Am J Phys Med Rehabil. 2001 May;80(5):346-50.

147. Dorsam RT, Gutkind JS. G-protein-coupled receptors and cancer. Nat Rev Cancer. 2007 Feb;7(2):79-94.

148. Parker SL, Parker MS, Sah R, Sallee F. Angiogenesis and rhodopsin-like receptors: A role for N-terminal acidic residues? Biochem Biophys Res Commun. 2005 Oct 7;335(4):983-92.

149. Chaudhry MA, Sachdeva H, Omaruddin RA. Radiation-induced micro-RNA modulation in glioblastoma cells differing in DNA-repair pathways. DNA Cell Biol. 2010 Sep;29(9):553-61.

150. Li Y, Guessous F, Zhang Y, Dipierro C, Kefas B, Johnson E, et al. MicroRNA-34a inhibits glioblastoma growth by targeting multiple oncogenes. Cancer Res. 2009 Oct 1;69(19):7569-76.

151. Ji Q, Hao X, Meng Y, Zhang M, Desano J, Fan D, et al. Restoration of tumor suppressor miR-34 inhibits human p53-mutant gastric cancer tumorspheres. BMC Cancer. 2008 Sep 21;8:266.

152. Volinia S, Galasso M, Costinean S, Tagliavini L, Gamberoni G, Drusco A, et al. Reprogramming of miRNA networks in cancer and leukemia. Genome Res. 2010 May;20(5):589-99.

153. Mees ST, Mardin WA, Sielker S, Willscher E, Senninger N, Schleicher C, et al. Involvement of CD40 targeting miR-224 and miR-486 on the progression of pancreatic ductal adenocarcinomas. Ann Surg Oncol. 2009 Aug;16(8):2339-50.

154. Dong H, Siu H, Luo L, Fang X, Jin L, Xiong M. Investigating gene and MicroRNA expression in glioblastoma. International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing. 2009:17-21.

155. Guan Y, Mizoguchi M, Yoshimoto K, Hata N, Shono T, Suzuki SO, et al. MiRNA-196 is upregulated in glioblastoma but not in anaplastic astrocytoma and has prognostic significance. Clin Cancer Res. 2010 Aug 15;16(16):4289-97.

156. Stecca B, Ruiz I Altaba A. Context-dependent regulation of the GLI code in cancer by HEDGEHOG and non-HEDGEHOG signals. J Mol Cell Biol. 2010 Apr;2(2):84-95.

157. Chen F, Zhu HH, Zhou LF, Wu SS, Wang J, Chen Z. Inhibition of c-FLIP expression by miR-512-3p contributes to taxol-induced apoptosis in hepatocellular carcinoma cells. Oncol Rep. 2010 May;23(5):1457-62.

158. Rauhala HE, Jalava SE, Isotalo J, Bracken H, Lehmusvaara S, Tammela TL, et al. miR-193b is an epigenetically regulated putative tumor suppressor in prostate cancer. Int J Cancer. 2010 Sep 1;127(6):1363-72.

159. Lodes MJ, Caraballo M, Suciu D, Munro S, Kumar A, Anderson B. Detection of cancer with serum miRNAs on an oligonucleotide microarray. PLoS One. 2009 Jul 14;4(7):e6229.

160. Samlos MA. Identification and functional analysis of micro-rnas encoded by kaposi's sarcoma-associated herpevirus. Case Western Reserve University, Ph.D. thesis. 2007.

161. Bandres E, Cubedo E, Agirre X, Malumbres R, Zarate R, Ramirez N, et al. Identification by real-time PCR of 13 mature microRNAs differentially expressed in colorectal cancer and non-tumoral tissues. Mol Cancer. 2006 Jul 19;5:29.

162. Wyman SK, Parkin RK, Mitchell PS, Fritz BR, O'Briant K, Godwin AK, et al. Repertoire of microRNAs in epithelial ovarian cancer as determined by next generation sequencing of small RNA cDNA libraries. PLoS One. 2009;4(4):e5311.

163. Wurdinger T, Tannous BA. Glioma angiogenesis: Towards novel RNA therapeutics. Cell Adh Migr. 2009 Apr;3(2):230-5.

164. Role of MicroRNAs in the development of cancer stem cells into glioblastomas [Internet].; 2008. Available from: http://realscience.breckschool.org/upper/research/Research2008/Tara.pdf.

165. Skalsky RL, Samols MA, Plaisance KB, Boss IW, Riva A, Lopez MC, et al. Kaposi's sarcoma-associated herpesvirus encodes an ortholog of miR-155. J Virol. 2007 Dec;81(23):12836-45.

166. Yang H, Kong W, He L, Zhao JJ, O'Donnell JD, Wang J, et al. MicroRNA expression profiling in human ovarian cancer: miR-214 induces cell survival and cisplatin resistance by targeting PTEN. Cancer Res. 2008 Jan 15;68(2):425-33.

167. Shih I, Davidson B. Pathogenesis of ovarian cancer: Clues from selected overexpressed genes. Future Oncol. 2009 Dec;5(10):1641-57.

168. Zaret KS, Carroll JS. Pioneer transcription factors: Establishing competence for gene expression. Genes Dev. 2011 Nov 1;25(21):2227-41.

169. von Gruenigen VE, Huang HQ, Gil KM, Gibbons HE, Monk BJ, Rose PG, et al. A comparison of quality-of-life domains and clinical factors in ovarian cancer patients: A gynecologic oncology group study. J Pain Symptom Manage. 2010 May;39(5):839-46.

170. Delfino KR, Serao NV, Southey BR, Rodriguez-Zas SL. Therapy-, gender- and race-specific microRNA markers, target genes and networks related to glioblastoma recurrence and survival. Cancer Genomics Proteomics. 2011 Jul-Aug;8(4):173-83.

171. Serao NV, Delfino KR, Southey BR, Beever JE, Rodriguez-Zas SL. Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival. BMC Med Genomics. 2011 Jun 7;4:49.

172. Delfino KR, Southey BR, Sweedler JV, Rodriguez-Zas SL. Genome-wide census and expression profiling of chicken neuropeptide and prohormone convertase genes. Neuropeptides. 2010 Feb;44(1):31-44.

173. Cox proportional-hazards regression for survival data: Web appendix to*An R and S-PLUS companion to applied regression*. [Internet]. Available from: http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf.

174. Schaubel DE, Wei G. Fitting semiparametric additive hazards models using standard statistical software. Biom J. 2007 Aug;49(5):719-30.

175. Le Bechec A, Portales-Casamar E, Vetter G, Moes M, Zindy PJ, Saumet A, et al. MIR@NT@N: A framework integrating transcription factors, microRNAs and their targets to identify sub-network motifs in a meta-regulation network model. BMC Bioinformatics. 2011 Mar 4;12:67.

176. Gene ontology [Internet]. Available from: http://www.geneontology.org/.

177. KEGG [Internet]. Available from: http://www.genome.jp/kegg/.

178. Dragon database of ovarian cancer genes [Internet]. Available from: http://apps.sanbi.ac.za/ddoc/.

179. Martinez I, Gardiner AS, Board KF, Monzon FA, Edwards RP, Khan SA. Human papillomavirus type 16 reduces the expression of microRNA-218 in cervical carcinoma cells. Oncogene. 2008 Apr 17;27(18):2575-82.

180. Miles GD, Seiler M, Rodriguez L, Rajagopal G, Bhanot G. Identifying microRNA/mRNA dysregulations in ovarian cancer. BMC Res Notes. 2012 Mar 27;5:164.

181. Zhou X, Zhao F, Wang ZN, Song YX, Chang H, Chiang Y, et al. Altered expression of miR-152 and miR-148a in ovarian cancer is related to cell proliferation. Oncol Rep. 2012 Feb;27(2):447-54.

182. Eitan R, Kushnir M, Lithwick-Yanai G, David MB, Hoshen M, Glezerman M, et al. Tumor microRNA expression patterns associated with resistance to platinum based chemotherapy and survival in ovarian cancer patients. Gynecol Oncol. 2009 Aug;114(2):253-9.

183. Yang N, Kaur S, Volinia S, Greshock J, Lassus H, Hasegawa K, et al. MicroRNA microarray identifies let-7i as a novel biomarker and therapeutic target in human epithelial ovarian cancer. Cancer Res. 2008 Dec 15;68(24):10307-14.

184. Boren T, Xiong Y, Hakam A, Wenham R, Apte S, Chan G, et al. MicroRNAs and their target messenger RNAs associated with ovarian cancer response to chemotherapy. Gynecol Oncol. 2009 May;113(2):249-55.

185. Dahiya N, Sherman-Baust CA, Wang TL, Davidson B, Shih I, Zhang Y, et al. MicroRNA expression and identification of putative miRNA targets in ovarian cancer. PLoS One. 2008 Jun 18;3(6):e2436.

186. Laios A, O'Toole S, Flavin R, Martin C, Kelly L, Ring M, et al. Potential role of miR-9 and miR-223 in recurrent ovarian cancer. Mol Cancer. 2008 Apr 28;7:35.

187. Zhang L, Huang J, Yang N, Greshock J, Megraw MS, Giannakakis A, et al. microRNAs exhibit high frequency genomic alterations in human cancer. Proc Natl Acad Sci U S A. 2006 Jun 13;103(24):9136-41.

188. Wang W, Peng B, Wang D, Ma X, Jiang D, Zhao J, et al. Human tumor microRNA signatures derived from large-scale oligonucleotide microarray datasets. Int J Cancer. 2011 Oct 1;129(7):1624-34.

189. Marchini S, Cavalieri D, Fruscio R, Calura E, Garavaglia D, Nerini IF, et al. Association between miR-200c and the survival of patients with stage I epithelial ovarian cancer: A retrospective study of two independent tumour tissue collections. Lancet Oncol. 2011 Mar;12(3):273-85.

190. Yamada Y, Enokida H, Kojima S, Kawakami K, Chiyomaru T, Tatarano S, et al. MiR-96 and miR-183 detection in urine serve as potential tumor markers of urothelial carcinoma:

Correlation with stage and grade, and comparison with urinary cytology. Cancer Sci. 2011 Mar;102(3):522-9.

191. Baffa R, Fassan M, Volinia S, O'Hara B, Liu CG, Palazzo JP, et al. MicroRNA expression profiling of human metastatic cancers identifies cancer gene targets. J Pathol. 2009 Oct;219(2):214-21.

192. Kobayashi D, Yamada M, Kamagata C, Kaneko R, Tsuji N, Nakamura M, et al. Overexpression of early growth response-1 as a metastasis-regulatory factor in gastric cancer. Anticancer Res. 2002 Nov-Dec;22(6C):3963-70.

193. Unoki M, Nakamura Y. EGR2 induces apoptosis in various cancer cell lines by direct transactivation of BNIP3L and BAK. Oncogene. 2003 Apr 10;22(14):2172-85.

194. Mahner S, Baasch C, Schwarz J, Hein S, Wolber L, Janicke F, et al. C-fos expression is a molecular predictor of progression and survival in epithelial ovarian carcinoma. Br J Cancer. 2008 Oct 21;99(8):1269-75.

195. Tanaka Y, Kobayashi H, Suzuki M, Kanayama N, Terao T. Transforming growth factor-beta1-dependent urokinase up-regulation and promotion of invasion are involved in src-MAPK-dependent signaling in human ovarian cancer cells. J Biol Chem. 2004 Mar 5;279(10):8567-76.

196. Sunde JS, Donninger H, Wu K, Johnson ME, Pestell RG, Rose GS, et al. Expression profiling identifies altered expression of genes that contribute to the inhibition of transforming growth factor-beta signaling in ovarian cancer. Cancer Res. 2006 Sep 1;66(17):8404-12.

197. Do TV, Kubba LA, Du H, Sturgis CD, Woodruff TK. Transforming growth factor-beta1, transforming growth factor-beta2, and transforming growth factor-beta3 enhance ovarian cancer metastatic potential by inducing a Smad3-dependent epithelial-to-mesenchymal transition. Mol Cancer Res. 2008 May;6(5):695-705.

198. Inan S, Vatansever S, Celik-Ozenci C, Sanci M, Dicle N, Demir R. Immunolocalizations of VEGF, its receptors flt-1, KDR and TGF-beta's in epithelial ovarian tumors. Histol Histopathol. 2006 Oct;21(10):1055-64.

199. Guan Y, Kuo WL, Stilwell JL, Takano H, Lapuk AV, Fridlyand J, et al. Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. Clin Cancer Res. 2007 Oct 1;13(19):5745-55.

200. Wisman GB, Hollema H, Helder MN, Knol AJ, Van der Meer GT, Krans M, et al. Telomerase in relation to expression of p53, c-myc and estrogen receptor in ovarian tumours. Int J Oncol. 2003 Nov;23(5):1451-9.

201. Suzuki F, Akahira J, Miura I, Suzuki T, Ito K, Hayashi S, et al. Loss of estrogen receptor beta isoform expression and its correlation with aberrant DNA methylation of the 5'-untranslated region in human epithelial ovarian carcinoma. Cancer Sci. 2008 Dec;99(12):2365-72.

202. Li AJ, Baldwin RL, Karlan BY. Estrogen and progesterone receptor subtype expression in normal and malignant ovarian epithelial cell cultures. Am J Obstet Gynecol. 2003 Jul;189(1):22-7.

203. Lurie G, Wilkens LR, Thompson PJ, McDuffie KE, Carney ME, Terada KY, et al. Genetic polymorphisms in the estrogen receptor beta (ESR2) gene and the risk of epithelial ovarian carcinoma. Cancer Causes Control. 2009 Feb;20(1):47-55.

204. Heckman CA, Duan H, Garcia PB, Boxer LM. Oct transcription factors mediate t(14;18) lymphoma cell survival by directly regulating bcl-2 expression. Oncogene. 2006 Feb 9;25(6):888-98.

205. Xu D, Dwyer J, Li H, Duan W, Liu JP. Ets2 maintains hTERT gene expression and breast cancer cell proliferation by interacting with c-myc. J Biol Chem. 2008 Aug 29;283(35):23567-80.

206. Al-azawi D, Ilroy MM, Kelly G, Redmond AM, Bane FT, Cocchiglia S, et al. Ets-2 and p160 proteins collaborate to regulate c-myc in endocrine resistant breast cancer. Oncogene. 2008 May 8;27(21):3021-31.

207. Lourenco D, Brauner R, Lin L, De Perdigo A, Weryha G, Muresan M, et al. Mutations in NR5A1 associated with ovarian insufficiency. N Engl J Med. 2009 Mar 19;360(12):1200-10.

208. Dodson H. "Women with variants in "CLOCK" gene have higher risk of breast cancer". Yale Office of Public Affairs and Communications [Internet]:April 2011. Available from: http://opac.yale.edu/news/article.aspx?id=7261.

209. Hoffman AE, Yi CH, Zheng T, Stevens RG, Leaderer D, Zhang Y, et al. CLOCK in breast tumorigenesis: Genetic, epigenetic, and transcriptional profiling analyses. Cancer Res. 2010 Feb 15;70(4):1459-68.

210. Nakagawa M, Oda Y, Eguchi T, Aishima S, Yao T, Hosoi F, et al. Expression profile of class I histone deacetylases in human cancer tissues. Oncol Rep. 2007 Oct;18(4):769-74.

211. Ohuchida K, Mizumoto K, Lin C, Yamaguchi H, Ohtsuka T, Sato N, et al. MicroRNA-10a is overexpressed in human pancreatic cancer and involved in its invasiveness partially via suppression of the HOXA1 gene. Ann Surg Oncol. 2012 Jul;19(7):2394-402.

212. Filippova GN, Qi CF, Ulmer JE, Moore JM, Ward MD, Hu YJ, et al. Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter tts DNA-binding specificity. Cancer Res. 2002 Jan 1;62(1):48-52.

213. Docquier F, Farrar D, D'Arcy V, Chernukhin I, Robinson AF, Loukinov D, et al. Heightened expression of CTCF in breast cancer cells is associated with resistance to apoptosis. Cancer Res. 2005 Jun 15;65(12):5112-22.

214. Hiranuma C, Kawakami K, Oyama K, Ota N, Omura K, Watanabe G. Hypermethylation of the MYOD1 gene is a novel prognostic factor in patients with colorectal cancer. Int J Mol Med. 2004 Mar;13(3):413-7.

215. Muller HM, Fiegl H, Widschwendter A, Widschwendter M. Prognostic DNA methylation marker in serum of cancer patients. Ann N Y Acad Sci. 2004 Jun;1022:44-9.

216. Azuma T, Seki N, Yoshikawa T, Saito T, Masuho Y, Muramatsu M. cDNA cloning, tissue expression, and chromosome mapping of human homolog of SOX18. J Hum Genet. 2000;45(3):192-5.

217. Young N, Hahn CN, Poh A, Dong C, Wilhelm D, Olsson J, et al. Effect of disrupted SOX18 transcription factor function on tumor growth, vascularization, and endothelial development. J Natl Cancer Inst. 2006 Aug 2;98(15):1060-7.

218. Johnson SA, Dubeau L, Kawalek M, Dervan A, Schonthal AH, Dang CV, et al. Increased expression of TATA-binding protein, the central transcription factor, can contribute to oncogenesis. Mol Cell Biol. 2003 May;23(9):3043-51.

219. Chen X, Thiaville MM, Chen L, Stoeck A, Xuan J, Gao M, et al. Defining NOTCH3 target genes in ovarian cancer. Cancer Res. 2012 May 1;72(9):2294-303.

220. Zoref-Shani E, Lavie R, Bromberg Y, Beery E, Sidi Y, Sperling O, et al. Effects of differentiation-inducing agents on purine nucleotide metabolism in an ovarian cancer cell line. J Cancer Res Clin Oncol. 1994;120(12):717-22.

221. Gao Z, Xu X, McClane B, Zeng Q, Litkouhi B, Welch WR, et al. C terminus of clostridium perfringens enterotoxin downregulates CLDN4 and sensitizes ovarian cancer cells to taxol and carboplatin. Clin Cancer Res. 2011 Mar 1;17(5):1065-74.

222. Iorio MV, Visone R, Di Leva G, Donati V, Petrocca F, Casalini P, et al. MicroRNA signatures in human ovarian cancer. Cancer Res. 2007 Sep 15;67(18):8699-707.

223. Zhang L, Volinia S, Bonome T, Calin GA, Greshock J, Yang N, et al. Genomic and epigenetic alterations deregulate microRNA expression in human epithelial ovarian cancer. Proc Natl Acad Sci U S A. 2008 May 13;105(19):7004-9.

224. Eitan R, Kushnir M, Lithwick-Yanai G, David MB, Hoshen M, Glezerman M, et al. Tumor microRNA expression patterns associated with resistance to platinum based chemotherapy and survival in ovarian cancer patients. Gynecol Oncol. 2009 Aug;114(2):253-9.

225. Wyman SK, Parkin RK, Mitchell PS, Fritz BR, O'Briant K, Godwin AK, et al. Repertoire of microRNAs in epithelial ovarian cancer as determined by next generation sequencing of small RNA cDNA libraries. PLoS One. 2009;4(4):e5311.

226. Unoki M, Nakamura Y. Growth-suppressive effects of BPOZ and EGR2, two genes involved in the PTEN signaling pathway. Oncogene. 2001 Jul 27;20(33):4457-65.

227. Mayr D, Hirschmann A, Marlow S, Horvath C, Diebold J. Analysis of selected oncogenes (AKT1, FOS, BCL2L2, TGFbeta) on chromosome 14 in granulosa cell tumors (GCTs): A comprehensive study on 30 GCTs combining comparative genomic hybridization (CGH) and fluorescence-in situ-hybridization (FISH). Pathol Res Pract. 2008;204(11):823-30.

228. Chen CH, Shen J, Lee WJ, Chow SN. Overexpression of cyclin D1 and c-myc gene products in human primary epithelial ovarian cancer. Int J Gynecol Cancer. 2005 Sep-Oct;15(5):878-83.

229. Carlson NR. Physiology of behavior 11th edition. Boston: Pearson; 2013.

230. Gorwood P, Le Strat Y, Ramoz N, Dubertret C, Moalic JM, Simonneau M. Genetics of dopamine receptors and drug addiction. Hum Genet. 2012 Jun;131(6):803-22.

231. Gene expression omnibus (GEO) repository [Internet]. Available from: http://ncbi.nlm.nih.gov/geo.

232. Lemberger T, Parkitna JR, Chai M, Schutz G, Engblom D. CREB has a context-dependent role in activity-regulated transcription and maintains neuronal cholesterol homeostasis. FASEB J. 2008 Aug;22(8):2872-9.

233. Affymetrix [Internet].; 2012. Available from: http://www.affymetrix.com.

234. Delfino KR, Southey BR, Sweedler JV, Rodriguez-Zas SL. Genome-wide census and expression profiling of chicken neuropeptide and prohormone convertase genes. Neuropeptides. 2010 Feb;44(1):31-44.

235. Delfino KR, Serao NV, Southey BR, Rodriguez-Zas SL. Therapy-, gender- and race-specific microRNA markers, target genes and networks related to glioblastoma recurrence and survival. Cancer Genomics Proteomics. 2011 Jul-Aug;8(4):173-83.

236. Delfino KR, Rodriguez-Zas SL. Transcription factor-microRNA-target gene networks associated with ovarian cancer survival and recurrence. PLoS One. 2013;8(3):e58608.

237. Serao NV, Delfino KR, Southey BR, Beever JE, Rodriguez-Zas SL. Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival. BMC Med Genomics. 2011 Jun 7;4:49.

238. Porter KI, Southey BR, Sweedler JV, Rodriguez-Zas SL. First survey and functional annotation of prohormone and convertase genes in the pig. BMC Genomics. 2012 Nov 15;13:582.

239. Ko Y, Zhai C, Rodriguez-Zas SL. Discovery of gene network variability across samples representing multiple classes. Int J Bioinform Res Appl. 2010;6(4):402-17.

240. Ko Y, Zhai C, Rodriguez-Zas S. Inference of gene pathways using mixture bayesian networks. BMC Syst Biol. 2009 May 19;3:54.

241. Adams HA, Southey BR, Everts RE, Marjani SL, Tian CX, Lewin HA, et al. Transferase activity function and system development process are critical in cattle embryo development. Funct Integr Genomics. 2011 Mar;11(1):139-50.

242. SAS version 9.3 [Internet]. Available from: http://support.sas.com/documentation.

243. DAVID functional annotation cluserting [Internet]. Available from: http://david.abcc.ncifcrf.gov/manuscripts/fuzzy_cluster/.

244. Piechota M, Korostynski M, Solecki W, Gieryk A, Slezak M, Bilecki W, et al. The dissection of transcriptional modules regulated by various drugs of abuse in the mouse striatum. Genome Biol. 2010;11(5):R48,2010-11-5-r48. Epub 2010 May 4.

245. Grueter BA, Robison AJ, Neve RL, Nestler EJ, Malenka RC. FosB differentially modulates nucleus accumbens direct and indirect pathway function. Proc Natl Acad Sci U S A. 2013 Jan 29;110(5):1923-8.

246. Peakman MC, Colby C, Perrotti LI, Tekumalla P, Carle T, Ulery P, et al. Inducible, brain region-specific expression of a dominant negative mutant of c-jun in transgenic mice decreases sensitivity to cocaine. Brain Res. 2003 Apr 25;970(1-2):73-86.

247. Ziolkowska B, Korostynski M, Piechota M, Kubik J, Przewlocki R. Effects of morphine on immediate-early gene expression in the striatum of C57BL/6J and DBA/2J mice. Pharmacol Rep. 2012 Sep;64(5):1091-104.

248. Caffino L, Racagni G, Fumagalli F. Stress and cocaine interact to modulate arc/Arg3.1 expression in rat brain. Psychopharmacology (Berl). 2011 Nov;218(1):241-8.

249. Tourino C, Valjent E, Ruiz-Medina J, Herve D, Ledent C, Valverde O. The orphan receptor GPR3 modulates the early phases of cocaine reinforcement. Br J Pharmacol. 2012 Oct;167(4):892-904.

250. Marie-Claire C, Salzmann J, David A, Courtin C, Canestrelli C, Noble F. Rnd family genes are differentially regulated by 3,4-methylenedioxymethamphetamine and cocaine acute treatment in mice brain. Brain Res. 2007 Feb 23;1134(1):12-7.

251. Chen J, Berry MJ. Selenium and selenoproteins in the brain and brain diseases. J Neurochem. 2003 Jul;86(1):1-12.

252. Rodd ZA, Kimpel MW, Edenberg HJ, Bell RL, Strother WN, McClintick JN, et al. Differential gene expression in the nucleus accumbens with ethanol self-administration in inbred alcohol-preferring rats. Pharmacol Biochem Behav. 2008 Jun;89(4):481-98.

253. Anderson GR, Cao Y, Davidson S, Truong HV, Pravetoni M, Thomas MJ, et al. R7BP complexes with RGS9-2 and RGS7 in the striatum differentially control motor learning and locomotor responses to cocaine. Neuropsychopharmacology. 2010 Mar;35(4):1040-50.

254. Wu H, Sun YE. Reversing DNA methylation: New insights from neuronal activity-induced Gadd45b in adult neurogenesis. Sci Signal. 2009 Mar 31;2(64):pe17.

255. Naegele J. Epilepsy and the plastic mind. Epilepsy Curr. 2009 Nov-Dec;9(6):166-9.

256. Wang KS, Liu X, Aragam N, Jian X, Mullersman JE, Liu Y, et al. Family-based association analysis of alcohol dependence in the COGA sample and replication in the australian twin-family study. J Neural Transm. 2011 Sep;118(9):1293-9.

257. Biswas SC, Zhang Y, Iyirhiaro G, Willett RT, Rodriguez Gonzalez Y, Cregan SP, et al. Sertad1 plays an essential role in developmental and pathological neuron death. J Neurosci. 2010 Mar 17;30(11):3973-82.

258. Horita H, Wada K, Rivas MV, Hara E, Jarvis ED. The dusp1 immediate early gene is regulated by natural stimuli predominantly in sensory input neurons. J Comp Neurol. 2010 Jul 15;518(14):2873-901.

259. Lillo P, Hodges JR. Cognition and behaviour in motor neurone disease. Curr Opin Neurol. 2010 Dec;23(6):638-42.

260. Fulceri F, Bartalucci A, Paparelli S, Pasquali L, Biagioni F, Ferrucci M, et al. Motor neuron pathology and behavioral alterations at late stages in a SMA mouse model. Brain Res. 2012 Mar 9;1442:66-75.

261. Comai G, Boutet A, Neirijnck Y, Schedl A. Expression patterns of the wtx/amer gene family during mouse embryonic development. Dev Dyn. 2010 Jun;239(6):1867-78.

262. Tondreau T, Dejeneffe M, Meuleman N, Stamatopoulos B, Delforge A, Martiat P, et al. Gene expression pattern of functional neuronal cells derived from human bone marrow mesenchymal stromal cells. BMC Genomics. 2008 Apr 11;9:166.

263. Lee CT, Ma YL, Lee EH. Serum- and glucocorticoid-inducible kinase1 enhances contextual fear memory formation through down-regulation of the expression of Hes5. J Neurochem. 2007 Mar;100(6):1531-42.

264. Carulli D, Pizzorusso T, Kwok JC, Putignano E, Poli A, Forostyak S, et al. Animals lacking link protein have attenuated perineuronal nets and persistent plasticity. Brain. 2010 Aug;133(Pt 8):2331-47.

265. Yamamoto H, Kokame K, Okuda T, Nakajo Y, Yanamoto H, Miyata T. NDRG4 protein-deficient mice exhibit spatial learning deficits and vulnerabilities to cerebral ischemia. J Biol Chem. 2011 Jul 22;286(29):26158-65.

266. Hitzler JK, Soares HD, Drolet DW, Inaba T, O'Connel S, MG R, et al. Expression patterns of the hepatic leukemia factor gene in the nervous system of developing and adult mice. Brain Res. 1999 Feb 27;820(1-2):1-11.

267. Kuroiwa M, Snyder GL, Shuto T, Fukuda A, Yanagawa Y, Benavides DR, et al. Phosphodiesterase 4 inhibition enhances the dopamine D1 receptor/PKA/DARPP-32 signaling cascade in frontal cortex. Psychopharmacology (Berl). 2012 Feb;219(4):1065-79.

268. Romagnoli A, Oliverio S, Evangelisti C, Iannicola C, Ippolito G, Piacentini M. Neuroleukin inhibition sensitises neuronal cells to caspase-dependent apoptosis. Biochem Biophys Res Commun. 2003 Mar 14;302(3):448-53.

269. Reyes BA, Vakharia K, Ferraro TN, Levenson R, Berrettini WH, Van Bockstaele EJ. Opiate agonist-induced re-distribution of wntless, a mu-opioid receptor interacting protein, in rat striatal neurons. Exp Neurol. 2012 Jan;233(1):205-13.

270. Moratalla R, Robertson HA, Graybiel AM. Dynamic regulation of NGFI-A (zif268, egr1) gene expression in the striatum. J Neurosci. 1992 Jul;12(7):2609-22.

271. Moratalla R, Vickers EA, Robertson HA, Cochran BH, Graybiel AM. Coordinate expression of c-fos and jun B is induced in the rat striatum by cocaine. J Neurosci. 1993 Feb;13(2):423-33.

272. Moratalla R, Xu M, Tonegawa S, Graybiel AM. Cellular responses to psychomotor stimulant and neuroleptic drugs are abnormal in mice lacking the D1 dopamine receptor. Proc Natl Acad Sci U S A. 1996 Dec 10;93(25):14928-33.

273. Kerscher O, Felberbaum R, Hochstrasser M. Modification of proteins by ubiquitin and ubiquitin-like proteins. Annu Rev Cell Dev Biol. 2006;22:159-80.

274. Ciechanover A, Iwai K. The ubiquitin system: From basic mechanisms to the patient bed. IUBMB Life. 2004 Apr;56(4):193-201.

275. Singh P, Mohammad F, Sharma A. Transcriptomic analysis in a drosophila model identifies previously implicated and novel pathways in the therapeutic mechanism in neuropsychiatric disorders. Front Neurosci. 2011 Mar 31;5:161.

276. Jordan KW, Carbone MA, Yamamoto A, Morgan TJ, Mackay TF. Quantitative genomics of locomotor behavior in drosophila melanogaster. Genome Biol. 2007;8(8):R172.

277. Vucetic Z, Totoki K, Schoch H, Whitaker KW, Hill-Smith T, Lucki I, et al. Early life protein restriction alters dopamine circuitry. Neuroscience. 2010 Jun 30;168(2):359-70.

278. Kabashi E, Lin L, Tradewell ML, Dion PA, Bercier V, Bourgouin P, et al. Gain and loss of function of ALS-related mutations of TARDBP (TDP-43) cause motor deficits in vivo. Hum Mol Genet. 2010 Feb 15;19(4):671-83.

279. Huang W, Li MD. Nicotine modulates expression of miR-140*, which targets the 3'-untranslated region of dynamin 1 gene (Dnm1). Int J Neuropsychopharmacol. 2009 May;12(4):537-46.

Appendix

Description of supplementary file: Delfino, KR_SupplementaryTables.xlsx

Table S1. Significant differentially expressed transcripts between the acute drug exposure and saline treatment identified by the MetaS meta-analysis and associated with addiction, behavioral or neurological disorders. Includes Probe ID, Sequence ID, Gene Symbol, Estimate or log2(fold change), standard error, p-value, and FDR p-value.

Table S2. Significant differentially expressed transcripts between the chronic drug exposure and saline treatment identified by the MetaM meta-analysis and associated with addiction, behavioral or neurological disorders. Includes Probe ID, Sequence ID, Gene Symbol, Estimate or log2(fold change), standard error, p-value, and FDR p-value.

Table S3. Significant (FDR-adjusted P-value < 0.1) enriched biological processes and molecular functions among the differentially expressed transcripts between acute drug exposure and saline treatment identified by the meta-analysis MetaS. Includes enrichments score, original database repository, number of genes involved in each term, percentage of the involved genes/total genes in the list, the actual genes in the biological processes and molecular functions, p-value, and FDR p-value.

Table S4. Significant (FDR-adjusted P-value < 0.1) enriched biological processes and molecular functions among the differentially expressed transcripts between chronic drug exposure and saline treatment identified by the meta-analysis MetaM. Includes enrichments score, original database repository, number of genes involved in each term, percentage of the involved genes/total genes in the list, the actual genes in the biological processes and molecular functions, p-value, and FDR p-value.

Table S5. Significant (P-value < 0.05) enriched biological processes and molecular functions among the differentially expressed transcripts common to the chronic and acute drug exposure contrasts relative to saline treatment identified by the MetaS and MetaM meta-analyses. Includes enrichments score, original database repository, number of genes involved in each term, percentage of the involved genes/total genes in the list, the actual genes in the biological processes and molecular functions, p-value, and FDR p-value.