Exploring microRNA Biology using Integrative Bioinformatics

By
Yotsawat Pomyen

A thesis submitted in fulfilment of requirements for the degree of Doctor of Philosophy of Imperial College London

Computational and Systems Medicine
Department of Surgery and Cancer
Imperial College London
2014

Copyright Declaration

Statement of originality

I hereby declare that all the works in this thesis is the product of my own original work, and to the best of my knowledge the material in this thesis has not been submitted to any other academic institution for any degree. The works of others used in this thesis are fully acknowledged according to academic standard practices.

Abstract

Deregulation of energy metabolism is one of the emerging hallmarks of cancer required for proliferation and metastasis. MicroRNAs are small RNA molecules that have crucial roles in the regulation of biological processes in organisms, including metabolism. Due to recent discovery of miRNAs in humans, roles of miRNAs in metabolism of tumour cells, and effects these have on cancer patients, are still obscure and in need of expansion. Currently, experimental and computational data on the miRNAs are being analysed by a wide range of statistical methods; however, these methods in their original forms posses many limitations. Therefore, new ways of utilising these statistical methods are needed in order to unravel the roles of miRNAs in cancer metabolism. In this thesis, the roles of a specific miRNA, miR-22, and the three metabolic target genes were investigated through the use of classical statistical methods, revealed that miR-22, the metabolic target genes, and the interactions between them, were beneficial to survival outcome of breast cancer patients. Furthermore, novel combinations of the conventional statistical methods were invented in order to investigate the global miRNA regulations on metabolic target genes. These new procedures were demonstrated by using publicly available data sets. In one analysis, it was found that miRNAs could be divided into six clusters according to the metabolic target genes through a novel combination of statistical methods. A new statistical method was also invented to provide a generalised means to test for clustering based on sets of correlations.

Acknowledgement

First of all, I would like to thank my supervisors Dr. Hector Keun and Dr. Timothy Ebbels who gave me the opportunity to work with them. I am grateful for their advice over the course of my entire PhD study. I would also like to thank Dr. Toby Athersuch who introduced me to the section of Biomolecular Medicine (now the Computational and Systems Biology) and led me to Dr. Keun's group. I am also grateful for the discussion and advice from Dr. Rachel Cavill at the initial stage of my PhD.

I would then like to thank the members and ex-members of Keun's group for being helpful during my PhD. In particular, I am grateful to Costas, Katya, Gabriel, Chung-Ho, Shyam, and Tianlai for discussion sessions that were useful to my work presented in this thesis.

I am forever grateful to the Thai Government that fully supported all the expenses and tuition for my study, and the opportunity to explore the world through my study. I would also like to thank all my good friends whom I have the chance to meet and doing interesting activities together. I also forever indebted to my parents who always support me in everyway possible to make this journey possible. Finally, I am thankful for the love, patience and support of Fha in every step of my PhD life.

TABLE OF CONTENTS

List of figures

List of tables

List of Abbreviations

Abbreviation	Meaning
2D-DIGE	2 Dimensional differentiation in-gel electrophoresis
5'RLM-RACE	5' RNA ligase mediated-rapid amplification of cDNA ends
ACLY	ATP citrate lyase
AGO	Argonaute
CI	Confidence interval
CLASH	Crosslinking, immunoprecipitation and sequencing of hybrids
CLIP	Crosslinking and immunoprecipitation
DGCR8	DiGeorge Syndrome Critical Region 8
eIF6	eukaryotic Initiation factor 6
ELOVL6	Elongase 6
ER-A	Estrogen receptor alpha
FDR	False discovery rate
GCRMA	GC-content, Robust multichip Average
GMUCT	Genome-wide mapping of uncapped transcripts
GO	Gene ontology
HER2	human epidermal growth factor receptor 2
HITS-CLIP	High-throughput sequencing of RNA isolated by CLIP
HMM	Hidden-Markov model
HR	Hazard ratio
iCLIP	individual-nucleotide resolution CLIP
KEGG	Kyoto encyclopedia of genes and genomes
LNA	Locked nucleic acid
MIR22HG	miR-22 host gene

Abbreviation	**Meaning**
miRNA	microRNA
mRNA	messenger RNA
MTA	5’-S-methyl-5’-thioadenosine
MTHFD2	Methylenetetrahydrofolate dehydrogenase/cyclohydrolase
NCI	National Cancer Institute
NGS	Next generation sequencing
ORA	Over-representation analysis
ORCA	Over-representation of correlation analysis
PACT	Protein activator of the interferon
PAR-CLIP	Photoactivatable ribonucleoside-enhanced CLIP
PARE	Parallel analysis of RNA ends
P_{ct}	Probability of preferentially conserved targeting
PR	Progesterone receptor
qRT-PCR	Quantitative real-time polymerase chain reaction
RECON1	human metabolic network reconstruction 1
RISC	RNA-induced silencing complex
RPKM	Read per kilobase per million mapped reads
SAM	S-(5’-adenosyl)-L-methioine
SILAC	Stable isotope labelling with amino acids in cell culture
TCGA	The cancer genome atlas
TET2	Ten eleven translocation 2
TMM	Trimmed-mean of M values
TNBC	Triple negative breast cancer
TRBP	TAR RNA-binding protein

Abbreviation	Meaning
UTR	Untranslated region
XPO5	Exportin 5

Chapter 1 : Introduction

1.1 Dysregulated metabolism is one of the established hallmarks of cancer

In 2000, Hanahan & Weinberg published a seminal work on the hallmarks of cancer, which are the six main characteristics that cancer cells have in common (Hanahan & Weinberg 2000). Eleven years later, the same authors published an updated article, including four more hallmarks that cancer cells would have to acquire in order to support proliferative capacity. One of those four hallmarks is the capability of the cancer to modify or reprogram the energy metabolism of the cell (Hanahan & Weinberg 2011). It is widely known that cancer cells have very different requirements compared to normal cells in terms of metabolism. Otto Warburg was the first to observe that cancer cells have abnormal energy metabolism (Koppenol et al. 2011I). It has been proposed that cancer cells use a "glycolytic phenotype" or aerobic glycolysis, in which the cancer cells largely choose for an inefficient way to produce energy from glucose even under sufficient oxygen supply, because the cells can divert intermediate metabolites from the glycolysis to generate other essential biomolecules for the new daughter cells (Heiden et al. 2009). Main components required in any cells that can be synthesized from glycolysis intermediates include nucleotides, amino acids, and lipids. Since energy metabolism mainly occurs in mitochondria, biological pathways within the mitochondria will be altered to accommodate the modified metabolite requirement for highly proliferative cells such as cancer cells (Ward & Thompson 2012; Seyfried et al. 2014).

1.2 c-Myc is an oncogene heavily involved in cancer cell metabolism

Transcription factor c-Myc is a proto-oncogene that plays a very important role in cell malignancy, especially in modifying cell metabolism (Cairns et al. 2011; Miller et al.

2012). c-Myc can activate or inhibit a wide range of genes that are involved in cell metabolism, including energy metabolism, nucleotide biosynthesis, lipid synthesis and mitochondrial biosynthesis (Dang 2013). A myriad of genes that control these biological pathways are under the control of this transcription factor, which means that slight change in the expression of c-Myc can produce huge metabolic consequences in the cells, leading to carcinogenesis. However, protein-coding genes are not the only type of genes that are under the regulation of c-Myc. Recently, several studies have identified another class of genes that is also under the sphere of influence of this transcription factor, which are non-coding microRNA genes or miRNAs (Chen et al. 2012; Psathas & Thomas-Tikhonenko 2014). Currently, several miRNAs have been identified to be under c-Myc control. There have been substantial evidence that c-Myc regulates processes, through miRNAs, that are crucial for carcinogenesis, including cell cycle progression, apoptosis, angiogenesis, metastasis, and metabolism (Psathas & Thomas-Tikhonenko 2014). Therefore, by identifying microRNAs under this transcription factor, which in turn regulate metabolic enzymes that have vital roles in cancer, a more complete picture of how cellular energetics works could be unraveled.

1.3 miRNA roles in cancer metabolism

Since the discovery of the first miRNA in *C. elegans* in 1993, thousands of miRNAs were identified, through various means, in all kinds of organisms, ranging from virus to animals and plants. After the discovery of the let-7 miRNA family, which has homologs in humans, an explosion of miRNA-related findings was soon took place in the biomedical research community, including cancer research. Several miRNAs have been identified to play various roles in all sorts of cancer hallmarks, ranging from tumour development (Malumbres 2013), invasion and metastasis (Garzon et al. 2009;

Melo & Esteller 2011), angiogenesis (Melo & Esteller 2011), and proliferation (Esquela-Kerscher & Slack 2006). An emerging cancer hallmark of reprogramming energy metabolism also has several miRNAs that were identified to have important roles (Chen et al. 2012; Dumortier et al. 2013; Tomasetti et al. 2014). Metabolic pathways that have been identified to be affected by miRNA dysregulation are glycolysis, TCA cycle, lipid metabolism, amino acid metabolism, most of them are taking place in mitochondria (Tomasetti et al. 2014; Dumortier et al. 2013). Therefore, the understanding of miRNA in cancer metabolism is far from complete and is also becoming more complex with the growing number of newly discovered metabolic targets of miRNAs (Rottiers & Näär 2012).

1.4 What is microRNA?

microRNAs (miRNAs) are a class of short, non-coding RNA molecules, approximately 20 – 30 nucleotides in length, that regulate gene expression in eukaryotes from plants to worms to humans. The first microRNA, *lin-4*, was discovered in *Caenorhabditis elegans* in 1993. The expression of the gene *lin-4* is required in *C. elegans* larvae to progress from early development stage of L1 to late stage L1. It was found that antisense binding of *lin-4* in 3’UTR of *lin-14* could regulate the expression of *lin-14*, a gene that is highly expressed during the early L1 developmental stage L1 of worm larvae, so that it can complete specific growth pattern in the late L1 stage and then progress to L2 stage (Lee et al. 1993). It was also independently confirmed that the repression of *lin-14* occurred post-transcriptionally, and only the binding in 3’UTR is sufficient for the repression (Wightman et al. 1993). Although *lin-4* was the first miRNA to be discovered, the sequence of *lin-4* was not conserved cross other species. It was not until 2000 that the *let-7* miRNA was discovered in *C. elegans* (Reinhart et al. 2000), and was also found that the sequence of *let-7* was conserved in other metazoan, from flies, mouse, and humans (Pasquinelli et al. 2000). The orthologs of *lin-4,* miR-125, were found later in mouse (Lagos-Quintana et al. 2002) and human (Sempere et al. 2004).

1.5 Genomic location of microRNA

There are two main classes of microRNAs divided according to their genomic location, which are intergenic and intragenic microRNAs. An intergenic microRNA is a microRNA that has an independent transcription unit without a host gene. Intergenic microRNA can be subdivided into monocistronic with a single primary transcript of microRNA such as miR-210 (Zhang et al. 2009), bicistronic, i.e. two adjacent primary transcripts of microRNA such as miR-222/211 cluster (Di Leva et al. 2010), and polycistronic with multiple primary transcripts of microRNAs in the same transcription unit such as miR-17/92 cluster with six different microRNAs (Mogilyansky & Rigoutsos 2013)

An intragenic microRNA is a microRNA that is located in the intron or exon of another gene. The host gene of an intragenic microRNA can be either a noncoding or coding gene. Therefore, there are four possible genomic locations that intragenic microRNAs can be located, which are: intronic of a noncoding gene such as miR-15a/16-1 in *DLEU* gene (Veronese et al. 2014); intronic of a coding gene such as miR-26a in *CTDSP2* gene (Zhu et al. 2012); exonic of a noncoding gene such as miR-155 in *BIC* gene (Elton et al. 2013); and exonic of a coding gene such as miR-985 in *CACNG8* gene (Kim et al. 2009);

A study had calculated number of intragenic miRNAs to be approximately 48% (~43% intronic and ~5% exonic miRNAs), the rest 52% are intergenic miRNAs (Hinske et al. 2010).

1.6 Biogenesis of microRNA in animals

Typically, biogenesis of microRNAs usually begins by transcription from genomic DNA to generate primary miRNA sequence, or pri-miRNA (Bartel 2004; Kim et al. 2009). The canonical pathway for miRNA biogenesis (**Figure 1.1A**) is accomplished through RNA polymerase II in the nucleus (Ameres & Zamore 2013), although some miRNAs can also be transcribed by RNA polymerase III (Borchert et al. 2006). The length of pri-miRNA can be range from several hundred base pairs to several thousands base pairs, which forms hairpin-structured stem loops from the transcript (Bartel 2004; Kim 2005). In the canonical pathway of miRNA biogenesis, heterodimer of endonuclease RNAse III Drosha and DGCR8 (DiGeorge Syndrome critical region 8) called the microprocessor complex then crops the primary transcript of miRNA at the stem loops, releasing one or more precursor miRNAs or pre-miRNAs according to pri-miRNA (Ameres & Zamore 2013). In an alternative miRNA biogenesis pathway (**Figure 1.1B**), a hairpin structure resembles pre-miRNA, which called a mirtron can be produced from host gene of miRNA through alternative splicing by spliceosome, by passing the need for microprocessor complex (Carthew & Sontheimer 2009; Kim et al. 2009). Next, the hairpin structure of pre-miRNA or mirtron, now 60-100 nucleotides, is exported out of the nucleus to cytoplasm by a complex of nucleo/cytoplasmic transporter exportin-5 (XPO5) and Ran-GTP (He & Hannon 2004; Ha & Kim 2014). After the transport out of the nucleus, pre-miRNA will then be further cleaved at the stem loop structure by another complex of endonuclease Dicer, TAR RNA-binding protein (TRBP) and protein activator of the interferon (PACT, also know PRKRA), and argonaute protein (AGO), resulting in a double-stranded miRNA-miRNA* duplex approximately 22 nucleotides (Kim et al. 2009; Ameres & Zamore 2013; Ha & Kim 2014). The miRNA-miRNA* duplex is then loaded onto AGO protein, which is followed by the liberation of miRNA* strand, and forms mature RNA-induced silencing complex (RISC), with mature miRNA sequence as the guide strand for mRNA target recognition (Kim et al. 2009; Ha & Kim 2014). Some small pre-miRNAs can bypass the dicing process, in which the pre-miRNA is directly loaded onto the AGO protein and undergo maturation process into a mature RISC (Kim et al. 2009; Ameres & Zamore 2013; Ha & Kim 2014).

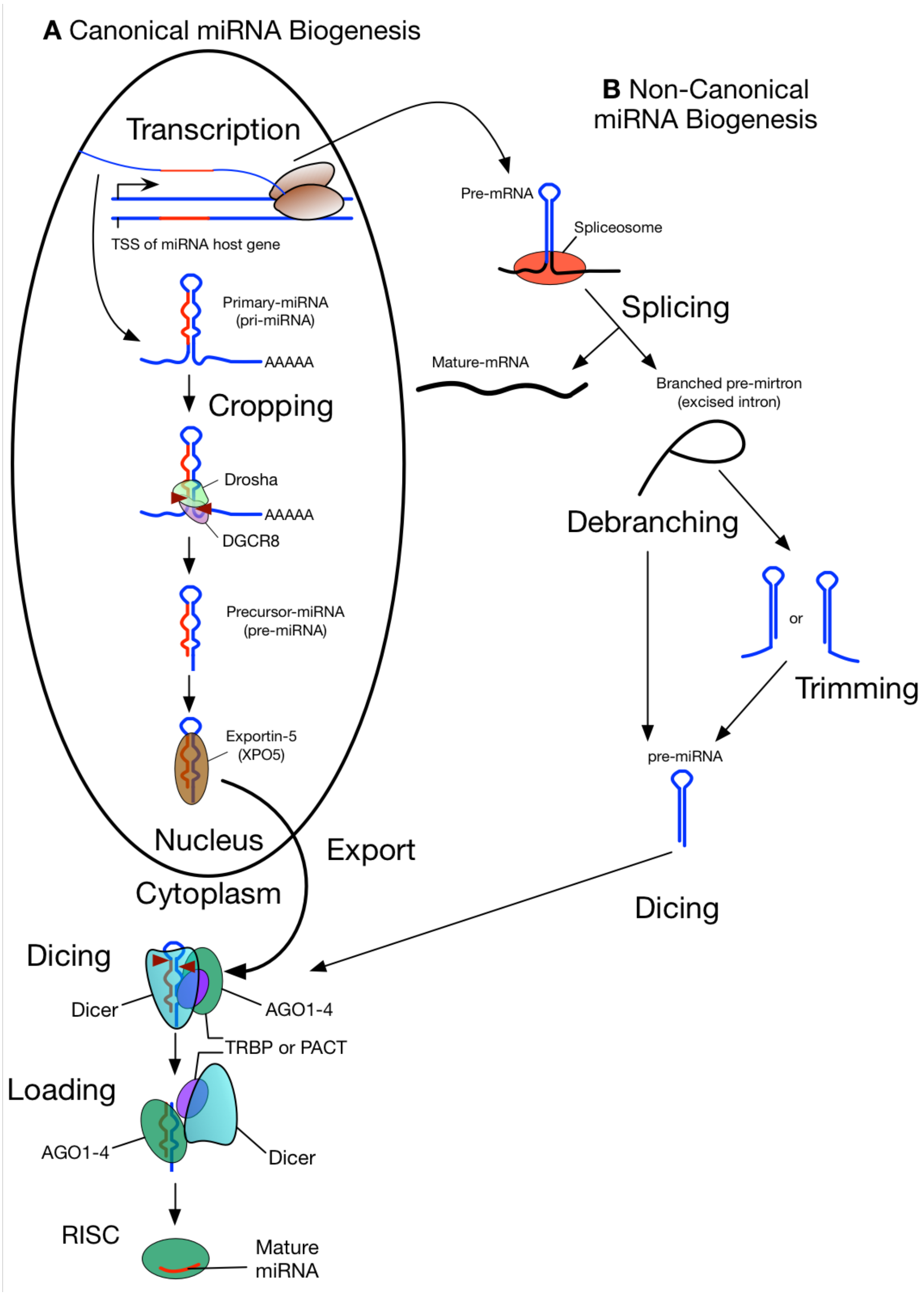


Figure 1.1 miRNA biogenesis pathways. A) Canonical miRNA biogenesis pathway. miRNA biogenesis starts with RNA transcription to produce primary miRNA molecule (pri-miRNA), then the secondary structure (hairpin loop) is then cropped by Drosha and DGCR8 proteins to produce precursor miRNA (pre-miRNA). The pre-miRNA is then exported out of the nucleus by Exportin-5 (XPO5) protein. After the export, pre-miRNA is further processed by Dicer protein. The mature miRNA sequence is finally loaded onto AGO protein, resulting in RNA-induced silencing complex (RISC). This figure is adapted from Kim et al. 2009.

1.7 Mechanisms of actions of miRNA

Several groups have been trying to ascertain how miRNA regulate the target genes, but there is no unifying theory of how miRNA suppress the expression of the target mRNAs (He & Hannon 2004; Pillai et al. 2007; Liu 2008; Sun et al. 2010; Eulalio et al. 2008; Ameres & Zamore 2013; Sarkies & Miska 2014). A recent study on kinetic signatures of mechanisms of actions of miRNA has gathered nine different mechanisms, and proposed a unifying model of how gene expression repression is carried out by a miRNA (Morozova et al. 2012). These nine mechanisms of actions of miRNA are:

1. Cap-40S inhibition – inhibition of translational initiation by cap-40S association with miRISC
2. 60S joining inhibition – inhibition of translational initiation by recruiting eukaryotic initiation factor 6 (eIF6) protein to prevent association between 40S-60S ribosome subunits
3. Inhibition of elongation
4. Ribosome drop-off (premature termination)
5. Cotranslational protein degradation
6. Sequestration in P-bodies
7. mRNA decay (degradation or destabilisation)
8. mRNA cleavage
9. Transcriptional inhibition (miRNA-mediated chromatin reorganisation following by gene silencing)

1.8 Experimental methods for miRNA target identification

Since the first miRNA was identified back in 1993, there has been numerous advances on how the targets of miRNA are identified. Currently, these methods are widely used to experimentally identify (or validate) the miRNA-target interaction:

1. Genetic screening – identification of mutated genes that rescue miRNA loss-of-function phenotype
2. Gene expression profiling after miRNA transfection/inhibition – comparing differentially expression genes after overexpression or inhibition of a certain miRNA
3. Translation (or polysome) profiling after miRNA transfection/inhibition – mRNAs undergoing elongation process by ribosomes are trapped by cyclohexamide, and then separated out by sucrose-gradient centrifugation.
4. Immunoprecipitation-based methods – mRNA transcripts are crosslinked with the miRISC proteins (typically AGO or TNRC6) by UV light then the crosslinked molecules are separated out by immunoprecipitation (crosslinking and immunoprecipitation or CLIP). The transcripts are then identified by microarray or sequencing. This category can be subdivided into four different methods:
    a. High-throughput sequencing of RNA isolated by CLIP (HITS-CLIP), where the transcripts obtained from immunoprecipitation undergo high-throughput sequencing (or deep sequencing or next-generation sequencing).
    b. Photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP), where photoactive 4-thiouridine is used as RNA substrate in RNA synthesis step.

   c. Crosslinking, immunoprecipitation and sequencing of hybrids (CLASH), where CLIP process is followed by ligation miRNA to the target sites within the miRISC.
   d. Individual-nucleotide resolution CLIP (iCLIP), where reverse transcriptase is used to block the polymerization that takes place before the crosslinking.

5. Biotin-tagged pull-down – miRNA is tagged with biotin at the 3' end and the miRNA-mRNA complex is then captured by streptavidin beads, which are then purified and the mRNA sequences analysed.
6. Proteomic-based methods – protein profiles are determined after miRNA transfection/inhibition. This category can be subdivided into two methods:
   a. Stable isotope labelling with amino acids in cell culture (SILAC), where protein profiles are measured by high-throughput mass spectrometry in the two samples (normal vs miRNA transfected/inhibited) that are labelled by different isotopes.
   b. Two-dimensional differentiation in-gel electrophoresis (2D-DIGE), where two samples (normal vs miRNA transfected/inhibited) are labelled with different fluorescent dyes then undergo iso-electric focusing and SDS-PAGE, and the protein spots are finally identified by mass spectrometry.
7. Direct target cleavage detection – high-throughput sequencing is performed globally on the product of a modified 5' RNA ligase mediated-rapid amplification of cDNA ends (5' RLM-RACE). This method is called parallel analysis of RNA ends (PARE) or degradom-seq or genome-wide mapping of uncapped transcripts (GMUCT).

These method were extensively reviewed by Thomson et al., 2011; Martinez-Sanchez & Murphy, 2013; and Hausser & Zavolan, 2014.

1.9 Several features used in miRNA target prediction algorithms

Before the advances on the experimental methods of target identification reviewed in the previous section, there was no economically viable means to globally identify the targets of miRNA. Researchers were resorting to a bottom-up approach, where only a handful of genes or miRNAs can be tested at a time. Therefore, bioinformatics algorithms and tools based-on unique characteristics (or features) of miRNA target recognition sites and/or miRNA-mRNA complex were invented to facilitate miRNA target identification. These algorithms and tools are immensely valuable to biomedical researchers because the researchers can use these tools to focus their work on likely candidates, and do not waste their precious time and resources to unnecessarily testing genes that might not be target of the miRNAs of their choice.

Several miRNA target prediction algorithms are online resources, and most of them are freely available. There are a number of reviews that introduce each algorithm regarding their main capabilities, methodologies, performances, and the differences between them (Hammell 2010; Saito & Saetrom 2010; Witkos et al. 2011; Peterson et al. 2014; Ritchie & Rasko 2014). These reviews mostly focused on the 'features' that the algorithms used in order to predict the target genes of miRNAs. Based on these reviews, common features that miRNA prediction software used to identify target genes are as follows:

1. Seed sequence complementarity – most of miRNA target prediction software require the target sites to have at least 6 or 7 nucleotides 'seed' matched to the first 2-7 or 2-8 nucleotides starting at 5' end of mature miRNA sequence.

2. Conservation among species – another feature that most miRNA target prediction algorithms used to assess potential target sites. If the target sites are conserved among several species, then they are considered evolutionarily important to living organisms.
3. miRNA:mRNA duplex thermodynamic stability or free energy – if a miRNA:mRNA duplex has low minimum (Gibbs) free energy, it is considered to be highly stable (or favourable) duplex, and thus more likely to be a real interaction.
4. Target site accessibility – mRNA secondary structure affects the binding of the miRISC to the target site. In this regard, minimum free energy can be calculated to assess whether the target site is accessible, and indeed a functional target site.
5. Target site abundance – to achieve effective targeting, a mRNA usually has more than one miRNA target sites for a miRNA. Some groups suggested that multiple target sites might result in additive or even synergistic effects.
6. miRNA:mRNA interaction types – certain types of miRNA:mRNA interactions are more likely to be used. Nevertheless, there exist several non-canonical site types for miRNA:mRNA interactions. These site types are:
    a. 6-mer: a site with perfect Watson-Crick match between 2-7 nucleotide seed sequence at 5’ end of miRNA. This site only has modest regulatory effect on the target.
    b. 7mer-A1: a site with perfect Watson-Crick match between 2-7 nucleotide seed sequence at 5’ end of miRNA and has an A across from nucleotide 1 of miRNA. It was proposed that the A is recognised by miRISC.

c. 7mer-m8: a site with perfect Watson-Crick match between 2-8 nucleotide seed sequence at 5' end of miRNA. This is the most abundance site types in the genomes and considered to be highly conserved among species.

d. 3'-supplementary: a site with additional Watson-Crick base-pairing at nucleotides 13-16 from 5' end of miRNA. This site type is rare in the genome.

e. 3'-compensatory: a site that has a mismatch in the seed region (nucleotides 2-7) with additional Watson-Crick base-paring at nucleotides 13-16 from 5' end of miRNA, which thought to compensate the mismatch in the seed region. This site type is rare in the genome.

7. Target site location – target sites of miRNAs are usually residing in 3' untranslated region (UTR), but recently there are emerging evidence that miRNA target sites can be in 5'UTR, or even on the open reading frames (ORFs) of the target genes.
8. miRNA:mRNA expression profiles – the target genes and miRNAs usually have negative correlation between them. With this fact, some algorithms take into account the expression profiles of the miRNAs and the target mRNAs in order to increase the accuracy of the predictions. It should be noted that this is not a main feature, and only a few algorithms incorporated this approach into the calculation.

1.10 Commonly used miRNA prediction algorithms

As previously mentioned in the previous section, there are quite a number of miRNA target prediction software, implementing different methodologies and take into account different number of features in the prediction methods. However, there are only a few that have been used extensively by researchers and worth mentioning in this study. These miRNA prediction software are as follow:

1. TargetScan – a prediction algorithm based on stringent seed pairing and target site conservation (Friedman et al. 2009).
2. miRanda – this tool considers three main features in the target predictions, which are stringent seed match, site conservation and minimum free energy (John et al. 2004).
3. DIANA-microT – one of several tools that consider a bulge in the seed match and require 7-9 nucleotides long to supplement or compensate the bulge (Reczko et al. 2012).
4. PITA – Probability of interaction by target accessibility (PITA) uses the site accessibility as the main feature in the target prediction. This tool also considers other features in the prediction (Kertesz et al. 2007).
5. PicTar – This algorithm is the first one to include the co-expression between miRNA and the target genes. The potential target sites are identified through sequence alignment and scored by using Hidden-Markov Model (HMM) (Krek et al. 2005).
6. RNAhybrid – Uses the minimum free energy between miRNAs (small RNAs) and mRNAs (large RNAs) duplex as the main feature. The first algorithm that use statistical models to assess the features that the algorithm takes into account, which are seed match, minimum free energy and target site abundance (Rehmsmeier et al. 2004).

7. RNA22 – Uses pattern recognition technique to search for potential target sites in mRNAs and then identify the miRNAs that could form duplex with the target sites from minimum free energy (Miranda et al. 2006).

There was a comprehensive review comparing these target prediction algorithms on the basis of precision, sensitivity, and the ratio between predicted targets and experimentally validated targets, and showed that these target prediction software have comparable performance (Alexiou et al. 2009).

1.11 Emerging miRNA that might play a significant role in cancer: miR-22

MiR-22 is an exonic miRNA, which is transcribed from a host gene called *MIR22HG* residing on the short arm of chromosome 17 in human, and has the genomic location at 1,617,197-1,617,281 according to UCSC Genome Browser human genome version hg19 (Karolchik et al. 2014). MiR-22 was previously found to be under the positive feedback loop of transcription factor c-Myc (Chang et al. 2008; J Xiong et al. 2010; Polioudakis et al. 2013), but another study showed that c-Myc might be a negative regulator of miR-22 (Kong et al. 2014). This miRNA has been shown to have potentially protective effects in neurodegenerative disease (Jovicic et al. 2013) and a cardiomyocyte hypertrophy regulator in heart tissue (Huang et al. 2013). Recent studies showed that the role of miR-22 in cancer is controversial, with conflicting evidence that miR-22 is a tumour-suppressor gene (J Xiong et al. 2010; Jianhua Xiong et al. 2010; Zhang et al. 2010; Zhang et al. 2012; X. Li et al. 2014) and an oncogene (Song, Ito, et al. 2013; Song, Poliseno, et al. 2013). Several hallmarks in multiple cancer types identified to be under the regulation of miR-22 are cancer cell invasion and metastasis (Li et al. 2010; Guo et al. 2013), apoptosis , cancer stem cell (Song, Ito, et al. 2013), cell cycle progression (Ting et al. 2010; Ling et al. 2012; D.

Xu et al. 2011), and cancer cell proliferation (Lenkala et al. 2014; Polioudakis et al. 2013). However, another hallmark that is currently under explored for miR-22 is cancer cell metabolism.

1.12 Breast cancer and miR-22: conflicting evidence

The role of miR-22 in several types of cancer is controversial as has been shown in the previous section, especially in breast cancer. In early work, miR-22 was suggested to be a tumour-suppressor gene in breast cancer. However, several recent studies yielded different results. Two of the very first breast cancer-related studies on miR-22 were that miR-22 repressed the expression of Estrogen receptor α (ERα), and thereby inhibited proliferation of breast cancer cell line MCF-7SH and in patient samples (Pandey & Picard 2009; Jianhua Xiong et al. 2010). Nevertheless, another study showed that miR-22 was not statistically associated with ERα expression in breast cancer tissues from patients (Yoshimoto et al. 2011). In another study, miR-22 was found to be suppressing the growth and cell invasiveness of breast cancer, both in *in vitro* and *in vivo* models through reprogramming of the senescent pathway by directly targeting *SIRT1*, *Sp1*, and *CDK6* genes (D. Xu et al. 2011). Another recently published study showed that *Sp1* was also suppressing the promoter region of miR-22. Taking into account the study by Xu et al. (2011), miR-22 is in a feed back loop control with *Sp1* and the transcription factor c-Myc (Kong et al. 2014). Another c-Myc-related finding in the miR-22 breast cancer study was that miR-22 downregulates c-Myc binding protein (*MYCBP*), and therefore suppressing the activity of c-Myc by depletion of its binding protein partner (J Xiong et al. 2010). Although it was found in another cancer cell line, c-Myc was found to be directly promoting miR-22 expression in cervical cancer cell line (Polioudakis et al. 2013). MiR-22 was also found to be repressing other two oncogenes, epidermal growth factor receptor 3 (*ERBB3* or *HER3*) and ecotropic viral integration site 1 (*EVI-1*) in five different breast cancer cell lines, which in turn reduce the activity of PI3K/AKT signalling pathway and also reduce the metastatic potential of those breast cancer cell lines (Patel et al. 2011). Another twist in the breast cancer-related story of miR-22 came in 2013, in which a study showed that miR-22 promotes cancer stemness and metastasis (through EMT) through chromatin remodeling via a demethylation of another tumour-suppressor gene miR-200 in the *in vivo* models (Song, Poliseno, et al. 2013). There is no consensus reached for the exact roles of miR-22 in breast cancer. As for the roles of miR-22 in emerging hallmark like energy metabolism, there is no study linking miR-22 to any metabolic target. Therefore, this is a very interesting aspect of miR-22 that should be explored in order to fill the gap for the role of miR-22 in cancer metabolism.

1.13 The overall extent of miRNA regulation for the metabolic target genes is still unknown

Despite the growing number of metabolic genes being discovered to be direct targets of several miRNAs and multiple miRNA predicted target databases, there are very few studies trying to probe the extent of metabolic consequences from miRNA control. The reasons for the lack of this kind of study are as follow:

1. The real target genes of miRNAs and exact roles of miRNAs on the metabolic target genes are not completely understood. This renders most of the attempts to model the interactions between miRNAs and the (predicted or experimentally verified) target genes less useful than it should be.
2. The multiple levels of interactions between metabolic genes, proteins, metabolites, and the miRNAs involved are so complex that there is no

computational or statistical approaches that can accommodate all of the variables at once.

3. Even using only two layers of data, either miRNA-mRNA, or miRNA-metabolites combinations, the models generated will still be very complex and comprise of many assumptions and constraints, render the usefulness of the model to be limited or very specific to fixed conditions (Resendis-Antonio et al. 2015).

Nevertheless, a few studies have been attempting to perform a global analysis of miRNA-metabolic association. One study by Tibiche & Wang (2008) was trying to find regulatory patterns of miRNAs on metabolic enzymes by mapping the miRNAs onto metabolic pathways according to miRNA predicted target genes. This study identified the ‘Hub nodes’ that are highly connected and ‘cut points’ that are the rate-limiting points for metabolite flows, and found that these hub nodes and cut points were under the control of miRNAs more than expected by chance. Although the approach was a very simple one, one should be cautious of the result of the study because of the fact that the most of the predicted targets of miRNAs might be false positives. Another study was using network-based Bayesian model called Gaussian Graphical Model (GGM) to model the correlation between miRNA, mRNA and patient survival time (Wang et al. 2013). One drawback of this study was that by using correlation between miRNA and mRNA, true targets of miRNAs that might not be correlated due to systematic errors of the data would be missed in the analysis. Another drawback of this approach is that it is a complex method and cumbersome to implement.

Therefore, a new statistical approach that is simpler to implement and able to reduce the chance of false positives in the pool of predicted target genes, and a novel combination of classical statistical methods to verify the findings from such approach is needed in order to better understand the interaction bewteen the miRNAs and target genes under the regulation of miRNAs, especially the metabolic target genes.

1.14 Hypothesis of the thesis

A better understanding of the biology of miRNAs can be achieved through integration of two types of biological data by using novel combinations of conventional statistical methods.

1.15 The aims of this thesis

1. Ascertain the role of miR-22 and the metabolic target genes under miR-22 regulation in survival outcome of breast cancer patients
2. Explore the global network of miRNA regulation on metabolic target genes by using statistical approach
3. Create a novel statistical procedure for data integration from combination of two classical statistical methods.

1.16 Statistical methods used in this thesis

Several classical statistical methods are used throughout this thesis in order to answer specific and broader research questions that were explored about the miRNA biology and metabolism. The implementation of the statistical methods in this thesis were ranging from purely exploratory, i.e. simply use the methods to answer a specific question, to a method development, i.e. inventing new ways to use classical statistical methods to unravel novel information from old data sets.

1.16.1 Spearman's Rank correlation

Rank correlation is a nonparametric statistical dependence test between two variables. This type of test does not assume that the variables are normally distributed. The Spearman's rank correlation in particular can be considered a nonparametric version of Pearson's product moment correlation (or Pearson correlation) (Kendall & Gibbons 1990). The main difference between these two tests is that the Pearson correlation probes the linear dependence of the two quantitative variables, while the Spearman's rank correlation probes the dependence of the two variables using the ranks between data in the two variables. Spearman's rank correlation is essentially Pearson correlation on the ranks of the two variables. In this thesis, the Spearman's rank correlation was used to determine the degree of association between two variables in order to avoid making assumptions on the distribution of the data. This method was utilised extensively in the Chapter 2 and 4. The Spearman Rho (ρ) can be calculated by the following equation:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where d_i is the difference between ranks of the two variables, and n is the number of samples in a variable.

1.16.2 Kaplan-Meier estimate and Log-rank test

When collecting the survival data of multiple patients, it is difficult to start data collection at the same exact time for all the patients. It is impossible for the researcher to be able to trace all the patients to the very end of the study period, especially in the study of disease such as cancer, where patients might be out of the study, either by death or just disappear from the study. Kaplan-Meier estimate is a method that calculates the probabilities of occurrence of an event, e.g. death, cancer relapse, metastasis, etc., at a certain point in time (Kleinbaum & Klein 2005a). Each probability at one point in time is multiplied to each and everyone of the probability together to yield the final estimate. There are three assumptions on the data, which are 1) a patient who is censored (i.e. lost-to-follow up) has the same survival prospect as other patients at a specified time point, 2) survival probabilities of every patient is the same, either recruited early or late in the study, and 3) event happened at the time specified. A survival probability at any time point is calculated by the following equation:

$$S_t = \frac{\#\, subjects\ alive\ at\ the\ start - \#\, subjects\ died\ at\ the\ time\ point}{\#\, subjects\ alive\ at\ the\ start}$$

To determine if the two or more KM estimates are statistically different, Log-Rank test is used to test the hypothesis that the KM estimates between groups are not different. Log-rank test is calculated by summing the ratio between squared of summed observed number of events minus the summed expected number of events in each group and the summed expected number of events in the group (Kleinbaum & Klein 2005a). The equation for Log-rank test statistic is the following equation:

$$Log - rank\ test\ statistic = \sum_{i}^{\#groups} \frac{(O_i + E_i)^2}{E_i}$$

The log-rank statistic is approximately Chi-square with the degree of freedom equal to the number of groups minus 1. Therefore, the p-value from the log-rank test can be determined from Chi-square distribution table based on the significant level required.

1.16.3 Cox proportional hazard model for hazard ratio calculation

Cox proportional hazard model (or Cox regression) is another survival analysis method based on regression model. The Cox regression models the incidence or hazard ratio, i.e. the numbers of new events per population at-risk per unit time

(Kleinbaum & Klein 2005a). The hazard function is the probability that if a person survives to a time point *t*, the person will experience the event in the next time point. Let $h(\mathrm{t} \mid \mathrm{X}_{1i}, \mathrm{X}_{2i}, \dots, \mathrm{X}_{Ki})$ be the hazard function of i^{th} person at a time point *t*, where $i = 1, 2, \dots, n$ and $\mathrm{X}_{1i}, \mathrm{X}_{2i}, \dots, \mathrm{X}_{Ki}$ denote *K* predictors. $h_0(\mathrm{t})$ denotes the baseline hazard function at time point *t*, where all the predictors equal 0. The hazard ratio $h(\mathrm{t}) / h_0(\mathrm{t})$ can be expressed as a hazard function at time *t* as follow:

$$\frac{h(t|X_{1i}^*, X_{2i}^*, \dots, X_{Ki}^*)}{h_0(t|X_{1i}, X_{2i}, \dots, X_{Ki})} = \mathrm{e}^{(\beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki})}$$

The usefulness of the Cox model is that the baseline hazard does not have to be specified in order to estimate the predictor *β*.

1.16.4 Meta-analysis

Meta-analysis is a statistical technique used for combining the results from different and individual studies in order to increase the statistical power of the estimated effect from interventions (in the case of clinical trials of drug or treatments) or specific molecular entities (in the case of gene expression measurements). This thesis used a random effect model proposed by Cochran (1954) to combine the estimates from individual studies. The weighted estimator m_w can be calculated from the following equation:

$$m_w = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

where y_i is the effect from *i*-th study, and w_i is the weight of *i*-th study. The weight of each study can be calculated by the following equation:

$$w_i = \frac{1}{(t^2 + s_i^2)}$$

where t^2 is inter-study variance estimate, and s_i^2 is sampling variance from *i*-th study. The inter-study variance is calculated from the following equation:

$$t^2 = max\left\{0, \frac{1}{k-1}\sum\nolimits_i (y_i - \bar{y})^2 - \frac{1}{k}\sum\nolimits_i s_i^2\right\}$$

where k is the number of studies, and $\bar{y}$ is the mean of effect estimates from all the studies.

1.16.5 Data normalisation

In high-throughput gene expression experiments, by microarray or high-throughput sequencing, there are usually variations that arise from the process of the experiments rather than from the underlying biological differences between samples. Such variations (or biases) include non-uniformity in the hybridization process, scanner settings, the time of the experiments, ambient environment during the experiments, experimentation batches, and variations from different operators/experimenters. These variations can be minimised or removed by a process called normalisation. This thesis used microarray data generated from a specific single-channel microarray platform called Affymetrix GeneChip, which are short oligonucleotide arrays constructed using photolithography technology and used RNA from only one sample to be hybridised to the microarray. The process of normalisation in this type of microarray is used interchangeably with the term pre-processing, which includes three steps of background noise correction, normalisation, and probe level summarisation. The method that was used to normalise the microarrays in this thesis is called GCRMA (GC-content, Robust Multichip Average). GCRMA corrects the background noise in the microarray by using probe sequence information to take into account the GC-content in the sequence to calculate non-specific binding in the probe. Next, GCRMA normalises the data by using quantile normalisation, which is a non-parametric statistical method that transforms the different data distributions between samples to be identical in every samples. Finally, the probes that measure the same gene, which is collectively called a probe set, is summarised by using Tukey Median polish algorithm, where the median of each probe from every sample (row median) and the median of each sample from every probe (column median) are calculated, then calculate the median of medians for each row and column and finally add these median of medians together to yield a grand effect on each sample. This procedure is repeated until the sample median becomes zero for all the samples.

The RNA high-throughput sequencing or next generation sequencing or RNA-seq experiments also suffer from same biases as the microarray experiments. However, the quantile normalisation that performs well in removing the biases in microarray data was found to be not suitable in high-throughput sequencing. The main reason being that the quantile normalisation tends to classify weakly expressed genes as differentially expressed, thereby increase the false-positive rate in the analysis (Dillies et al. 2013). Another commonly used approach in RNA-seq data normalisation is called read per kilobase per million mapped reads or RPKM, which rescales the gene counts based on the library size and the length of the gene. Unfortunately, this approach introduces bias in the per-gene variance, which in turn also increase the false-positive rate from weakly expressed genes similar to quantile normalisation (Dillies et al. 2013). This thesis, therefore, utilised another approach called trimmed mean of M-values or TMM normalisation. TMM calculates a nomalisation factor, which is defined as a weighted trimmed mean of the log ratios between test samples and a reference sample (Robinson & Oshlack 2010). This method was found to perform better than RPKM and quantile normalisation in differential expressed gene analysis in the real samples (Dillies et al. 2013).

1.16.6 Cochran-Armitage trend test

Cochran-Armitage trend test is used to test whether there is an association between two categories in one variable and two or more categories (ordinal) in another variable. The data can be formulated in to a 2 x k contingency table, where the two-category variable is the response variable that has binomial distribution and the ordinal variable is the explanatory variable. The null hypothesis of this test is that there is no association (or trend) between these two variables. The test statistic for the Cochran-Amitage trend test has Chi-squared distribution with the degree of freedom equals 1 (the number of two response categories in response variable minus one).

1.16.7 Jaccard index and Jaccard distance

Similarity between two sets can be compared by a similarity measure called the Jaccard index, defined by the ratio between the intersection of the two sets and the union of the two sets (Levandowsky & Winter 1971). The Jaccard index can be expressed in the form of the following equation:

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Distance between sets is the one-complement of the Jaccard index in the form of the following equation:

$$D_J(X,Y) = 1 - J(X,Y)$$

The Jaccard distance is used as the distance metric to compare the distance (or dissimilarity) between two miRNAs families based on the predicted target genes that the two miRNAs have in common.

1.16.8 Hierarchical clustering

Clustering is an unsupervised classification technique extensively used in data mining. The main objective of the clustering is to group the data, which can be objects, records or observations, into classes of similar kinds. A cluster can be defined as a group of data that are similar to other objects in the cluster and dissimilar to the objects in other clusters. The similarity (or dissimilarity) of the data can be measured by a variety of metrics, such as correlation coefficients, distance between the data in the Euclidean space (i.e. Euclidean distance), or the Jaccard distance. Hierarchical clustering is an algorithm that uses recursive steps in order to group the data, either through a partitioning (divisive methods) or combining (agglomerative methods) of existing clusters. Agglomerative hierarchical clustering (a bottom-up approach) needs a criterion, which is termed linkage, in order to determine the distance between the data and group the data into clusters (Larose 2005). Several linkages for grouping the data are:

1. Single linkage or the nearest-neighbour criterion, where the minimum distance between each data points from two clusters determines the distance between two clusters.
2. Complete linkage or the farthest-neighbour criterion, where the maximum distance between each data points from two clusters determines the distance between the two clusters.
3. Average linkage, where the average distance between data points from two clusters determines the distance between the two clusters.
4. Ward’s linkage, where a predefined objective function is used to create clusters that maximise the similarity between the data in the clusters being combined, and also maximise the external separation between other clusters.

This thesis will utilise an agglomerative clustering algorithm based on Ward’s criterion.

1.16.9 Over-representation analysis or Fisher's Exact test

In a situation of sampling without replacement, to test whether occurrences of a certain number of objects in one class by random sampling from a universe of two classes of objects is more than expected by chance, the problem can be formulated into a 2 x 2 contingency table as follow:

		Selected by random sampling		
		Yes	No	Total
Member of class A	Yes	a	b	n_1
	No	c	d	n_2
	Total	m_1	m_2	N

This analysis of determining if the number in cell *a* is more than expected by chance is called over-representation analysis. The statistical test used to determine the probability value from this type of analysis is a categorical data analysis method called Fisher's exact test or hypergeometric test, derived from the name of the p-value distribution of the problem, which is called the hypergeometric distribution (Leonard 2000). The p-value from the 2 x 2 contingency table is a cumulative distribution function of the p-value, and can be calculated by using the following equation:

$$p(i \geq x) = 1 - \sum_{i=0}^{x} \frac{\binom{a}{i}\binom{N-m_1}{n_1-i}}{\binom{N}{n_1}}$$

where $p(i \geq x)$ is the cumulative probability of obtaining the number of objects, a, that are both selected by random sampling and members of class A, $\binom{t}{u} = \frac{t!}{u!(t-u)!}$ is a binomial coefficient, N is the total number of objects, n_1 is the number of objects in the class A, and m_1 is the number of randomly selected objects by random sampling.

1.16.10 Multiple hypothesis testing

In high-throughput gene expression analysis, each gene is subjected to individual hypothesis testing for differential expression between two conditions with small sample size. The number of the genes measured in one experiment usually exceeds 10,000 genes. The goal of multiple hypothesis testing procedures is to control the number of false-positives from individual hypothesis testing. Family-wise error rate (FWER) multiple hypothesis testing procedure such as Bonferroni correction tests if there is one or more false positives in all the hypotheses being tested. This FWER procedure is considered too stringent in biological experiments, since the number of truly differentially expressed genes is unknown. Another multiple hypothesis testing procedure called false discovery rate (FDR) calculates the probability of expected false-positive proportion at a certain pre-specified level. Commonly used FDR correction procedure is the Benjamini-Hochberg correction procedure (Benjamini & Hochberg 1995), where the proportion of expected false positives *E*(*Q*) is defined as follow:

$$E(Q) = \frac{m_0}{m}\alpha$$

where, m_0 is the number of true null hypothesis from individual hypothesis testing, m is the total number of hypotheses being tested, and α is the level of error rate desired to control. This thesis used Benjamini-Hochberg FDR correction procedure in all of the multiple hypothesis testing.

1.16.11 Shannon's information entropy

In information theory, entropy is the amount of information or uncertainty in the data. For a discrete random variable *X* with a certain probability distribution, the entropy of the probability distribution of the variable *H(X)* is defined as follow:

$$H(X) = -\sum_{i} p(x_i)\ \log p(x_i)$$

where x_i is the i-th random variable, and log is natural logarithm. This concept of information entropy was proposed by Shannon (1948). This thesis used the concept to determine the information entropy in the data in order to find the maximum contrast in the data, by calculating a Shannon entropy-like score based on this equation.

Chapter 2 : Roles of miR-22 in metabolic regulation and survival outcome of breast cancer patients

2.1 Introduction

As part of an ongoing effort to identify miRNAs that regulate metabolism in tumour cells, previous work in our laboratory identified miR-22 as a possible post-transcriptional regulator of three metabolic genes, *MTHFD2*, *ACLY*, and *ELOVL6*. An initial bioinformatics analysis using correlation analysis and O2-PLS modeling of molecular profiles from the NCI-60 cell line panel identified several miRNAs in which the expression were strongly correlated to metabolite levels: of these miR-22 had the highest variance explained by cross validation in the model (**Figure 2.1A**).

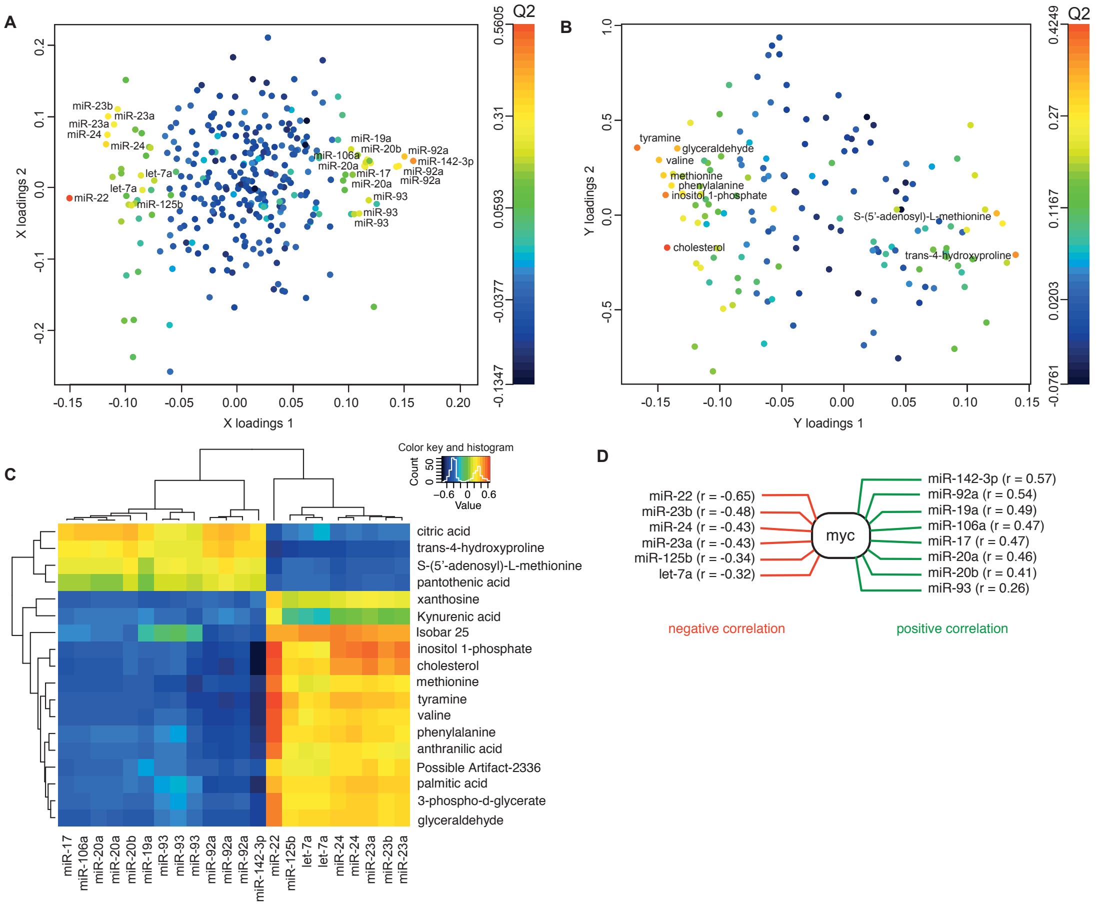


Figure 2.1 O2-PLS modeling of metabolite-miRNA associations in the NCI-60 panel (From Koufaris et al., submitted). A) O2-PLS model correlated component 'loadings' and individual variable cross validated predicted explained variance (Q^2) for miRNAs. B) O2-PLS model correlated component 'loadings' and Q^2 for metabolites. C) Heatmap showing univariate Pearson correlation of selected miRNAs with metabolites in the NCI-60 panel. D) Pearson correlations of miRNAs identified from multivariate model with MYC mRNA expression across the NCI-60 panel.

The models also identified several statistically significant metabolites, which were intermediates in biological pathways such as glycolysis, TCA cycle, one-carbon metabolism, lipid metabolism and nucleotide metabolism (**Figure 2.1B-C**). The set of metabolite-associated miRNA were also highly correlated to c-Myc expression consistent with regulation of these microRNAs by c-Myc (**Figure 2.1D**). Further bioinformatic analysis identified miR-22 predicted target genes consistent across three or more target prediction algorithms that were down regulated in two independent studies of the effects of miR-22 transfection. This analysis yielded seven target genes, in which two were metabolic target genes both involved in fatty acid synthesis: *ELOVL6* and *ACLY* (**Figure 2.2A-B**). Another metabolic target gene identified by overlapping output from multiple target prediction software was *MTHFD2*, involved in mitochondrial one-carbon metabolism. This gene was of interest since a high degree of negative correlation between miR-22 and the metabolite S-(5'-adenosyl)-L-methioine (SAM), an intermediate in the one-carbon metabolism was observed (**Figure 2.1C**). A series of experiments conducted in breast cancer cell line MCF-7 confirmed that these three metabolic genes were indeed the genuine targets of post-transcriptional regulation of miR-22 (**Figure 2.2C-D** and **Figure 2.3A-D**). Hence, it was hypothesised that miR-22 and the target genes might have effects on the survival outcome of the breast cancer patients. This study aim was to determine the extent that these three metabolic target genes have on the survival outcome of breast cancer patients, and effect modification of miR-22 on the three target genes *in vivo*, by using the miRNA and mRNA expression profiles measured by microarray and next generation sequencing (NGS) experiments from publicly available breast cancer data sets.

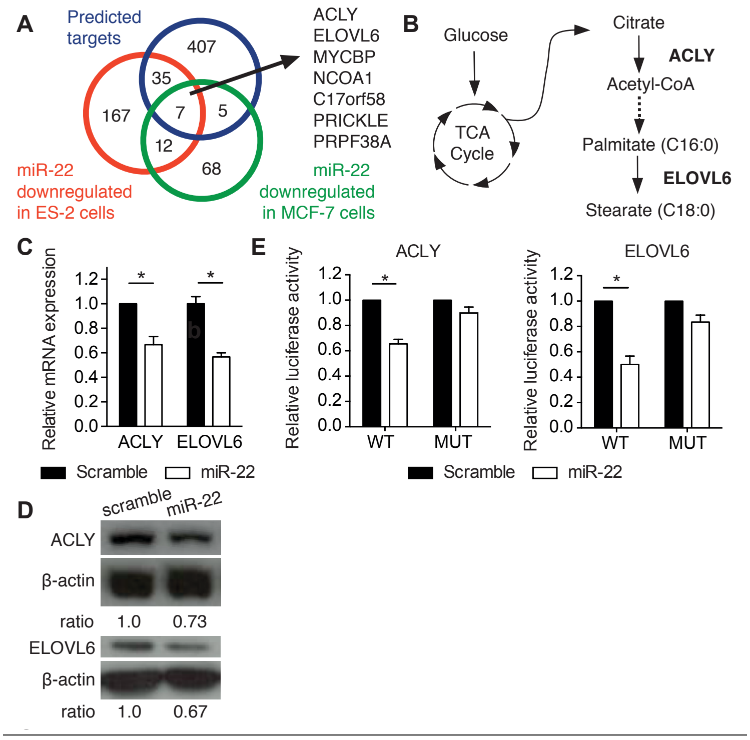


Figure 2.2 miR-22 represses fatty acid synthesis and elongation by direct targeting of ACLY and ELOVL6 (From Koufaris et al., submitted). A) Identification of putative miR-22 direct targets, identifying common targets among genes from two independent studies with 1.5-fold downregulation after miR-22 transfection of ES2 cells, MCF-7 cells, and genes predicted to be miR-22 targets by three or more target prediction algorithms. B) Scheme of *de novo* synthesis and elongation of fatty acid and involvement of ACLY and ELOVL6 in these pathways. C) qRT-PCR detection of *ACLY* and *ELOVL6* mRNA in MCF-7 cells transfected with either miR-22 mimic, or a negative control (scramble) after 48 hours. Data are normalised to β-actin and shown as mean ± s.e.m., n=3. *p <0.05. D) Immunoblotting analysis of ACLY and ELOVL6 in miR-22-transfected MCF-7 cells after 48 hours.

2.2 Materials and methods

The data used in this study were all retrieved from publicly available sources. The breast cancer data sets used can be divided into two main groups, one which reports directly levels of the mature miR-22 and another which examined expression of the miR-22 host gene (*MIR22HG*), a potential surrogate marker of mature miR-22 expression. More details of these data can be found in the results section. The correlation coefficient metric used throughout this study was Spearman's rank correlation coefficient (Spearman ρ). The combined correlation coefficients were calculated by a vote counting procedure proposed by Bushman & Wang (1995).

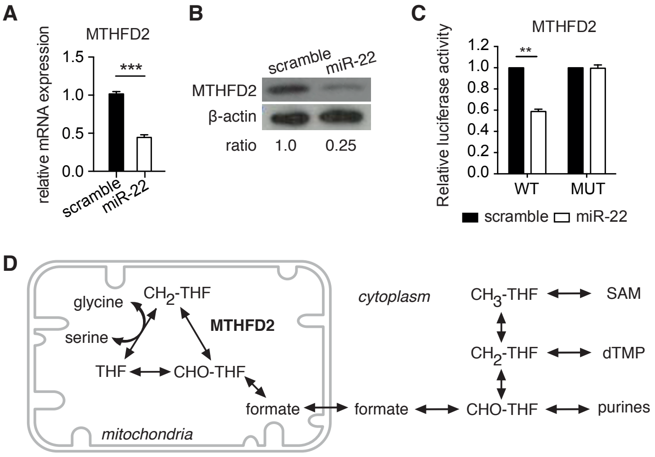


Figure 2.3 The mitochondrial enzyme MTHFD2 is repressed by miR-22 and affects one-carbon metabolism in cancer cells (From Koufaris et al., submitted). A) Representative immunoblots for MTHFD2 in MCF-7 transfected with miR-22 mimic. B) qRT-PCR for *MTHFD2.* Data were normalised to β-actin. C)Validation of direct targeting of *MTHFD2* by miR-22 using a luciferase reporter assay. Vector containing the luciferase reporter harboring the wildtype 3'UTR of gene (WT) or the 3'UTR with specific mutation of miR-22 binding sites (MUT) were co-transfected with miR-22 or negative mimic control oligonucleotide into MCF-7 cells. D) Scheme of mf mammalian mitochondrial one-carbon metabolism and the metabolic role of MTHFD2. All data are shown as mean ± s.e.m., n = 3. ** p<0.01. *** p<0.001.

Survival analyses were performed using Kaplan-Meier estimator to create survival curves from the follow-up data (Kleinbaum & Klein 2005a). The Cox proportional hazards model was used to calculate the hazard ratio of the gene of interest on the survival data (Kleinbaum & Klein 2005b). Meta-analysis was performed using random effect model on the Cox proportional hazard estimates (Cochran 1954), and using Fisher's method on the log-rank p-values associated with Cox proportional hazard estimates (Whitlock 2005). The Cochran-Armitage test for trend was used to test if the association between miR-22 and the target genes resulted in effect modification (Armitage 1955). All statistical calculations and plots were performed in R statistical environment version 3.0 with the following packages: `survival` for Kaplan-Meier estimates, Cox proportional hazard estimates and associated log-rank p-values calculations, and all KM-plots in survival analyses; `survcomp` and `rmeta` for combining Cox proportional estimates and log-rank p-values, and all forest plots in meta-analyses; and `coin` for Cochran-Armitage test for trend p-values calculations in effect modification determination between miR-22 and the target genes.

2.3 Results

2.3.1 Correlation between the expression of the mature miR-22 gene product and the host gene MIR22HG mRNA

To test if expression of the host gene *MIR22HG* (previously known as C17orf91) was sufficiently associated to levels of the mature miR-22 to act as a surrogate marker, Spearman's Rank correlation coefficient (Rho or ρ) between mature miR-22 and the host gene was calculated for several data sets where expression data on both were available (**Table 2.1**). Overall, very high degrees of agreement were observed. Spearman's ρ ranged from 0.4-0.53, and all ρ were statistically significant. Given these observations, in several independent data sets, it was decided that *MIR22HG* could be used as a surrogate for mature miR-22 expression.

2.3.2 Normalisation effect on correlation coefficients of in vivo data sets

Variation in the platforms used in measuring both mature miR-22 and the host genes (i.e. high-throughput sequencing and microarray technologies) did not have significant effect on the correlation. However, data normalisation on high-throughput sequencing data did have a very significant effect on the correlation coefficient. The current normalisation method employed in most miRNA-seq data analysis with multiple samples is read per kilobase per million (RPKM), which normalises each miRNA species based on their transcript length on each sample.

Table 2.1 Spearman's rank correlation coefficients between mature miR-22 and the host gene.

	Data set	Spearman's ρ	FDR adjusted p-value
NCI-60 data	miR-Israel Lab x Gx-GeneLogic (HG-U133)	0.528	2.71E-5
	miR-Israel Lab x Gx-GeneLogic (HG-U95A)	0.424	8.00E-4
	miR-Weinstein Lab x Gx-GeneLogic (HG-U133)	0.398	0.002
	miR-Weinstein Lab x Gx-GeneLogic (HG-U95A)	0.396	0.002
Breast Cancer data (TCGA)	miR-UNC (miRNAseq + HiSeq2000) x Gx-BCGSC (HiSeq2000)	0.512	< 2.20E-16

For *in vitro* (NCI-60) data sets, mature miR-22 was measured by RT-PCR (Israel Lab [Gaur et al., 2007]) or custom spotted microarray (Weinstein Lab [Blower et al., 2007]), and the host gene expression was measured by two separate Affymetrix microarray designs. NCI-60 microarray experiments were carried out by GeneLogic, Inc (Shankavaram et al. 2007). For TCGA data set, both mature miR-22 and the host gene were measured by high-throughput sequencing platforms by Illumina (The Cancer Genome Atlas Network 2012).

Nevertheless, in this study, it was found that the standard RPKM normalised mRNA and miRNA lead to underestimation of correlation between mature miR-22 and the target genes. **Table 2.2** and **Table 2.3** show the correlation coefficients after RPKM- and TMM-normalised between mature miR-22 and the three target genes from Farazi et al. (2011) and TCGA breast cancer data sets, respectively. For both data sets, the TMM-normalised data yielded statistically significant and higher negative correlation coefficients than RPKM-normalised data, except for *ELOVL6* in Farazi et al. data set. The ρ between miR-22 and the three metabolic target genes, *MTHFD2*, *ACLY*, and *ELOVL6,* from Farazi et al. data set using RPKM-normalised data were -0.24, -0.14 and -0.19, respectively, and for TMM-normalised data the ρ were, -0.33, -0.17 and -0.07, respectively. For TCGA data set, the ρ for RPKM-normalised data were -0.07, 0.06 and -0.04, and the ρ for TMM-normalised data were -0.29, -0.06 and -0.11, for *MTHFD2*, *ACLY* and *ELOVL6*, respectively. This clearly showed that normalisation of miRNA-seq data was affecting the differential miRNA analysis.

Table 2.2 Correlation coefficients comparison between RPKM- and TMM-normalised of Farazi et al. (2010) breast cancer data set.

Gene	RPKM-normalised		TMM-normalised	
	Spearman's ρ	p-value	Spearman's ρ	p-value
MTHFD2	-0.24	0.002	-0.33	2.11E-5
ACLY	-0.14	0.07	-0.17	0.02
ELOVL6	-0.19	0.01	-0.07	0.35

Table 2.3 Correlation coefficients comparison between RPKM- and TMM-normalised of TCGA breast cancer data set.

Gene	RPKM-normalised		TMM-normalised	
	Spearman's ρ	p-value	Spearman's ρ	p-value
MTHFD2	-0.07	0.15	-0.29	2.20E-16
ACLY	0.06	0.23	-0.06	0.02
ELOVL6	-0.04	0.46	-0.11	0.0003

2.3.3 Correlation between mature miR-22 and metabolites – in vitro data sets

Using microRNA and metabolite data sets based on NCI-60 cell panel, miR-22 has strong and negative correlations with several metabolites. The metabolites with strongest negative correlation coefficients with miR-22 were trans-4-hydroxy-L-proline (hydroxyproline), S-(5'-adenosyl)-L-methionine (SAM), 5'-S-methyl-5'-thioadenosine (MTA), and citric acid. **Table 2.4** shows Spearman's Rank correlations and adjusted p-values between miR-22 from two independent microRNA array data sets based on NCI-60 cell panel and the three metabolites. Hydroxyproline is the metabolite of amino acid proline, and both are major components in collagen proteins and connective tissues in human. SAM is the major donor of methyl groups for most biosynthetic methylation processes in the cells. Synthesis and recycling of SAM is coupled with another metabolic pathway, which is the folate-dependent one-carbon metabolism. MTA is the substrate for adenine synthesis, which is a substrate for adenosine synthesis, a building block of DNA, and also a one-carbon (methylation) pathway metabolite. Citric acid or citrate is a substrate in the TCA cycle and fatty acid biosynthesis.

Table 2.4 Correlation between miR-22 and metabolites from *in vitro* data sets.

Metabolite	Sokilde et al. (2011)		Liu et al. (2010)	
	Spearman's ρ	FDR adjusted p-value	Spearman's ρ	FDR adjusted p-value
Hydroxyproline	-0.657	1.98E-5	-0.504	0.009
SAM	-0.630	1.98E-5	-0.415	0.03
MTA	-0.599	5.97E-5	-0.331	0.08
Citric acid	-0.588	7.80E-5	-0.434	0.02

Strongest negative correlation coefficients between miR-22 (Søkilde et al. 2011; Liu et al. 2010) and metabolites (National Cancer Institute 2014) from data sets based on NCI-60 cell panel**.**

Another metabolic gene predicted to be a target of miR-22 by at least three microRNA target prediction software was methylenetetrahydrofolate dehydrogenase/cyclohydrolase (*MTHFD2*). This gene is the mitochondrial bifunctional enzyme in the SAM cycle, which was found to be a rate limiting step for synthesising 5,10-methylenetetrahydrofolate, a substrate for purine biosynthesis (Pawelek & MacKenzie 1998).

These results confirmed the previous work of our group that these metabolites have high degree of negative correlations with miR-22, and rationalised the three metabolic target genes that were proposed to be under the regulation of miR-22.

2.3.4 *Correlation between miR-22 and the three metabolic target genes – in vivo data sets*

To test the hypothesis that *ACLY*, *MTHFD2*, and *ELOVL6* were the targets of miR-22 *in vivo*, the correlation between the miR-22 expression and mRNA levels of these targets in a series of tumour data sets were calculated. The expressions of these metabolic target genes were extracted from breast cancer patient data sets from several breast cancer studies. The selected data sets for this work were based on the studies used in the Kaplan-Meier plotter website (kmplotter.com). The criteria for data sets to be selected were as follow: 1) survival data [i.e. relapse-free follow-up time] were available and 2) the gene expression platform used was Affymetrix HG-U133 family. The survival data, i.e. the relapse-free period of the patients, were used for subsequent survival analyses. The gene used as the surrogate of miR-22 expression was *MIR22HG* (previously known as C17orf91), which was presented in the Affymetrix HG-U133 family GeneChip platform. In total, nine breast cancer data sets from eight breast cancer studies were selected for this study. Affymetrix probeset names for *MIR22HG* and the target genes are shown in **Table 2.5**. Note that where multiple probesets were available, the probesets reported for the target genes of miR-22 were the best performing probesets, i.e. the ones that give the highest correlations and/or highest hazard ratios in subsequent survival analyses. This was based on the presumption that noise in these data sets were more likely to reduce a true correlation than create a false positive.

Two mature microRNA data sets were used in this work. First data set is from The Cancer Genome Atlas (TCGA) project. Briefly, TCGA is a comprehensive data repository for several cancers with multiple molecular measurements by microarray technology or high-throughput sequencing technology. The breast cancer data used in this analysis consist of mature microRNAs and gene expression measured by high-throughput sequencing and survival data (follow-up time and vital status) of the patients. The second data set is from Farazi et al. (2010), a collection of microRNA and gene expression of 161 breast carcinomas and 6 cell lines with survival data. microRNA expression was measured by high-throughput sequencing and gene expression was measured by custom spotted microarray. **Table 2.6** summarises the number of samples in each data set that was used in this study.

The results of the correlation analyses are shown in **Table 2.7**. miR-22 was negatively correlated with c-Myc (Spearman's ρ ranged from 0.003 to -0.46), consistent with previous findings that miR-22 is regulated by this transcription factor. miR-22 was also negatively correlated with its target genes that have been previously identified by microRNA target prediction software and confirmed by cell culture experiment. The trend was clear that miR-22 is negatively correlated with these target genes, albeit with non-statistically significant p-values in majority of the cases. Spearman's Rho for *ACLY*, *MTHFD*, and *ELOVL6* were ranging from -0.02 to -0.25, -0.08 to -0.33, and 0.0006 to -0.32, respectively. Combined correlation coefficients between miR-22 and the target genes were also calculated by using the combining sample correlation coefficients and vote counts procedure (Bushman & Wang 1995). It was evidently clear from the combined correlation coefficients that these target genes were negatively correlated with the miR-22, and that these target genes were possibly the real targets for post-transcriptional repression by miR-22 *in vivo*. The combined Spearman's ρ [with 95% confidence interval] for *MTHFD2*, *ACLY*, and *ELOVL6* were -0.24 [-0.20, -0.28], -0.07 [-0.03, -0.11], and -0.11 [-0.08, -0.15], respectively.

Figure 2.4 shows the relationship between mature miR-22 expression and the three metabolic target genes in detail. The patients in each data set were divided into four bins according to mature miR-22 expression quartiles (the first quartile and fourth quartile are the lowest and highest mature miR-22 expression, respectively), and the boxes and whiskers showed the expression level of its target genes in each bin. For *MTHFD2*, the median expression in each bin decreases with higher miR-22 expression, which is also evidence that miR-22 might be associated with *MTHFD2 in vivo*. The cell culture performed within Keun group (**Figure 2.3A-C**) showed that *MTHFD2* was repressed by miR-22, which confirmed that the association is genuine. For *ACLY* and *ELOVL6*, visual inspection did not show associations between the genes and miR-22 (**Figure 2.4**).

Table 2.5 Probeset names of *MIR22HG* and the target genes from Affymetrix GeneChip platforms used in breast cancer studies.

Gene Symbol	Selected probeset name	Affymetrix GeneChip platforms
MIR22HG	214696_at	HG-U133A, HG-U133plus2
MTHFD2	201761_at	HG-U133A, HG-U133plus2
ACLY	201128_s_at	HG-U133A, HG-U133plus2
ELOVL6	204256_at	HG-U133A, HG-U133plus2
c-MYC	202431_s_at	HG-U133A, HG-U133plus2

Note that there are multiple probesets for *ACLY* and *ELOVL6* and only best performing probesets were selected for these two genes for correlation and subsequent survival analyses.

Table 2.6 Number of samples from data sets used in this studies.

Data set		Number of samples	Median*/mean^ age at diagnosis (years)	Median*/mean^ Follow-up time (years)	Number of events (relapse)
Mature miR-22	TCGA	1126	NA	2.7^ [0-18.6]	37
	Farazi et al. (2010)+	161	50^ [26-83]	7.1^ [0.6-17.8]	31
miR-22 host gene	GSE1456	159	58^ [42-75]	6.2^ [0.2-8.4]	40
	GSE16391	55	59^ [46-78]	3^ [0.9-5.4]	55
	GSE2034	286	54^ [32-70]	6.4^ [0.1-14.2]	107
	GSE2990	102	58^ [32-86]	7^ [0.02-14.5]	40
	GSE3494	251	62^ [28-93]	7.1^ [0.08-12.7]	61
	GSE6532 (GPL570)	87	NA	12.7* [NA]	28
	GSE6532 (GP96)	72	NA	6.1* [NA]	19
	GSE7390	198	47* [NA]	14* [NA]	51
	GSE9195	77	63^ [42-82]	12.5^ [NA]	10

+ number of samples with microRNA measurement is 185 but samples with corresponding gene expression is 161.

^ marks the data with mean values.

* marks the data with median values. NA = Not available.

Table 2.7 Correlation coefficients (Spearman's ρ) between mature miR-22 or the host genes and the target genes.

Data set		Possible miR-22 Target genes				
		MTHFD2	*ACLY*	*ELOVL6*	c-MYC	
Mature miR-22	TCGA	-0.2932	-0.0647	-0.108	-0.1666	
	Farazi et al. (2010)	-0.3301	-0.1746	-0.0732	-0.2531	
miR-22 host gene	GSE1456	-0.1939	-0.2599	-0.1206	-0.2876	
	GSE16391	-0.1508	-0.0997	0.0738	-0.4639	p-value
	GSE2034	-0.1232	-0.0253	-0.1801	0.069	< 0.001
	GSE2990	-0.2481	-0.1549	-0.1415	-0.2163	< 0.01
	GSE3494	-0.3289	-0.0350	-0.1222	-0.0163	< 0.05
	GSE6532 (GPL570)	-0.1677	-0.0659	-0.2029	-0.1418	
	GSE6532 (GP96)	-0.2003	-0.1028	-0.2727	0.0034	
	GSE7390	-0.0891	-0.0419	-0.0006	-0.1786	
	GSE9195	-0.3159	-0.0219	-0.3256	0.0163	
Combined correlation		**-0.24**	**-0.07**	**-0.11**	**-0.13**	
[95% confidence interval]		**[-0.20, -0.28]**	**[-0.03, -0.11]**	**[-0.08, -0.15]**	**[-0.09, -0.17]**	

P-value levels associated with the correlation coefficients are shown on the right of the table. The combined correlation coefficients were calculated by using the combining sample correlation coefficients and vote counts procedure (Bushman & Wang 1995).

2.3.5 Patient survival outcome based on expression in upper and lower tertiles of MIR22HG and the metabolic target genes – individual in vivo data sets

Mature miR-22, the host gene *MIR22HG*, and its three target genes were used to perform survival analysis on the data sets with follow-up information of the patients. The main hypothesis of this study was that miR-22 and/or the three metabolic target genes have important roles in survival outcome of breast cancer patients. To this end, mature miR-22, *MIR22HG*, and the metabolic target genes were subjected to survival analysis by calculating survival functions using Kaplan-Meier estimator, based on previously mentioned variables. Each variable was divided into high and low expression groups according to the expression tertiles (i.e. the patients were divided into three equal bins). High expression group is the patients with the genes in upper tertile, whereas the low expression group is the patients with the genes in lower tertile. Survival curves were then plotted for the high and low groups in each data set. **Figure 2.5**, **Figure 2.6**, **Figure 2.7**, and **Figure 2.8** show survival curves from individual data sets between high and low expression of *MIR22HG*, *MTHFD2*, *ACLY* and *ELOVL6*, respectively. The hazard ratio (HR) calculated by Cox proportional hazards model with 95% confidence intervals (CI) and associated log-rank p-values for each data set is shown at the bottom right corner of each plot.

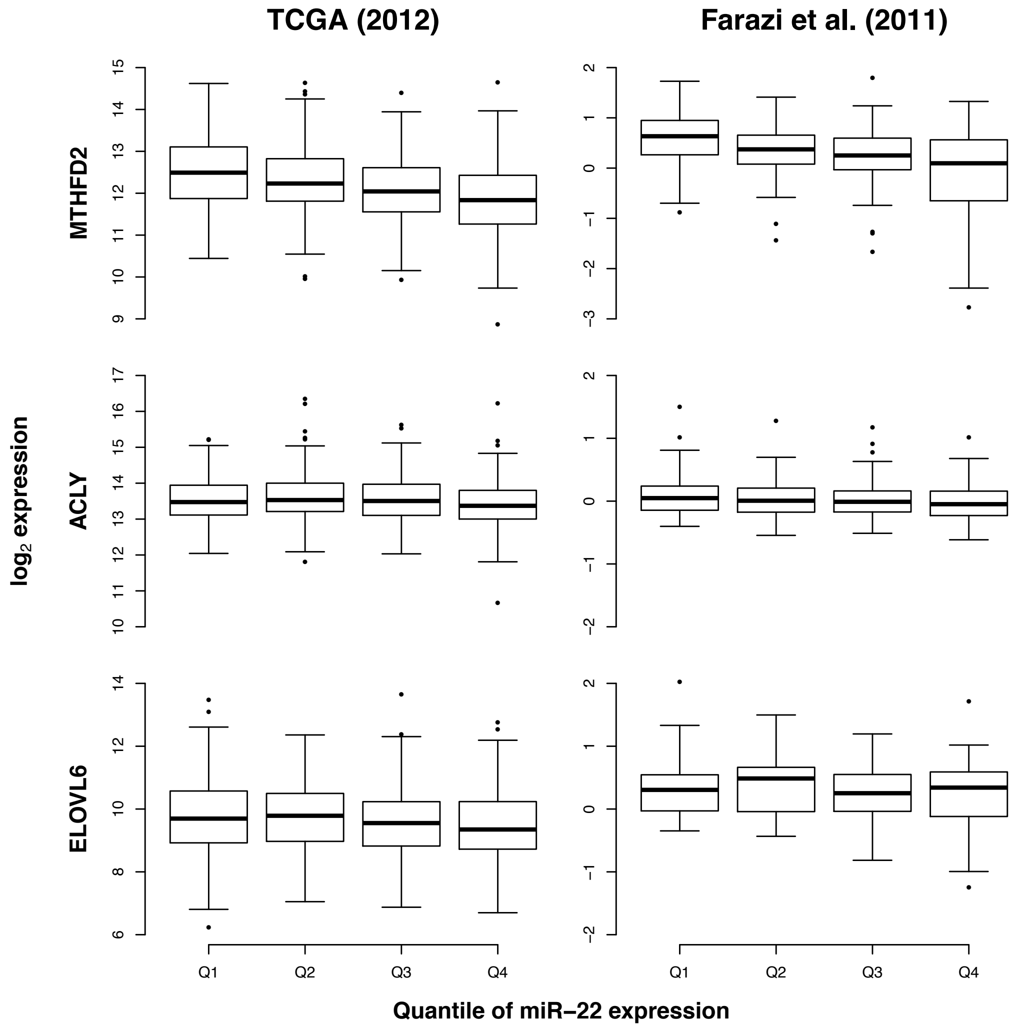


Figure 2.4 Box plots of miR-22 target genes expressions according to expression quartiles of mature miR-22 from two breast cancer data sets. Correlation coefficients and associated p-values between miR-22 target genes and mature miR-22 are corresponded to table 2.7.

The data showed that on individual data sets level, the majority of data sets do not yield any statistically significant results at p-value threshold of 0.05. For *MIR22HG*, only three data sets have statistically significant results. These three data sets showed that breast cancer patients with high expression of *MIR22HG* have better survival outcome than the patients with low *MIR22HG* expression (**Figure 2.5**). The data sets (with log-rank p-value) with statistically significant results for *MIR22HG* were GSE16391 (0.03), GSE3494 (0.002), and GSE6532.GPL96 (0.001). The opposite trends were observed in the analyses on the three target genes of miR-22. Where the data sets yield statistically significant results, all the patient groups with low expressions of *MTHFD2* and *ELOVL6* have better survival outcome than those with high expression of these two genes (**Figure 2.6** and **Figure 2.8**, respectively). The data sets with statistically significant results (with log-rank p-value) for *MTHFD2* were GSE2034 (0.02) and GSE3494 (0.02), and for *ELOVL6* were GSE2034 (0.005), GSE3494 (0.005) and GSE6532.GPL570 (0.03). For *ACLY*, there is no data set with statistically significant result at the p-value threshold of 0.05 (**Figure 2.7**). The results from individual data sets could not give a conclusive answer whether the miR-22 host gene and the target genes can be used as predictors for survival outcome of breast cancer patients.

2.3.6 Patient survival outcome based on expression in upper and lower tertiles of miR-22 host gene and the metabolic target genes – all in vivo data sets combined

The expression of miR-22 host gene, i.e. *MIR22HG*, and its target genes, i.e. *MTHFD2*, *ACLY* and *ELOVL6*, from each data set was divided into tertiles. Next, samples in each data set in upper and lower tertiles were labeled as high and low according to their expression values. Finally, samples in upper and lower tertiles from nine data sets were combined into one data set, together with its follow-up and vital status information of each sample. These combined data sets were then subjected to survival analysis. **Figure 2.9** shows the survival curves of the combined data sets on *MIR22HG*, *MTHFD2*, *ACLY* and *ELOVL6*.

For *MIR22HG*, *MTHFD2* and *ELOVL6*, the results are statistically significant at the p-value threshold of 0.05. The results suggest that breast cancer patients with high expression of *MIR22HG* and low expression of *MTHFD2* and *ELOVL6* have better survival outcome than the patients with low expression of *MIR22HG* and high expression of *MTHFD2* and *ELOVL6* (**Figure 2.9A**, **B** and **D**). From the survival analysis, high expression of *MIR22HG* has a protective effect in the breast cancer patients, with hazard ratio of 0.78 (95 % confidence interval [CI] = 0.63 – 0.97), log-rank p-value = 0.027. High expression of *MTHFD2* and *ELOVL6* has the opposite effect on the survival outcome of breast cancer patients. Patients with high expression of *MTHFD2* have a hazard ratio of 1.54 (95% CI = 1.24 – 1.92), log-rank p-value = 0.0001. Patients with high expression of *ELOVL6* have a hazard ratio of 1.58 (95% = 1.29 – 1.95), log-rank p-value = 1.35 x 10^{-5}. For *ACLY*, the result is not statistically significant at the same level of p-value cut-off (**Figure 2.9C**).

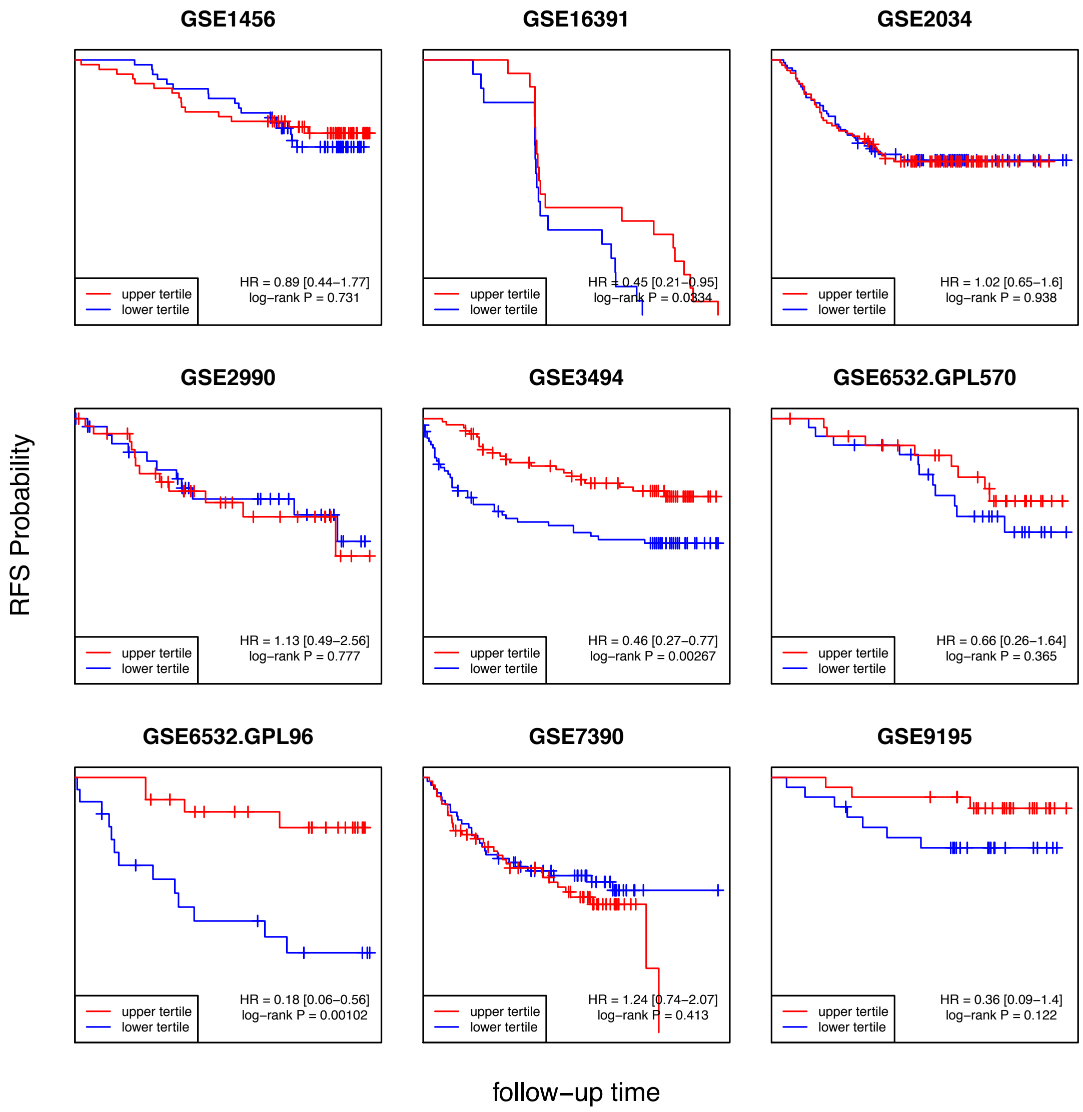


Figure 2.5 Survival curves of nine data sets for *MIR22HG*. The survival curves were calculated by using Kaplan-Meier estimator based on relapse-free survival time of the patients. The patients were divided by tertiles of *MIR22HG* expression. X axis is follow up time before censored in days and y axis is relapsed-free survival (RFS) probability calculated by Kaplan-Meier estimator. Blue and red lines represent survival probability trajectories of patients with high and low expression of *MIR22HG*, respectively. Vertical bars are censored times of patients. Hazard ratio for patients with high *MIR22HG* expression, with 95% confidence interval in square brackets, and its associated log-rank p-value of each data set is on the lower right corner of the each plot.

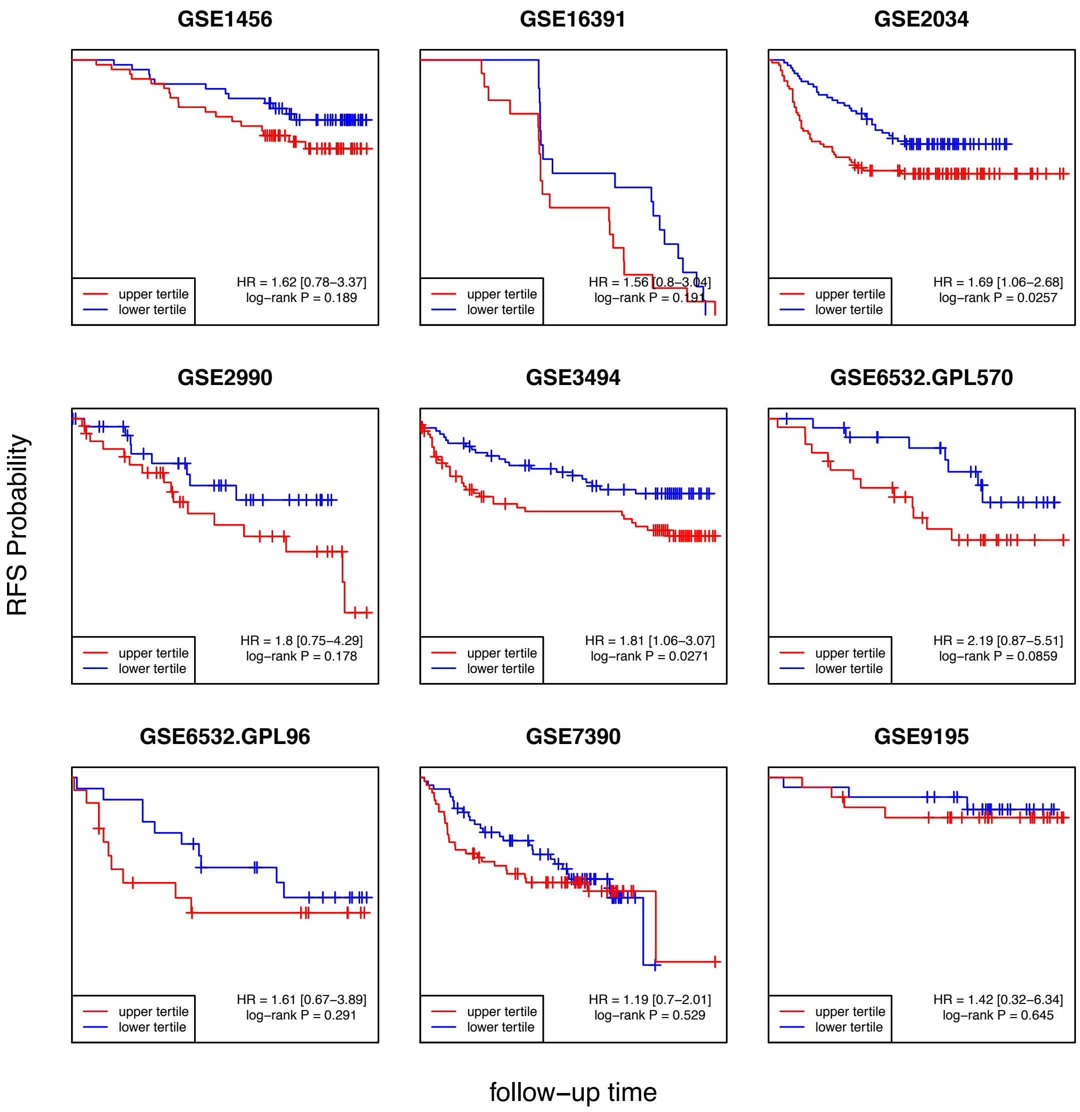


Figure 2.6 Survival curves of nine data sets for *MTHFD2*. The survival curves were calculated by using Kaplan-Meier estimator based on relapse-free survival time of the patients. The patients were divided by tertiles of *MTHFD2* expression. X axis is follow up time before censored in days and y axis is relapsed-free survival (RFS) probability calculated by Kaplan-Meier estimator. Blue and red lines represent survival probability trajectories of patients with high and low expression of *MTHFD2*, respectively. Vertical bars are censored times of patients. Hazard ratio for patients with high *MTHFD2* expression, with 95% confidence interval in square brackets, and its associated log-rank p-value of each data set is on the lower right corner of the each plot.

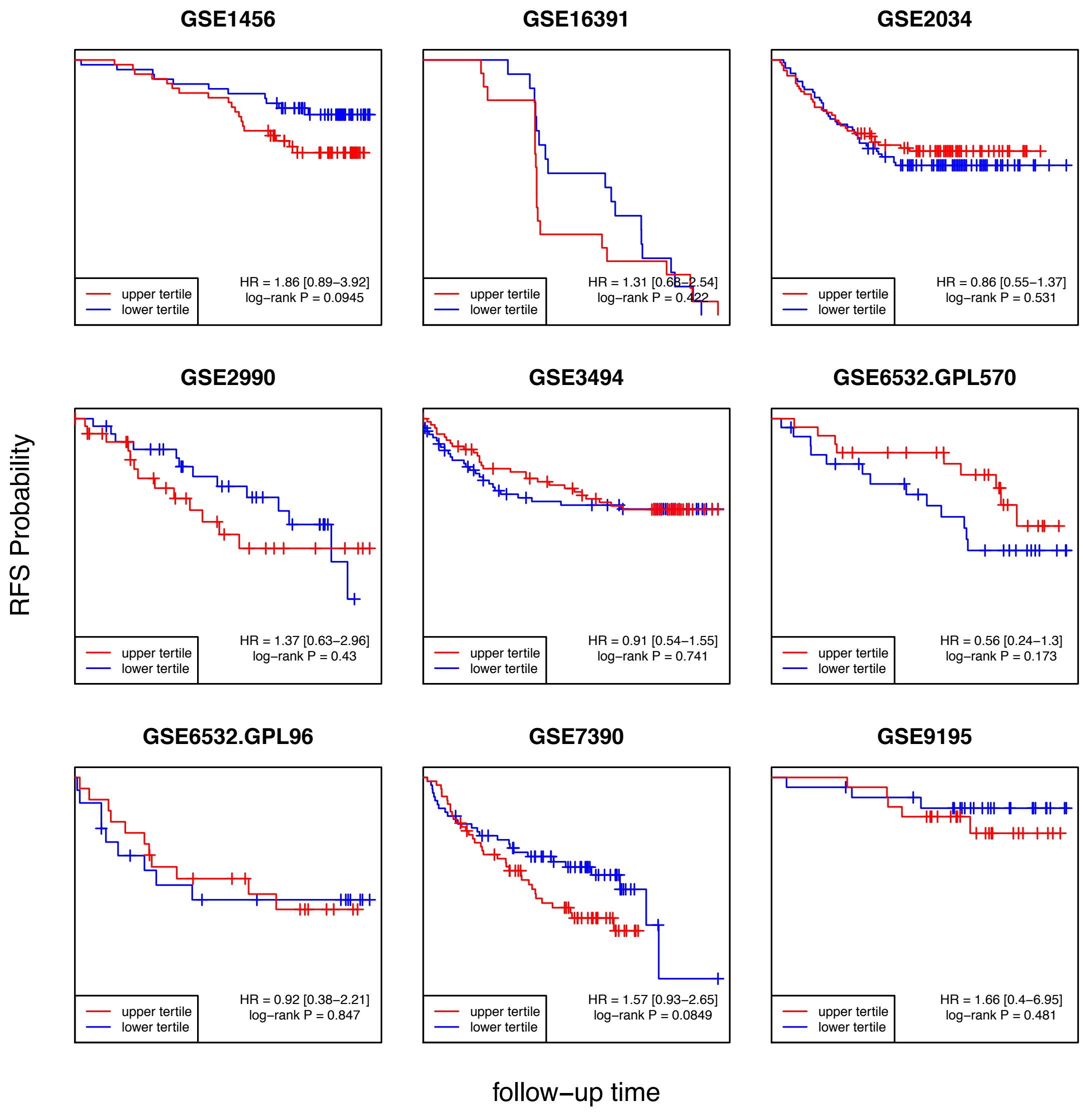


Figure 2.7 Survival curves of nine data sets for *ACLY*. The survival curves were calculated by using Kaplan-Meier estimator based on relapse-free survival time of the patients. The patients were divided by tertiles of *ACLY* expression. X axis is follow up time before censored in days and y axis is relapsed-free survival (RFS) probability calculated by Kaplan-Meier estimator. Blue and red lines represent survival probability trajectories of patients with high and low expression of *ACLY*, respectively. Vertical bars are censored times of patients. Hazard ratio for patients with high *ACLY* expression, with 95% confidence interval in square brackets, and its associated log-rank p-value of each data set is on the lower right corner of the each plot.

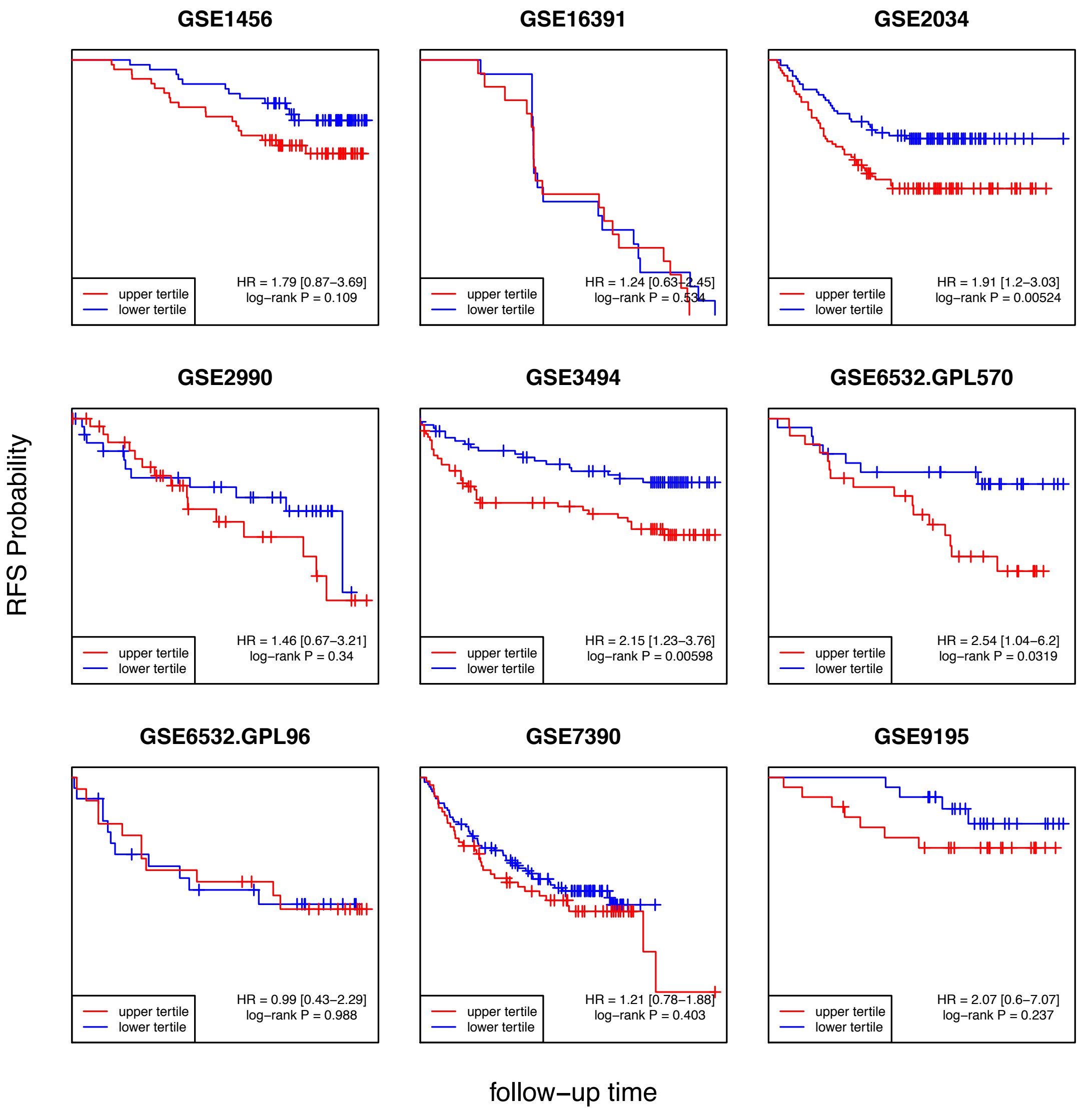


Figure 2.8 Survival curves of nine data sets for *ELOVL6*. The survival curves were calculated by using Kaplan-Meier estimator based on relapse-free survival time of the patients. The patients were divided by tertiles of *ELOVL6* expression. X axis is follow up time before censored in days and y axis is relapsed-free survival (RFS) probability calculated by Kaplan-Meier estimator. Blue and red lines represent survival probability trajectories of patients with high and low expression of *ELOVL6*, respectively. Vertical bars are censored times of patients. Hazard ratio for patients with high *ELOVL6* expression, with 95% confidence interval in square brackets, and its associated log-rank p-value of each data set is on the lower right corner of the each plot.

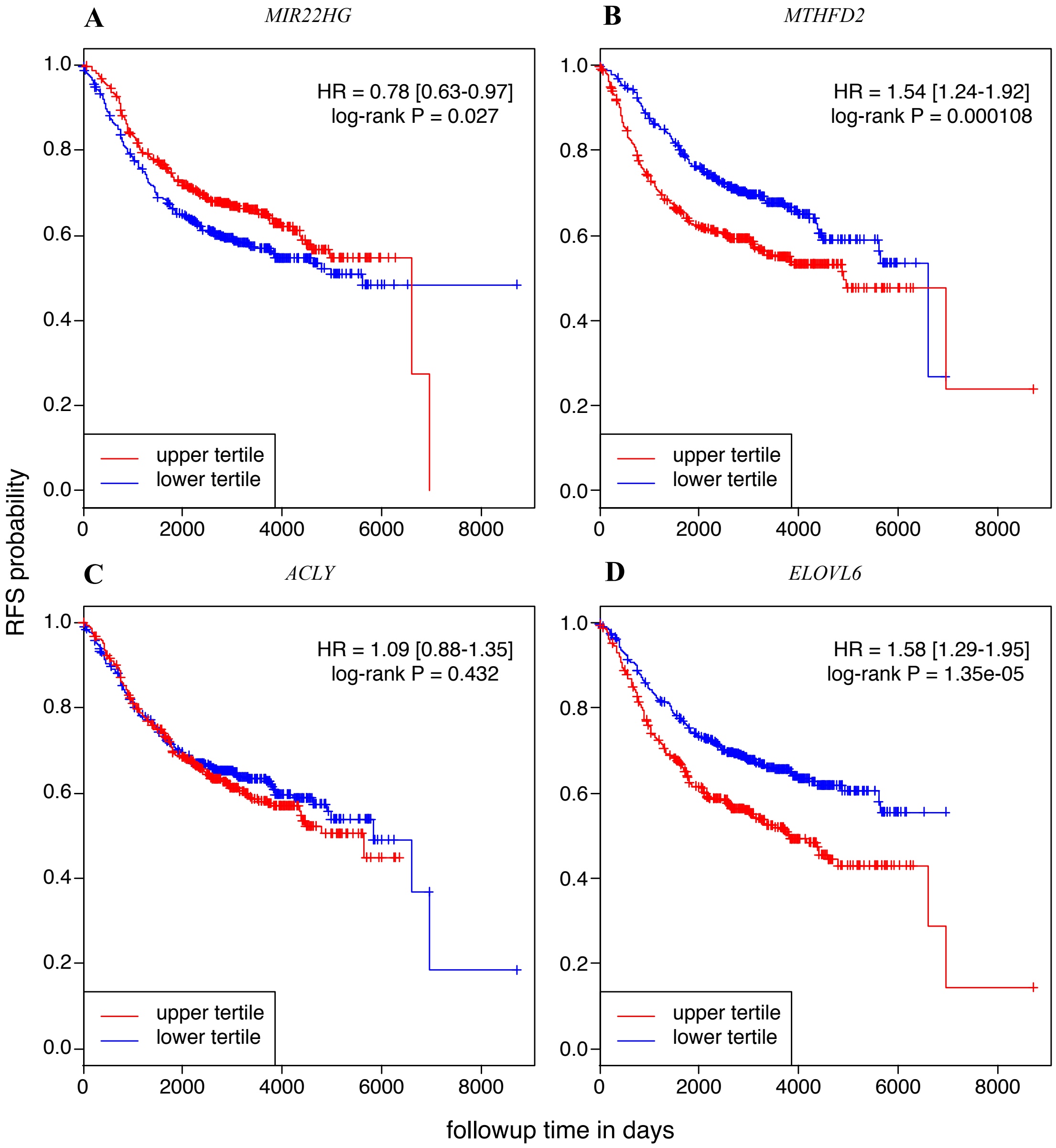


Figure 2.9 Survival curves from all data sets combined for *MIR22HG* and the three metabolic target genes. A) *MIR22HG*, B) *MTHFD2*, C) *ACLY*, D) *ELOVL6.* The survival curves were calculated by using Kaplan-Meier estimator based on relapse-free survival time of the patients. The patients were divided by tertiles of *MIR22HG* and the three target genes expressions. X axes are follow up time before censored in days and y axes are relapsed-free survival (RFS) probability calculated by Kaplan-Meier estimator. Blue and red lines represent survival probability trajectories of patients with high and low expression of *MIR22HG* and the three target genes, respectively. Vertical bars are censored times of patients. Hazard ratio for patients with high *MIR22HG* and the three target genes expression, with 95% confidence interval in square brackets, and its associated log-rank p-value of each data set is on the lower right corner of the each plot.

2.3.7 Meta-analysis

Because the combined KM analysis did not take into account the sizes of the different studies, the hazard ratios of *MIR22HG* and the metabolic target genes found by the combined KM analyses might be over- or underestimated. In order to compare and combine the effect of *MIR22HG* and the target genes on the survival outcome of the breast cancer patients from multiple data sets, taking into account the different effect sizes from different data sets, meta-analysis was used to combine the results from several breast cancer studies. The data sets were subjected to meta-analysis using random effects model. Forest plots were created to illustrate the effects that the genes have on patient survival outcome. Overall hazard ratios and associated log-rank p-values were calculated from Cox proportional hazards model between high and low expression groups divided by upper and lower tertiles expression of those genes. **Figure 2.10** shows forest plots on meta-analysis for *MIR22HG*, *MTHFD2*, *ACLY*, and *ELOVL6*.

The meta-analyses yielded the results similar to the survival analyses in the previous section, where *MIR22HG*, *MTHFD2* and *ELOVL6* were statistically significant at p-value cut-off of 0.05. From meta-analyses (**Figure 2.10**), high expression of *MIR22HG* has a protective effect on the survival outcome of breast cancer patients, with the hazard ratio of 0.68 (95% CI = 0.47 – 0.98), log-rank p-value = 0.001 (**Figure 2.10A**). High expressions of *MTHFD2* and *ELOVL6* have adverse health effect on the survival outcome of the breast cancer patients (**Figure 2.10B** and **D**). The hazard ratio of patients with high expression of *MTHFD2* is 1.58 (95% CI = 1.27 – 1.95), log-rank p-value = 0.019. The hazard ratio of patients with high expression of

ELOVL6 is 1.54 (95% CI = 1.25 – 1.88), log-rank p-value = 0.004. For *ACLY*, the result from meta-analysis was not statistically significant (**Figure 2.10C**).

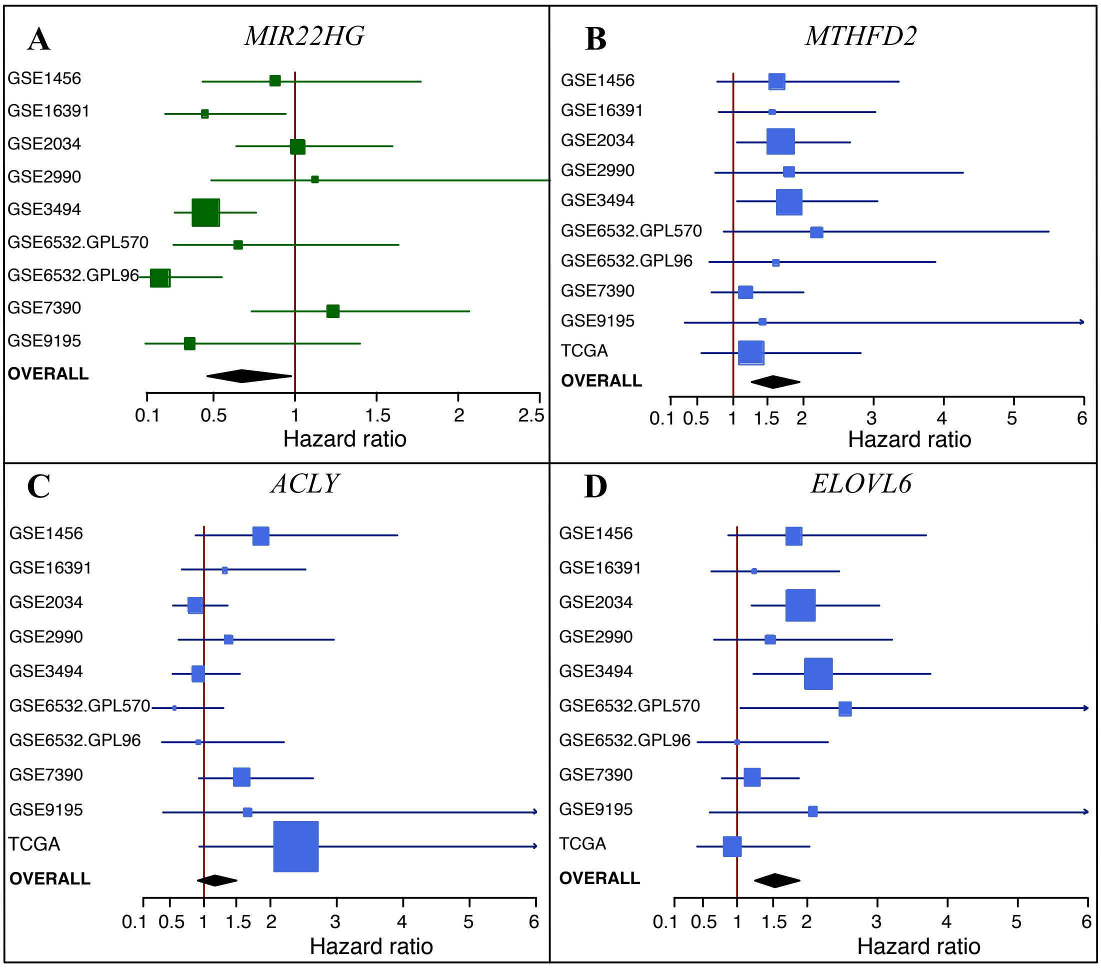


Figure 2.10 Meta-analyses of *MIR22HG* and the three metabolic target genes. To compare the differences between data sets, the forest plots were generated based on hazard ratios (HRs) and associated 95% CIs from the data sets. To achieve higher statistical power from more samples, the overall hazard ratios were calculated by using random-effects model. The boxes and the whiskers indicate HRs and 95% confidence intervals of separate studies, and sizes of the boxes signify the effect sizes of the studies. The diamond represents overall HR of *MIR22HG* and the three metabolic target genes and [upper and lower 95% confidence interval] on both ends of the diamond: A) *MIR22HG,* 0.68 [0.47-0.98], log-rank p-value = 0.001, n = 857; B) *MTHFD2,* 1.58 [1.27-1.95], log-rank p-value = 0.019, n = 1088; C) *ACLY,* 1.17 [0.92-1.50], log-rank p-value = 0.16, n= 1088; and D) *ELOVL6,* 1.54 [1.25-1.88], log-rank p-value = 0.004, n = 1134.

2.3.8 Effect modification between miR-22 and the metabolic target genes

Although miR-22 and the metabolic target genes on their own might have effects on the patient survival outcome, a combined effect between miR-22 and one of its target genes might even be stronger than the effect of miR-22 or the target genes on their own. Therefore, the effect modification of miR-22 on the three metabolic target genes on patient survival outcome were investigated. *MIR22HG* (for data sets from Kmplotter webtool) and mature miR-22 expression (from TCGA data set) were divided into three groups according to their expression tertiles, which are upper, middle and lower tertiles. The target genes, i.e. *MTHFD2 ACLY* and *ELOVL6*, were divided into two groups (high and low) according to upper and lower tertiles. Patients were then divided into six categories according to the expression of miR-22 and one of its target genes. **Table 2.8** shows the categories of the patients according to miR-22 and one of the metabolic target genes expression. Finally, survival curves were plotted for each target genes, comparing survival outcome between high and low expression groups in each expression tertile of miR-22.

In order to test whether or not the expression of miR-22 target genes in different miR-22 expression tertiles were statistically associated with the survival outcome of the breast cancer patients, Cochran-Armitage test for trend was used to calculate the probability of the association being occurred at random. For the data sets from Kmplotter webtool, *MIR22HG* was used as the surrogate for mature miR-22.

The effect modification of miR-22 on the three metabolic target genes found to be statistically significant were between miR-22+*MTHFD2* and miR-22+*ELOVL6*, with association p-values calculated by Cochran-Armitage trend test equal 0.001 and 3.46

x 10^{-5}, respectively. The hazard ratios for breast cancer patients that have high expression of these two target genes were increased after combining the effect of having low expression of miR-22. The hazard ratio of the patients who have high *MTHFD2* expression and low miR-22 expression was 1.72 (95% CI = 1.19 – 2.5), log-rank p-value = 0.003 (**Figure 2.11**). The hazard ratio of the patients with high expression of *ELOVL6* and low expression of miR-22 was 2.44 (95% CI = 1.64 – 3.63), log-rank p-value = 6.21 x 10^{-6} (**Figure 2.13**). The effect modification of miR-22 was stronger for *ELOVL6* than *MTHFD2* and the effect gradually lessen the lower expression groups of the metabolic target genes, albeit not statistically significant. However, the survival outcome of the patients with low expression of miR-22 were similar in every group of the metabolic target genes expressions for both *MTHFD2* and *ELOVL6*, suggesting that the high expression of miR-22 was associated with the breast cancer patient survival outcome.

The effect modification of miR-22 on *ACLY*, however, was not statistically significant, and combining the low expression of miR-22 did not significantly alter the hazard ratio for the patients with high expression of *ACLY* (**Figure 2.12**).

Table 2.8 Patient categories based on miR-22 expression tertiles and miR-22 target genes expression.

Patient Category	miR-22 expression (*MIR22HG* and mature miR-22)	miR-22 target genes expression (*MTHFD2*, *ACLY* and *ELOVL6*)
1	Upper tertile	Lower Tertile (Low expression group)
2	Middle tertile	
3	Lower tertile	
4	Upper tertile	Upper tertile (High expression group)
5	Middle tertile	
6	Lower tertile	

The categories with the same shading are used to calculate the hazard ratios for having high expression of target genes.

Based on the data presented, it is evident that miR-22 is a post-transcriptional regulator of *MTHFD2* and *ELOVL6*, but not *ACLY in vivo*. The interaction between miR-22 and these two metabolic target genes also has real effect on the survival outcome of the breast cancer patients, which means that miR-22 and the two metabolic target genes *MTHFD2* and *ELOVL6* might have significant clinical relevance in the future.

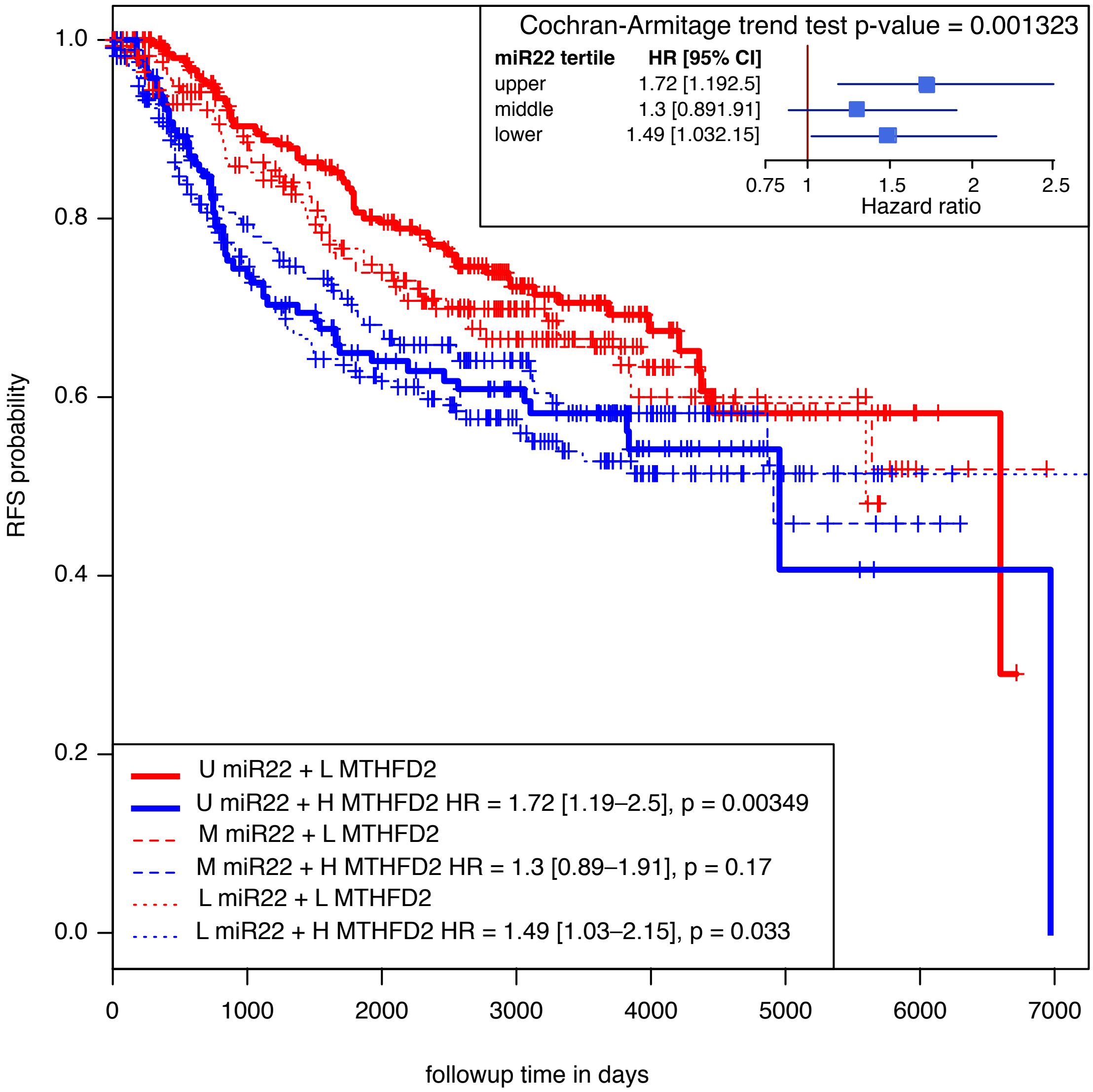


Figure 2.11 Effect modification of miR-22 on *MTHFD2.* A survival curve of combined data set for *MTHFD2* shows survival trajectories of breast cancer patients in different *MIR22HG* expression tertiles. The survival curve was calculated by using Kaplan-Meier estimator based on relapse-free survival time of the patients. U, M and L for miR-22 are survival trajectories for breast cancer patients in upper, middle and lower tertiles, respectively. H and L for *MTHFD2* are high and low expression group of breast cancer patients based on upper and lower tertiles of *MTHFD2*. Hazard ratios were calculated for the patients with high expression of *MTHFD2* in each miR-22 expression tertile. On the top right corner is a forest plot comparing hazard ratios of breast cancer patients with high *MTHFD2* expression (divided by upper and lower tertiles) between miR-22 expression tertiles.

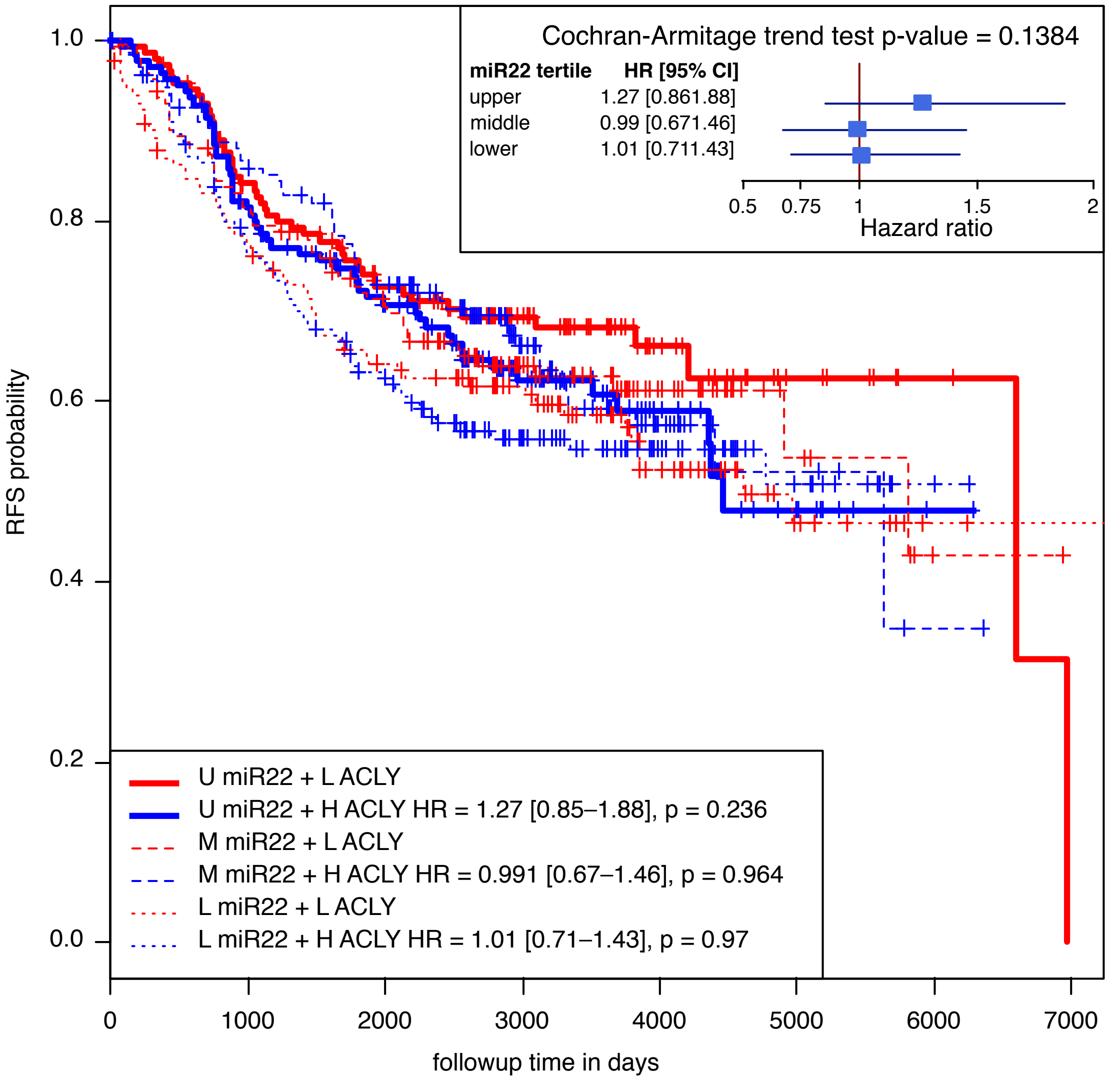


Figure 2.12 Effect modification of miR-22 on ACLY. A survival curve of combined data set for *ACLY* shows survival trajectories of breast cancer patients in different *MIR22HG* expression tertiles. U, M and L for miR-22 are survival trajectories for breast cancer patients in upper, middle and lower tertiles, respectively. H and L for *ACLY* are high and low expression group of breast cancer patients based on upper and lower tertiles of *ACLY*. Hazard ratios were calculated for the patients with high expression of *ACLY* in each miR-22 expression tertile. On the top right corner is a forest plot comparing hazard ratios of breast cancer patients with high *ACLY* expression (divided by upper and lower tertiles) between miR-22 expression tertiles.

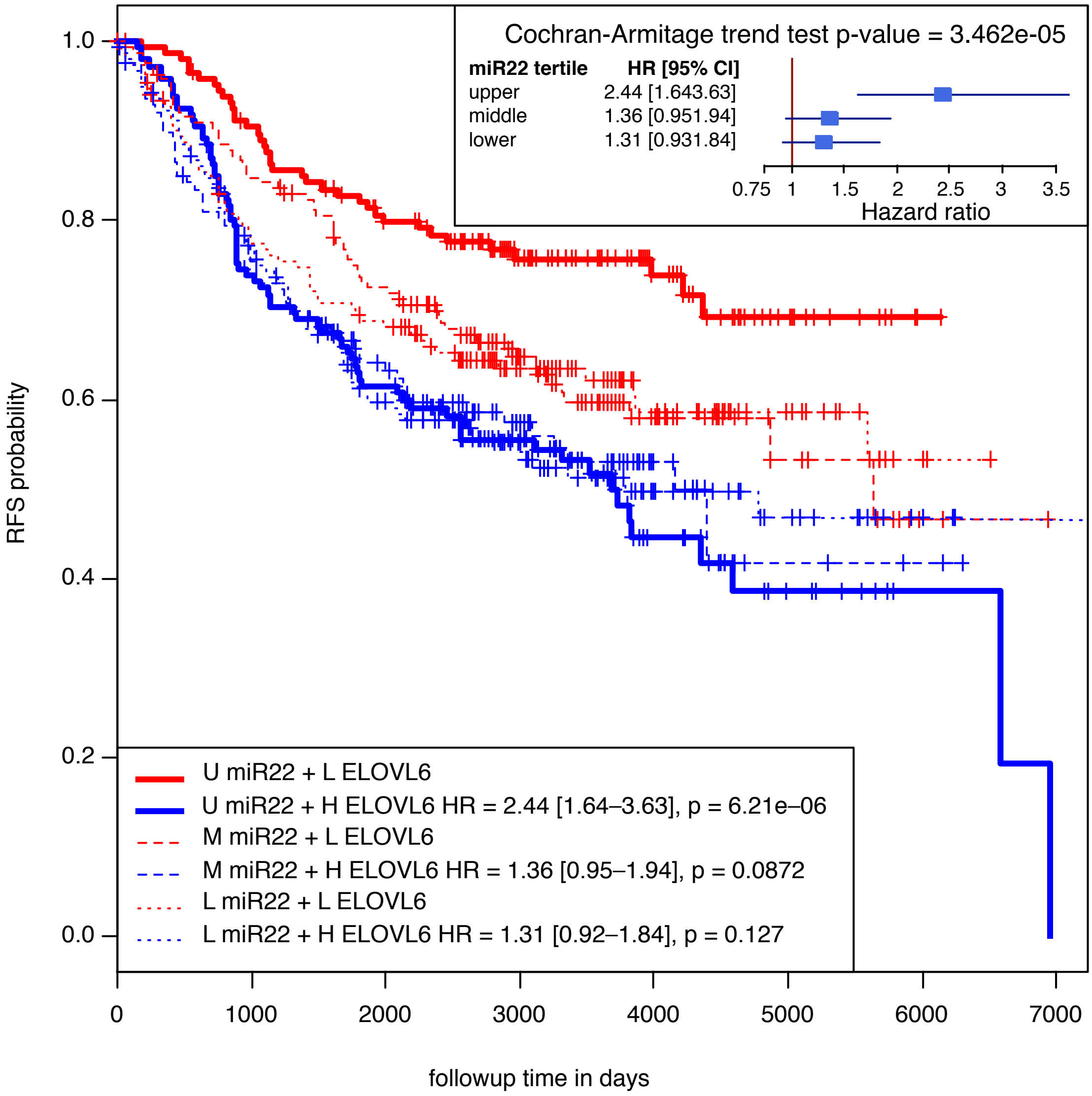


Figure 2.13 Effect modification of miR-22 on *ELOVL6.* A survival curve of combined data set for *ELOVL6* shows survival trajectories of breast cancer patients in different miR-22 expression tertiles. U, M and L for miR-22 are survival trajectories for breast cancer patients in upper, middle and lower tertiles, respectively. H and L for *ELOVL6* are high and low expression group of breast cancer patients based on upper and lower tertiles of *ELOVL6*. Hazard ratios were calculated for the patients with high expression of *ELOVL6* in each miR-22 expression tertile. On the top right corner is a forest plot comparing hazard ratios of breast cancer patients with high *ELOVL6* expression (divided by upper and lower tertiles) between miR-22 expression tertiles.

2.4 Discussion

2.4.1 MIR22HG as the surrogate of mature miR-22

The use of *MIR22HG* as the surrogate for mature miR-22 expression was justified by the fact that the correlation coefficients between the two in several independent data sets were very high (see **Table 2.1**). The NCI-60 *in vitro* data sets comprises of 59 different cancer cell lines, and therefore, the correlation from this data sets should be able to provide an accurate estimate of *MIR22HG* and the mature sequence. Several studies also confirmed that mammalian miRNAs usually correlate well with the host genes (Baskerville & Bartel 2005; Ruike et al. 2008), especially miR-22 (Rodriguez et al. 2004; Wang et al. 2011), and that the host gene could be used as a surrogate for mature miRNA expression (Gennarino et al. 2009). However, some studies suggested that, in some cell-specific contexts, host genes expressions are not reflecting those of their mature sequences (Sikand et al. 2009; Biasiolo et al. 2011). The causes for the inconsistency in host gene-mature sequence relationship might be due to: polymorphisms in the miRNA transcripts, affecting the biogenesis of the miRNA (Ha & Kim 2014); epigenetic control by DNA methylation and/or histone modification (Davis & Hata 2009; Ha & Kim 2014); and level of transcription factor c-Myc and the exact roles of c-Myc in cell-specific contexts of miR-22 transcription (J Xiong et al. 2010; Polioudakis et al. 2013; Kong et al. 2014). Nevertheless, *MIR22HG* is not the active molecule that is incorporated into the miRISC, which acts on the target genes. It is possible that *MIR22HG* transcript might be under all of the previously mentioned transcriptional controls and might not reflect the extent of the mature species.

Another drawback of using *MIR22HG* as a surrogate for mature miR-22 is the possibility of existence of alternatively spliced variants of the *MIR22HG* transcript. *MIR22HG* has four exons in the pri-miR-22 (Slezak-Prochazka et al. 2013), and it was found that the longer isoform of miR-22 is more effective than the shorter isoform in target repression (Nam et al. 2014). Therefore, the measured *MIR22HG* might not be an accurate estimate of mature miR-22 from the possibility that the *MIR22HG* measured did not measure the correct isoform of miR-22 by the microarray. As far as alternative splicing of miRNA is concerned, a recent report showed that the alternative splicing process itself negatively regulated the precursor miRNA levels (Melamed et al. 2013). It should also be noted that miR-22 is an exonic miRNA, which is different from majority of miRNAs that are intronic miRNAs. Currently there is no report on the overall extent of correlation between exonic miRNAs and the host genes. Unfortunately, the mature miR-22 breast cancer data set of Farazi et al. (2010) did not have *MIR22HG* (or *C17orf91*) measurement for *in vivo* correlation assessment to definitely confirm that *MIR22HG* and mature miR-22 is correlated.

2.4.2 Normalisation is a crucial step for differential miRNA comparison

The effect of normalisation for high-throughput sequencing has an impact on the values of mature miRNAs expression. Although the way high-throughput sequencing of RNA (whether mRNA or miRNA) are usually normalised, i.e. RPKM or read per kilobase exon model per million mapped reads (Mortazavi et al. 2008), provides good estimate of most of the miRNA species present in different samples, there are growing number of studies suggested that RPKM is not suitable for differential expression analysis (Bullard et al. 2010; Robinson & Oshlack 2010; Sun & Zhu 2012; Dillies et al. 2013). This might be due to the fact that there were different number of counts between different samples (Garmire & Subramaniam 2012). However, there are only a few studies comparing the effects and suitability of RNA-seq data normalisation and the results are still inconclusive, and it is even more controversial in the case of miRNA-seq data normalisation. Currently there is also no high-throughput mRNA and miRNA data sets that have both mature miR-22 and the host gene, *MIR22HG*, which can be used for a definitive test on the extent of the effect of normalisation.

2.4.3 Correlation does not equal causation

The correlation between miR-22 and the metabolic target genes, especially *MTHFD2* and *ELOVL6*, were quite convincing that these two genes might be affecting the survival outcome of breast cancer patients the most. However, it was not the case for *ACLY*, and the target gene of miR-22 with the strongest effect on the patient survival outcome was neither *MTHFD2* or *ACLY*, but *ELOVL6*. Despite the lowest correlation with miR-22 (**Table 2.7**), *ELOVL6* affected the patient survival outcome the most (**Figure 2.9**), especially when the effect of *ELOVL6* was modified by miR-22 expression (Figure 2.10). The effect of *ELOVL6* on patient survival outcome was the strongest when the patients also in the miR-22 high expression group. On the other hand, *ELOVL6* does not affect the survival outcome in the patient with miR-22 low expression, suggesting the effect modification of miR-22 on *ELOVL6*. *MTHFD2* also showed similar trend as *ELOVL6*, albeit at a lesser extent (**Figure 2.11**). Several approaches have been developed to identify effect modification among miRNAs or between miRNAs and their target genes based on their sequences (Yoon & De Micheli 2005) or their expression profiles (Boross et al. 2009; J. Xu et al. 2011; Y. Li et al. 2014), but the approach used in this study is simpler than the methods previously mentioned. However, this approach is quite limited on the number of miRNA and mRNA that can be tested in the analysis. As the number of miRNA and/or mRNA is increasing, the category numbers of the samples will also be increasing, which will complicate the analysis. Also the number of samples in each category will be decreasing, lessen the power of the analysis. Nevertheless, the approach in this study is suitable for targeted identification of the effect modification relationship between miRNA-mRNA because the number of samples used, which gives more power for the analysis.

2.4.4 High expression of miR-22 is beneficial to breast cancer patients

The results from this study suggested that miR-22 is beneficial to the breast cancer patients, possibly through the suppression of the metabolic target genes *MTHFD2,* and *ELOVL6*. These metabolic target genes of miR-22 have very important roles in cancer progression, especially in breast cancer. By suppressing these metabolic target genes, miR-22 is indirectly linked to several aspects of metabolic control in breast cancer.

A very recent study showed that patients with high expression of the protein product of *MTHFD2* have poorer survival outcome than the patients with low or no MTHFD2 expression (Liu et al. 2014). Another study showed that *MTHFD2* expression was increased in many types of cancer cells, and suppressing *MTHFD2* expression caused apoptosis in most of the cancer cell lines tested in the study, which imply that *MTHFD2* is needed for cancer cell proliferation and viability (Nilsson et al. 2014). Another study identified miR-9 as another post-transcriptional regulator of *MTHFD2*, and showed that *MTHFD2* depletion has a proliferative effect with overexpression of miR-9 (Selcuklu et al. 2012). A *MTHFD2* knock down experiment impaired cell migration and invasion into extracellular matrix in breast cancer cell lines (Lehtinen et al. 2013). Despite the weak evidence observed for an interaction between *ACLY* and miR-22, *ACLY* is also clearly important in cancer. It was found recently that *ACLY* gene product ATP citrate lyase (ACLY) has a role in reversing the epithelial-mesenchymal transition (EMT) process in non-small cell lung cancer cell lines and human mammary breast cancer lines, thereby promoting cell differentiation and reduction of cancer stem cells (Hanai et al. 2013). Furthermore, decreased ACLY leads to suppressed growth and/or apoptosis by blocking the triglyceride elongation process (Migita et al. 2014). The expression of *ELOVL6* was recently found to be down-regulated in mammary epithelial cells, promoting the cell differentiation process, and the low expression of *ELOVL6* was also found to be beneficial for breast cancer patients survival outcome (Dória et al. 2014). Note that previously mentioned study used the same data sets that were also used in this study. Another recent study showed that decrease *ELOVL6* reduce cell proliferation in liver cancer (Shiau et al. 2014). All these studies were in line with the previous findings in the Keun group laboratory and this study.

The effect modification of miR-22 on *ELOVL6* might be of interest in breast cancer therapy. This study showed that breast cancer patients with low expression of *ELOVL6* and high expression of miR-22 have the best survival outcome compared to other groups of patients. A combination therapy that enhances miR-22 expression and lower *ELOVL6* expression might be beneficial to the breast cancer patients, although the molecular mechanism of the therapy would have to be clearly elucidated. Other metabolic target genes of miR-22, which remain to be discovered in the future, might also help in the understanding of the roles of miR-22 in breast cancer.

2.4.5 The exact roles of miR-22 in cancer still remain controversial

Despite previously mentioned studies that were in line with the analysis presented here, there were also evidence that miR-22 is an oncogene, particularly in breast cancer (Song, Poliseno, et al. 2013). In this particular study, xenograft mouse models were used and identified another target gene, Ten eleven translocation 2 (*TET2*), which in turn inhibits demethylation of the promoter of a tumour suppressor miR-200 and promotes invasiveness and metastasis through EMT initiation. This study also showed that triple negative breast cancer (TNBC) patients with high expression of miR-22 have poorer survival outcome.

It is possible that miR-22 can regulate different processes through different sets of target genes. Although *TET2* is not a metabolic gene, it regulates another miRNA by remodeling the epigenetic landscape of the promoter of miR-200, suppressing the expression of this miRNA and in turn upregulate the transcription factors Zeb1, Zeb2 and SIP1 triggering the EMT process. It is also possible that miR-22 might behave differently in the *in vitro* and *in vivo* environments. Many studies previously described were mostly done in cancer cell lines, which might not be a representative environment for what is taking place in the actual human or other organisms. This goes to show that more experiments, especially *in vivo* studies such as mouse xenografts should be performed to ascertain the definitive roles of miR-22 and the mechanisms that it operates.

It was also a limitation of this study that the hormone status, estrogen receptor alpha (ER-α), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2 or ERBB2), were omitted, and patients were not divided based on hormone receptor status. The reason was partly due to the fact that some of the breast cancer data sets do not have hormone receptor status available. The breast cancer data that was used in Song, Poliseno, et al. (2013) study was among a few studies that have complete hormone receptor statuses. However, this data set has only around 200 samples, and after excluding other hormone receptors positive samples, the samples used in the study were hardly representative of the population. Another explanation is that TNBC are usually very heterogeneous samples, with completely different gene expression profiles among the TNBC samples. It is possible that in TNBC, miR-22 will behave completely different, and act as an oncogene instead of a tumour suppressor. A very recent study showed that miR-22 was expressed differently in different cell types (Nam et al. 2014). Therefore, it is possible that even in the same cell types but with different context (differentiated cell or stem cell-like phenotypes), miR-22 might be expressed differently. Nevertheless, only with more studies, complete with all the hormone receptor status, the question of whether miR-22 is indeed an oncogene in TNBC can be revealed. Evidently, miR-22 has multiple target genes, with different roles in cancer, and should be studied more thoroughly to elucidate the exact mechanisms of actions for the benefit of the breast cancer sufferers.

Chapter 3 : Global analysis on common predicted target genes of miRNA

3.1 Introduction

Metabolic control by miRNAs is poorly understood relative to transcriptional control of mRNAs. Because miRNAs directly suppress or are involved in degradation of mRNA, the relationship of miRNA-mRNA is more obvious than that of miRNA-metabolite. Metabolic pathways are potentially now more complicated than ever because of this additional layer of control by miRNAs. Currently there are very few attempts in global analysis of miRNA-associated metabolic consequences. This might be due to the fact that miRNA-mRNA relationship is already more complex than expected. One way of reducing the complexity between miRNA-gene-metabolite networks is to find a subset of miRNAs that potentially target the same groups of genes, and then focus the analysis on the pathway level to limit the scope of miRNAs, genes, and metabolites that are involved in the analysis. This approach is a top-down approach; the opposite way compared to the approach used in the chapter 2 where the hypothesis was specifically generated by focusing on one specific set of miRNA-gene interactions (bottom-up approach). In this study, all microRNA families and their predicted target genes were used to calculate the proportions of probable sets of predicted metabolic targets between every pair of miRNA families to find miRNA clusters that have similar predicted metabolic target genes. The predicted metabolic genes in these miRNA clusters were then used in pathway analysis to find pathways that might be associated with these miRNA clusters.

3.2 Materials and methods

The P_{CT} scores are the probability of preferentially conserved targeting of an mRNA by a miRNA family. The P_{CT} score was defined as the expected signal-to-background ratio of a miRNA target site in an mRNA, calculated through branch-length score from multiple species alignment of the target site (Friedman et al. 2009). Predicted conserved target genes and associated P_{CT} scores of miRNA families, release 6.2 (June 2012) are retrieved on 7 October 2013 from TargetScan website (Bioinformatics and research computing 2012). In total, there are 87 miRNA families presented in TargetScan based on seed sequences of miRNAs. The P_{CT} scores were used as the weights for the weighted Jaccard index later. Metabolic gene list was retrieved from a study of human metabolic reconstruction RECON 1 list (Duarte et al. 2007). The overlapped predicted target genes from two miRNA families were trimmed down to metabolic genes only according RECON 1 metabolic gene list. The Jaccard index is a similarity metric used for measuring the difference between two sets, *X* and *Y*, which is the ratio of overlapped members between *X* and *Y* of all the members in *X* and *Y* (Levandowsky & Winter 1971) and can be expressed in the following equation:

$$J = \frac{|c_X \cap c_Y|}{|c_X \cup c_Y|} \tag{1}$$

where *J* is the Jaccard index, and C_X and C_Y are the numbers of members in set *X* and *Y*, respectively. Jaccard distance is a dissimilarity metric derived from the Jaccard index used for measuring the distance between the two sets, which can be expressed as follow:

$$d_J = 1 - J \tag{2}$$

where d_J is the Jaccard distance. In this study, a weighted Jaccard index (Sethi & Alagiriswamy 2010) was used to calculate the similarity between two miRNA families. Instead of simply counting the number of overlapped predicted target genes between two miRNA families, P_{CT} score of each predicted target gene was used as the weight to take into account the probability of that predicted target gene being a real target for a certain miRNA. The weighted Jaccard index can be expressed as follow:

$$J_w = \frac{\sum_{x=1}^{M}(w_X \cap w_Y)}{\sum_{x=1}^{M} w_X \cup \sum_{y=1}^{N} w_Y} \tag{3}$$

where J_w is weighted Jaccard index, W_X and W_Y are the weights of the members in sets *X* and *Y*, respectively, which in this case set *X* and *Y* are sets of predicted gene targets from two miRNA families and W_X and W_Y are the P_{CT} scores of predicted target genes from two miRNA families. Weighted Jaccard distance of all miRNA family pairs was calculated from the weighted Jaccard index to produce the distance matrix between all miRNA families for the hierarchical clustering analysis. Hierarchical clustering was performed using Ward's minimum variance method or Ward's linkage (Ward 1963), where objective function is to obtain minimum distance between the miRNAs in the clusters. The clusters of miRNAs were defined by Dynamic Hybrid algorithm from R package 'Dynamic Tree Cut' (Langfelder et al. 2008). Pathway analysis was performed on predicted target gene lists of each miRNA family pair by using over-representation analysis (ORA) on KEGG pathway database (Kanehisa & Goto 2000). Finally, the number of pathways that are associated with miRNA families in each cluster was tested by ORA, to ascertain whether or not the number of pathways occurred more than expected by chance. The p-values associated with the over-represented pathways in each cluster was subjected to multiple hypothesis testing using the Benjamini-Hochberg false discovery rate procedure (Benjamini & Hochberg 1995). The pathways with FDR-adjusted p-value less than 0.05 were retained. All statistical calculations were performed in R statistical packages version 2.12. KEGG database was accessed through KEGGREST package (Tenenbaum 2013). Jaccard distance matrices were produced by R package ‘NeatMap’ (Rajaram & Oono 2010). **Figure 3.1** illustrates the analysis as a flow chart.

3.3 Results

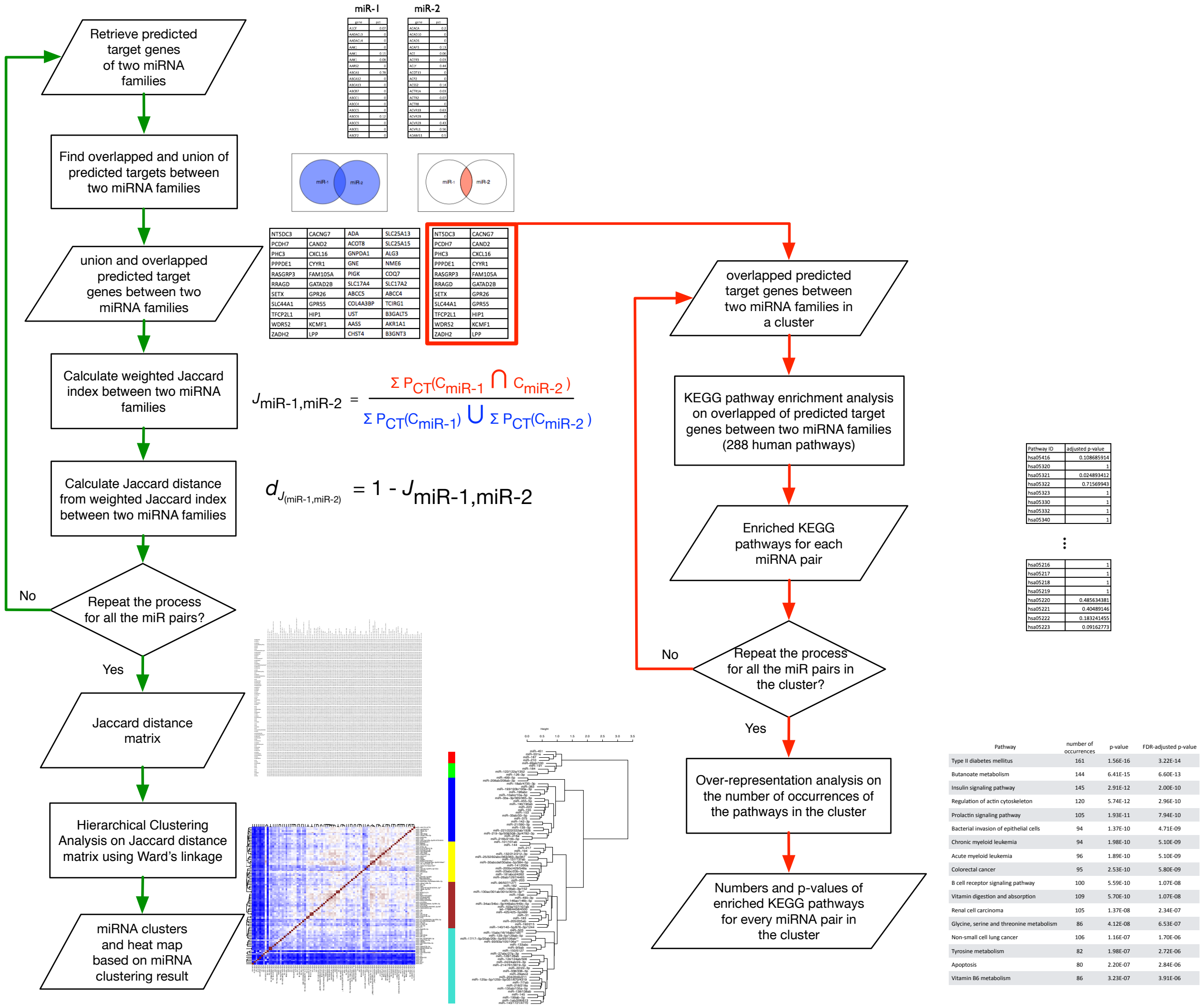
3.3.1 Basic statistics

Most of predicted target genes of miRNAs are evolutionarily conserved among mammals and the 3’UTR of those target genes are under selective pressure to be under the control of miRNAs. The estimate of the proportion of protein-coding genes under the control of miRNAs is more than 60% (Friedman et al. 2009). In this analysis, a crude estimate of the number of genes under the control of miRNAs is around 82%, based on the number of conserved predicted target genes in the current release of TargetScan database (Release 6.2, June 2012) and the estimated number of protein-coding genes of 20,687 genes (The ENCODE Project Consortium 2012). The number of overlapped metabolic genes based on RECON 1 metabolic network reconstruction study (Duarte et al. 2007) is around 8% of the predicted target genes. **Table 3.1** summarises the basic statistics of the raw data used in this study. The number of predicted target genes present in the database and the number of metabolic predicted target genes for each of the miRNA family are summarised in **Table 3.2**.

Table 3.1 Summary of basic statistics on the data used in this study.

	Number of Genes
Predicted protein-coding genes (ENCODE project)	20,678
Predicted target genes under miRNA control (TargetScan)	17,088
Metabolic genes (RECON 1)*	1,480
Overlapped between TargetScan and RECON 1 genes	1358

*The original number of metabolic genes in RECON 1 is 1496 genes. The genes that are note included in the calculation are due to the depreciated gene symbols (i.e. has been renamed or dropped) in the HUGO nomenclature.


miR-1
miR-2
Retrieve predicted target genes of two miRNA families
Find overlapped and union of predicted targets between two miRNA families
miR-1
miR-2
union and overlapped predicted target genes between two miRNA families
Calculate weighted Jaccard index between two miRNA families
Calculate Jaccard distance from weighted Jaccard index between two miRNA families
No
Repeat the process for all the miR pairs?
Yes
Jaccard distance matrix
Hierarchical Clustering Analysis on Jaccard distance matrix using Ward's linkage
miRNA clusters and heat map based on miRNA clustering result
overlapped predicted target genes between two miRNA families in a cluster
KEGG pathway enrichment analysis on overlapped of predicted target genes between two miRNA families (288 human pathways)
Enriched KEGG pathways for each miRNA pair
No
Repeat the process for all the miR pairs in the cluster?
Yes
Over-representation analysis on the number of occurrences of the pathways in the cluster
Numbers and p-values of enriched KEGG pathways for every miRNA pair in the cluster

$$J_{miR\text{-}1,miR\text{-}2} = \frac{\Sigma\, P_{CT}(C_{miR\text{-}1} \cap C_{miR\text{-}2})}{\Sigma\, P_{CT}(C_{miR\text{-}1}) \cup \Sigma\, P_{CT}(C_{miR\text{-}2})}$$

$$d_{J(miR\text{-}1,miR\text{-}2)} = 1 - J_{miR\text{-}1,miR\text{-}2}$$

NT5DC3	CACNG7	ADA	SLC25A13
PCDH7	CAND2	ACOT8	SLC25A15
PHC3	CXCL16	GNPDA1	ALG3
PPPDE1	CYYR1	GNE	NME6
RASGRP3	FAM105A	PIGK	COQ7
RRAGD	GATAD2B	SLC17A4	SLC17A2
SETX	GPR26	ABCC5	ABCC4
SLC44A1	GPR55	COL4A3BP	TCIRG1
TFCP2L1	HIP1	UST	B3GALT5
WDR52	KCMF1	AASS	AKR1A1
ZADH2	LPP	CHST4	B3GNT3

NT5DC3	CACNG7
PCDH7	CAND2
PHC3	CXCL16
PPPDE1	CYYR1
RASGRP3	FAM105A
RRAGD	GATAD2B
SETX	GPR26
SLC44A1	GPR55
TFCP2L1	HIP1
WDR52	KCMF1
ZADH2	LPP

Pathway ID	adjusted p-value
hsa05416	0.108685914
hsa05320	1
hsa05321	0.024893412
hsa05322	0.71569943
hsa05323	1
hsa05330	1
hsa05332	1
hsa05340	1
⋮	
hsa05216	1
hsa05217	1
hsa05218	1
hsa05219	1
hsa05220	0.485634381
hsa05221	0.40489146
hsa05222	0.183241455
hsa05223	0.09162773

Pathway	number of occurrences	p-value	FDR-adjusted p-value
Type II diabetes mellitus	161	1.56E-16	3.22E-14
Butanoate metabolism	144	6.41E-15	6.60E-13
Insulin signaling pathway	145	2.91E-12	2.00E-10
Regulation of actin cytoskeleton	120	5.74E-12	2.96E-10
Prolactin signaling pathway	105	1.93E-11	7.94E-10
Bacterial invasion of epithelial cells	94	1.37E-10	4.71E-09
Chronic myeloid leukemia	94	1.98E-10	5.10E-09
Acute myeloid leukemia	96	1.89E-10	5.10E-09
Colorectal cancer	95	2.53E-10	5.80E-09
B cell receptor signaling pathway	100	5.59E-10	1.07E-08
Vitamin digestion and absorption	109	5.70E-10	1.07E-08
Renal cell carcinoma	105	1.37E-08	2.34E-07
Glycine, serine and threonine metabolism	86	4.12E-08	6.53E-07
Non-small cell lung cancer	106	1.16E-07	1.70E-06
Tyrosine metabolism	82	1.98E-07	2.72E-06
Apoptosis	80	2.20E-07	2.84E-06
Vitamin B6 metabolism	86	3.23E-07	3.91E-06

Figure 3.1 Flow chart illustrating overall of the analysis. The genes used in the calculation were metabolic genes only.

Table 3.2 Summary of number of predicted target genes in each miRNA family.

miRNA family	All genes	Metabolic genes only	% of metabolic genes
let-7/98/4458/4500	2353	191	8.12
miR-101/101ab	3013	217	7.20
miR-103a/107/107ab	3366	246	7.31
miR-10abc/10a-5p	2475	179	7.23
miR-122/122a/1352	3161	244	7.72
miR-124/124ab/506	3482	277	7.96
miR-125a-5p/125b-5p*	3428	278	8.11
miR-126-3p	154	17	11.04
miR-128/128ab	4247	341	8.03
miR-129-5p/129ab-5p	5077	361	7.11
miR-130ac/301ab**	2633	195	7.41
miR-132/212/212-3p	2878	224	7.78
miR-133abc	2216	183	8.26
miR-135ab/135a-5p	2708	212	7.83
miR-137/137ab	2743	195	7.11
miR-138/138ab	2995	208	6.94
miR-139-5p	2535	189	7.46
miR-140/140-5p***	2106	158	7.50
miR-141/200a	3634	286	7.87
miR-142-3p	1601	113	7.06
miR-143/1721/4770	3288	245	7.45
miR-144	3332	245	7.35
miR-145	3247	243	7.48
miR-146ac/146b-5p	2773	210	7.57
miR-148ab-3p/152	2989	218	7.29
miR-150/5127	4538	376	8.29
miR-153	2430	176	7.24
miR-155	2390	177	7.41
miR-15abc/16****	3903	305	7.81
miR-17/17-5p#	4255	334	7.85
miR-181abcd/4262	3867	286	7.40
miR-182	3329	265	7.96
miR-183	2512	192	7.64
miR-184	731	54	7.39
miR-187	582	48	8.25
miR-18ab/4735-3p	2150	163	7.58
miR-190/190ab	1800	130	7.22
miR-191	568	39	6.87
miR-192/215	1759	132	7.50
miR-193/193b/193a-3p	2186	168	7.69
miR-194	2731	193	7.07
miR-196abc	1620	120	7.41
miR-199ab-5p	3071	245	7.98
miR-19ab	3025	214	7.07
miR-1ab/206/613	3117	266	8.53
miR-200bc/429/548a	3414	250	7.32
miR-203	4586	380	8.29
miR-204/204b/211	3985	285	7.15
miR-205/205ab	3175	246	7.75
miR-208ab/208ab-3p	1787	109	6.10
miR-21/590-5p	1920	142	7.40
miR-210	586	43	7.34
miR-214/761/3619-5p	4602	370	8.04
miR-216a	2849	210	7.37
miR-216b/216b-5p	2696	216	8.01
miR-217	2853	208	7.29
miR-218/218a	2940	226	7.69
miR-219-5p/508##	1622	122	7.52
miR-22/22-3p	2770	215	7.76
miR-221/222/222ab###	2275	165	7.25
miR-223	2332	162	6.95
miR-23abc/23b-3p	4206	322	7.66
miR-24/24ab/24-3p	4321	328	7.59
miR-25/32/92abc####	2779	213	7.66
miR-26ab/1297/4465	3419	265	7.75
miR-27abc/27a-3p	4075	315	7.73
miR-29abcd	2627	193	7.35
miR-30abcdef+	3261	237	7.27
miR-31	2901	224	7.72
miR-338/338-3p	3410	256	7.51
miR-33a-3p/365++	2099	157	7.48
miR-33ab/33-5p	2900	215	7.41
miR-34ac/34bc-5p+++	3372	244	7.24
miR-375	2267	182	8.03
miR-383	1978	160	8.09
miR-425/425-5p/489	2550	217	8.51
miR-451	397	36	9.07
miR-455-5p	2077	159	7.66
miR-490-3p	2539	177	6.97
miR-499-5p	2413	167	6.92
miR-503	1905	155	8.14
miR-551a	323	23	7.12
miR-7/7ab	3576	282	7.89
miR-9/9ab	3347	268	8.01
miR-93/93a/105++++	3668	286	7.80
miR-96/507/1271	3079	231	7.50
miR-99ab/100	258	25	9.69

Full name of miRNA families marked with symbols are:
* miR-125a-5p/125b-5p/351/670/4319
** miR-130ac/301ab/301b/301b-3p/454/721/4295/3666
*** miR-140/140-5p/876-3p/1244
**** miR-15abc/16/16abc/195/322/424/497/1907
# miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d
## miR-219-5p/508/508-3p/4782-3p
### miR-221/222/222ab/1928
#### miR-25/32/92abc/363/363-3p/367
+ miR-30abcdef/30abe-5p/384-5p
++ miR-33a-3p/365/365-3p
+++ miR-34ac/34bc-5p/449abc/449c-5p
++++ miR-93/93a/105/106a/291a-3p/294/295/302abcde

On average, the number of predicted target genes of a miRNA family is around 2700 genes, the number of predicted target genes that are also metabolic genes is around 200 genes, and the percentage of metabolic genes of the predicted targets of a miRNA family is 7.64%.

Table 3.3 Basic statistics of each miRNA cluster.

Cluster number	Number of miRNA families in the cluster	Number of predicted target genes in the cluster (% of all predicted targets)	Number of predicted metabolic genes in the cluster (% of present metabolic target)	Average Jaccard distance in the cluster
Cluster 1	4	1,724 (10.09%)	125 (9.20%)	0.92
Cluster 2	5	4,213 (24.65%)	317 (23.34%)	0.85
Cluster 3	22	12,716 (74.41%)	982 (72.31%)	0.78
Cluster 4	14	12,510 (73.21%)	962 (70.84%)	0.67
Cluster 5	16	12,707 (74.36%)	981 (72.24%)	0.73
Cluster 6	26	15,396 (90.10%)	1205 (88.73%)	0.69

These are number of miRNA families, number of all the predicted target genes present, number of predicted metabolic genes, and average Jaccard distance between the miRNA family pair in each cluster. The numbers in parentheses are percentages of genes in the cluster compared to all predicted targets and percentages of metabolic target genes compared to present metabolic targets.

3.3.2 Hierarchical clustering of miRNA families based on predicted metabolic target genes

After the predicted target genes of all miRNA families were retrieved, a weighted Jaccard distance matrix of all miRNA family pairs was then calculated. The squared Jaccard distance matrix was then subjected to hierarchical clustering by using Ward's minimal variance criterion or Ward's linkage, in order to group the miRNA family that have minimal distance between them. The clustering tree was further processed by a Dynamic Hybrid tree cut algorithm to identify the clusters from the hierarchical clustering results. The resulting clusters are shown in **Figure 3.2**. A heatmap that illustrates the distances between all miRNA family pairs is shown in **Figure 3.3**, with the clusters defined by the Dynamic Hybrid cut tree algorithm as multicolour-shading on top of the heatmap. **Table 3.3** summaries the number of miRNA families in each cluster. **Appendix A1 Table A1.1 – A1.6** list the miRNA familes in clusters 1 through 6. From the **Figure 3.3**, miRNA families in clusters 1 and 2 have very distinct sets of predicted target genes compared to other miRNA families, while miRNA families in clusters 4 and 6 have high proportions of similar predicted target genes within the clusters. According to **Table 3.3**, miRNA families in clusters 3, 4, 5, and 6 have good coverage of metabolic genes, which are around 76% of all metabolic genes present in the database. Despite the high coverage of predicted target genes, some of the genes might not be taking into account in the calculation of the hierarchical clustering because some genes have P_{CT} score of 0, which means that the number of binding sites in the 3'UTR of the genes do not exceed the background number of overall miRNA binding sites of that particular 3'UTR (Friedman et al. 2009). Clusters 1 and 2 cover less than 25% of the metabolic genes present in the database.

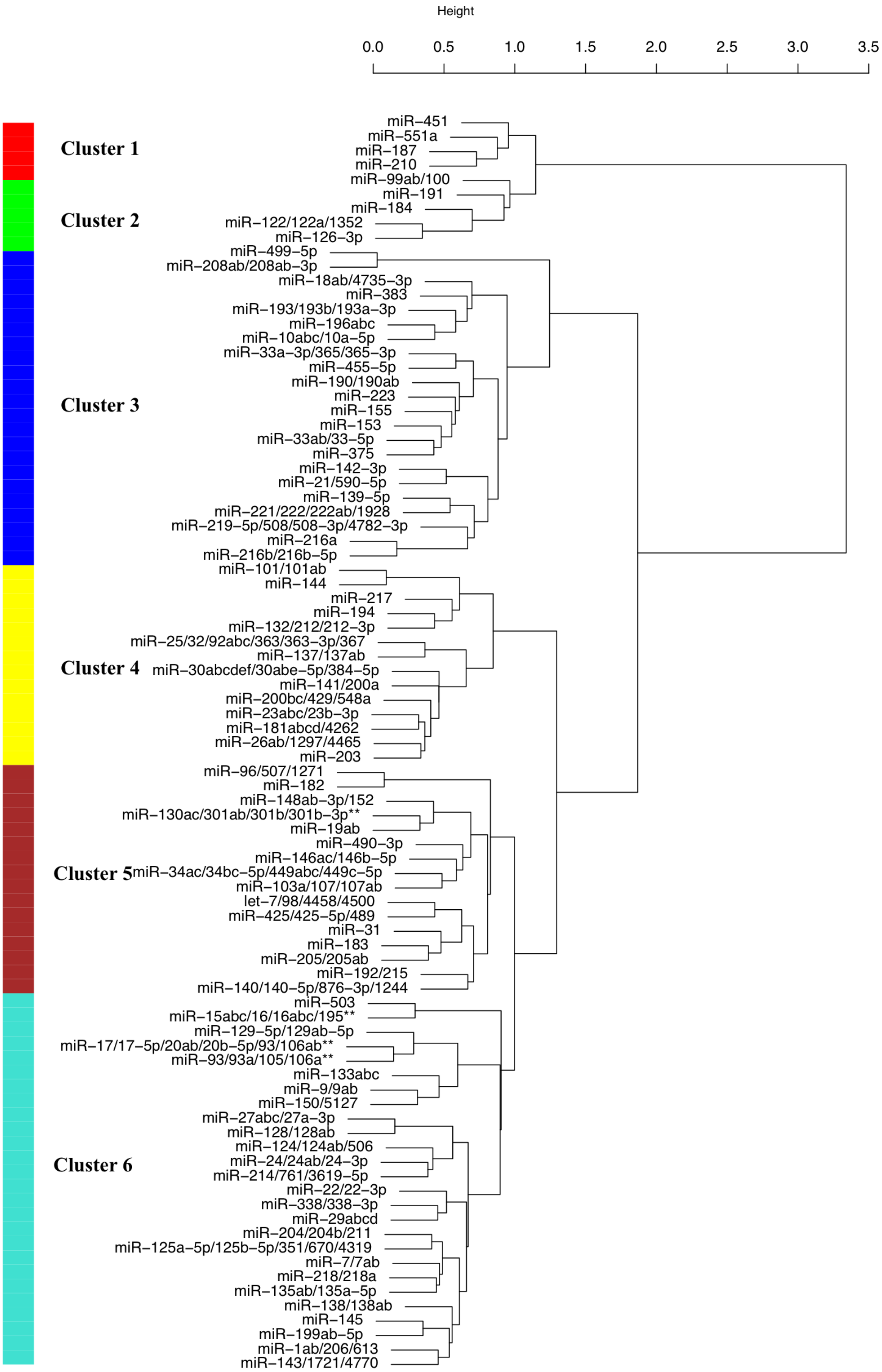


Figure 3.2 Hierarchical clustering of miRNA families based on predicted metabolic target genes results from Jaccard distance matrix using Ward's linkage. The six clusters are identified using the Dynamic Hybrid tree cut algorithm.

** Full miRNA family names are as follow: miR-130ac/301ab/301b/301b-3p/454/721/4295/3666, miR-15abc/16/16abc/195/322/424/497/1907, miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d, and miR-93/93a/105/106a/291a-3p/294/295/302abcde

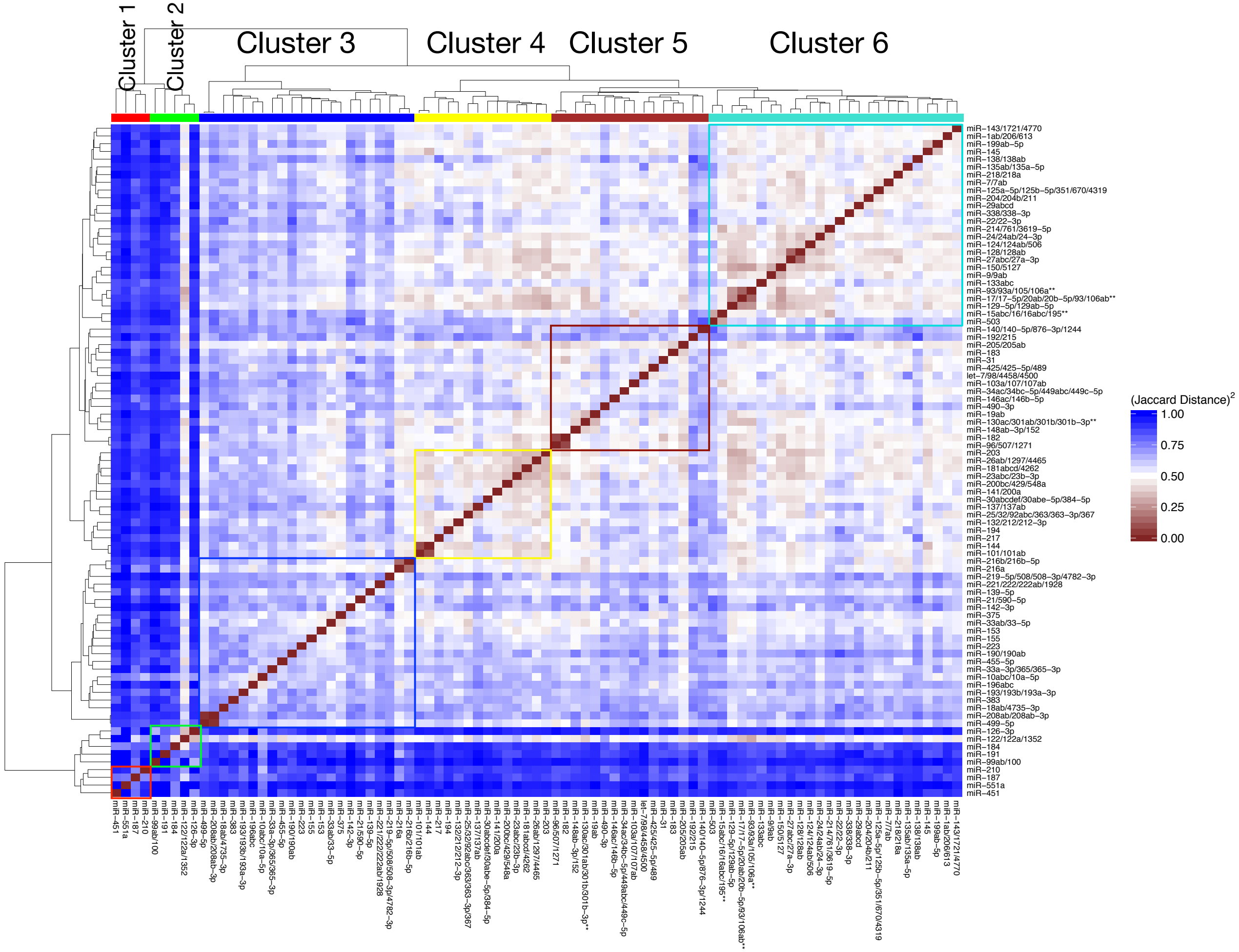


Figure 3.3 The heatmap illustrating squared Jaccard distance between all miRNA family pairs.

Figure 3.3 Dark red box signifies low distance between two miRNA families, i.e. high proportion of overlapped set of predicted target genes. Blue box signifies high distance between two miRNAs families, i.e. low proportion of overlapped set of predicted target genes. miRNA families in clusters 1 and 2 have very distinct sets of predicted target genes compared to other miRNA families in the database. In contrast, miRNA families in clusters 4 and 6 have higher proportions of overlapped predicted target genes within their clusters according to the average Jaccard distance of these two clusters.

3.3.3 KEGG pathway enrichment analysis on the overlapped predicted target genes in each cluster

After the clusters of miRNA families have been defined by hierarchical clustering and Dynamic Hybrid tree cut algorithm, the list of overlapped predicted target genes of each miRNA family pair was subjected to KEGG pathway enrichment analysis using over-representation analysis (ORA). The predicted gene list was tested against all 288 KEGG human pathways, resulting in a list of 288 p-values for each overlapped predicted target genes list between each miRNA family pair. The ORA uses the hypergeometric distribution to calculate probability of obtaining over-represented number of elements. The hypergeometric p-value is calculated using the following equation:

$$p(i \geq x) = 1 - \sum_{i=0}^{x} \frac{\binom{m}{i}\binom{N-m}{K-i}}{\binom{N}{K}} \tag{4}$$

where $p(i \geq x)$ is the cumulative probability of having the number of predicted target genes equal or more than x genes , $\binom{t}{u} = \frac{t!}{u!(t-u)!}$ is a binomial coefficient, m is the number of genes in a testing pathway, K is the number of predicted target genes between the two miRNA families, x is the number of overlapped between sets m and K, and N is the number of all the genes presented in the KEGG database. **Figure 3.4** summarises the formulation of ORA on KEGG pathways. The matrix of p-value lists of all miRNA family pairs is stored in an R object, which can be found in the Supplementary file S1. The p-value list on each miRNA family pairs was subjected to multiple hypothesis testing using the Benjamini-Hochberg false discovery rate (FDR) procedure. The FDR-adjusted p-values of the pathways less then 0.05 cutoff were retained for final analysis.

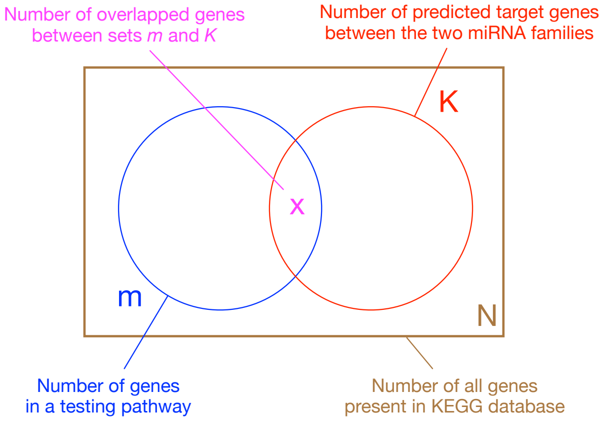


Figure 3.4 Venn diagram illustrating KEGG pathway enrichment formulation.

For the final analysis, the pathways that passed the FDR-adjusted p-values of 0.05 were counted in each cluster separately. The list of pathways that passed the FDR-adjusted p-value cutoff of 0.05 of each cluster with the number of occurrences was produced. The numbers of occurrences of over-represented pathways in were then subjected to another ORA. Again, the hypergeometric p-value of each over-represented pathway can be calculated with the equation 4:

$$p(i \geq x) = 1 - \sum_{i=0}^{x} \frac{\binom{m}{i}\binom{N-m}{K-i}}{\binom{N}{K}} \tag{4}$$

where $p(i \geq x)$ is the cumulative probability of having the number of occurrences of a pathway equal or more than x times , m is the number of occurrences of the pathway in all of the miRNA family pairs, K is the number of occurrences of all the pathways in the cluster of interest, x is the number of overlapped between sets m and K, and N is the number of all occurrences of all the over-represented pathways in all of miRNA family pairs. **Figure 3.5** summarises the formulation of ORA on over-represented KEGG pathways in each cluster.

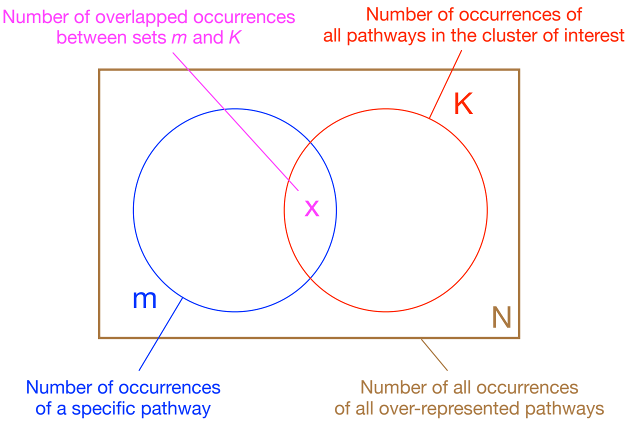

Number of overlapped occurrences between sets *m* and *K*
Number of occurrences of all pathways in the cluster of interest
K
x
m
N
Number of occurrences of a specific pathway
Number of all occurrences of all over-represented pathways

Figure 3.5 Venn diagram illustrating ORA formulation of over-represented KEGG pathway in each cluster.

After the ORA of pathways of all the clusters, the hypergeometric p-values of those pathways in each cluster were again subjected to the multiple hypothesis testing using Benjamini-Hochberg FDR procedure. The full table results of the ORA of all clusters are shown in **Appendix A1 table A1.7-A1.12**. Using the FDR-adjusted p-value cutoff of 0.05, clusters 1, 2, and 5 did not yield any over-represented pathways that occurred more than expected by chance. The ORA of over-represented pathways in clusters 3, 4, and 6 yielded several pathways that are occurred more than expected by chance, which are shown in **Table 3.4, 3.5**, and **3.6**, respectively.

Table 3.4 Over-represented pathways within miRNA family cluster 3.

KEGG Pathway	Number of occurrences	p-value	FDR-adjusted p-value
Purine metabolism	199	2.48E-12	4.76E-10
Phosphatidylinositol signaling system	141	0.0015	0.0244
Morphine addiction	132	6.60E-06	4.22E-04
Salivary secretion	127	0.0006	0.0112
Pancreatic secretion	120	0.0002	0.0058
cAMP signaling pathway	120	0.0006	0.0112
Glycosphingolipid biosynthesis - lacto and neolacto series	104	9.36E-07	8.99E-05
GABAergic synapse	104	0.0004	0.0103
Valine, leucine and isoleucine degradation	102	0.0019	0.0257
Glycosphingolipid biosynthesis - globo series	92	0.0002	0.0056
Pantothenate and CoA biosynthesis	70	0.0001	0.0049
Lysine degradation	68	0.0007	0.0124
Valine, leucine and isoleucine biosynthesis	61	3.68E-05	0.0018
Retinol metabolism	37	0.0034	0.043
alpha-Linolenic acid metabolism	31	0.0018	0.0257

According to **Table 3.4**, over-represented pathways in cluster 3 are involved in essential metabolite biosynthesis and metabolism within the cells. The metabolites that might be associated with the miRNA families in cluster 3 are purine (for DNA synthesis), glycans, amino acids, a fatty acid and vitamin A precursor retinol. This suggests that miRNA families in cluster 3 might have important roles in the regulation of these metabolites production/degradation.

Table 3.5 Over-represented pathways within miRNA family cluster 4.

KEGG Pathway	Number of occurrences	p-value	FDR-adjusted p-value
Peroxisome	58	3.02E-05	0.0007
Vibrio cholerae infection	57	1.50E-07	1.02E-05
Valine, leucine and isoleucine biosynthesis	51	0.0031	0.0248
Glycolysis / Gluconeogenesis#	51	0.0052	0.0319
Pyruvate metabolism#	48	1.55E-07	1.02E-05
Protein digestion and absorption	48	6.53E-05	0.0011
Synaptic vesicle cycle	47	1.59E-05	0.0004
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	46	6.31E-05	0.0011
Endometrial cancer*	44	3.21E-06	0.0001
Proximal tubule bicarbonate reclamation	42	0.0042	0.0299
Oxidative phosphorylation#	42	0.0049	0.031
mTOR signaling pathway#	40	4.95E-05	0.0011
Glioma*	39	0.0014	0.013
Melanoma*	37	0.0002	0.003
Collecting duct acid secretion	36	5.45E-05	0.0011
Starch and sucrose metabolism	36	0.001	0.0112
Small cell lung cancer*	34	0.001	0.0112
Fat digestion and absorption	34	0.0045	0.0301
Adipocytokine signaling pathway	33	0.002	0.0178
FoxO signaling pathway#	30	0.0061	0.0366
MicroRNAs in cancer*	28	2.91E-11	5.74E-09
Vitamin digestion and absorption	28	0.0013	0.013
Citrate cycle (TCA cycle)#	25	3.41E-07	1.68E-05
Tyrosine metabolism	25	0.0003	0.0038
Phagosome	25	0.0041	0.0296
Histidine metabolism	22	0.009	0.0491
Chemical carcinogenesis*	20	0.0014	0.013
Serotonergic synapse	20	0.0067	0.0389
Focal adhesion	15	0.0003	0.0043
Glutathione metabolism	14	0.0006	0.007
Pathways in cancer*	13	0.0023	0.0194
Phenylalanine metabolism	12	0.007	0.0391
Primary bile acid biosynthesis	11	0.0046	0.0301
Protein export	10	5.14E-06	0.0002
PI3K-Akt signaling pathway#	9	0.003	0.0248
Other glycan degradation	5	0.0035	0.0266

Asterisks (*) marked the cancer-specific pathways. Hashes (#) marked cancer-related pathways.

Table 3.6 Over-represented pathways within miRNA family cluster 6.

KEGG Pathway	Number of occurrences	p-value	FDR-adjusted p-value
Carbohydrate digestion and absorption	222	0.0024	0.0099
Sphingolipid metabolism	220	8.43E-05	0.0005
Lysine degradation	210	0.0055	0.0191
AMPK signaling pathway	205	1.02E-06	1.16E-05
Mineral absorption	203	0.0013	0.0058
Chemokine signaling pathway	185	0.0064	0.0215
Aldosterone-regulated sodium reabsorption	168	8.64E-06	6.59E-05
Type II diabetes mellitus	161	1.56E-16	3.22E-14
HIF-1 signaling pathway#	148	9.99E-05	0.0005
HTLV-I infection	146	0.0011	0.0052
Insulin signaling pathway	145	2.91E-12	2.00E-10
Glycosaminoglycan biosynthesis - keratan sulfate	145	1.28E-06	1.32E-05
Butanoate metabolism	144	6.41E-15	6.60E-13
beta-Alanine metabolism	140	7.92E-06	6.27E-05
Tryptophan metabolism	137	4.03E-05	0.0002
Drug metabolism - other enzymes	134	0.0096	0.0309
Regulation of actin cytoskeleton	120	5.74E-12	2.96E-10
Fructose and mannose metabolism	119	0.0015	0.0063
VEGF signaling pathway#	116	0.0013	0.0055
Glioma*	116	0.0129	0.0385
Fc gamma R-mediated phagocytosis	115	0.0112	0.0343
Endometrial cancer*	110	0.0022	0.0092
Vitamin digestion and absorption	109	5.70E-10	1.07E-08
Fc epsilon RI signaling pathway	107	1.53E-06	1.43E-05
Non-alcoholic fatty liver disease (NAFLD)	107	8.69E-05	0.0005
mTOR signaling pathway#	107	0.0034	0.0133
Non-small cell lung cancer*	106	1.16E-07	1.70E-06
Prolactin signaling pathway	105	1.93E-11	7.94E-10
Renal cell carcinoma*	105	1.37E-08	2.34E-07
Pancreatic cancer*	105	1.13E-06	1.23E-05
Propanoate metabolism	105	2.48E-05	0.0002
Melanoma*	103	0.0038	0.0141
B cell receptor signaling pathway	100	5.59E-10	1.07E-08
Pentose phosphate pathway	100	2.91E-05	0.0002
Acute myeloid leukemia*	96	1.89E-10	5.10E-09
Colorectal cancer*	95	2.53E-10	5.80E-09
Bacterial invasion of epithelial cells	94	1.37E-10	4.71E-09
Chronic myeloid leukemia*	94	1.98E-10	5.10E-09
Glycine, serine and threonine metabolism	86	4.12E-08	6.53E-07
Vitamin B6 metabolism	86	3.23E-07	3.91E-06
ErbB signaling pathway#	84	4.18E-05	0.0002
Prostate cancer*	84	0.0119	0.0361
Lysine biosynthesis	83	4.44E-06	3.81E-05
Tyrosine metabolism	82	1.98E-07	2.72E-06
Epstein-Barr virus infection	82	3.96E-06	3.54E-05
Lysosome	81	0.0028	0.011
Apoptosis#	80	2.20E-07	2.84E-06
Fatty acid biosynthesis	75	1.84E-05	0.0001
Histidine metabolism	73	0.0012	0.0055
Galactose metabolism	70	0.0096	0.0309

KEGG Pathway	Number of occurrences	p-value	FDR-adjusted p-value
Ras signaling pathway#	66	0.0005	0.0024
ABC transporters	58	0.0075	0.0251
T cell receptor signaling pathway	58	0.0102	0.0322
TNF signaling pathway#	57	0.0049	0.0174
Caffeine metabolism	56	4.63E-06	3.82E-05
Toll-like receptor signaling pathway	55	0.0007	0.0031
Phenylalanine, tyrosine and tryptophan biosynthesis	47	5.21E-05	0.0003
Butirosin and neomycin biosynthesis	44	0.0035	0.0134
Phenylalanine metabolism	36	0.0011	0.0052
Ascorbate and aldarate metabolism	33	0.0037	0.014
Focal adhesion#	32	0.0161	0.0473
Proteoglycans in cancer*	28	1.44E-06	1.42E-05
Jak-STAT signaling pathway#	27	0.0006	0.0028
Influenza A	20	0.0003	0.0016
Viral carcinogenesis*	19	4.10E-05	0.0002
Terpenoid backbone biosynthesis	17	3.50E-05	0.0002
Folate biosynthesis	14	0.0108	0.0338
Phototransduction	11	0.0046	0.0168
Nicotine addiction	8	3.22E-05	0.0002
Ubiquinone and other terpenoid-quinone biosynthesis	7	0.0063	0.0215

Asterisks (*) marked the cancer-specific pathways. Hashes (#) marked cancer-related pathways.

Table 3.5 and **Table 3.6** showed that over-represented pathways in miRNA family clusters 4 and 6 have several pathways that are cancer-specific, such as glioma, melanoma, lung cancer etc., suggesting that the miRNA families in these two clusters are having important roles in regulating the genes involving in these cancers. Over-represented pathways in clusters 4 and 6 also have several pathways that are cancer-related, i.e. pathways that are deregulated in cancer cells, such as mTOR signaling pathway, PI3K-Akt signaling pathway, Citrate cycle (TCA cycle), VEGF signaling pathway etc. **Figure 3.6** shows over-represented pathways from clusters 4 and 6 that are cancer-specific and cancer-related based on KEGG pathway diagram.

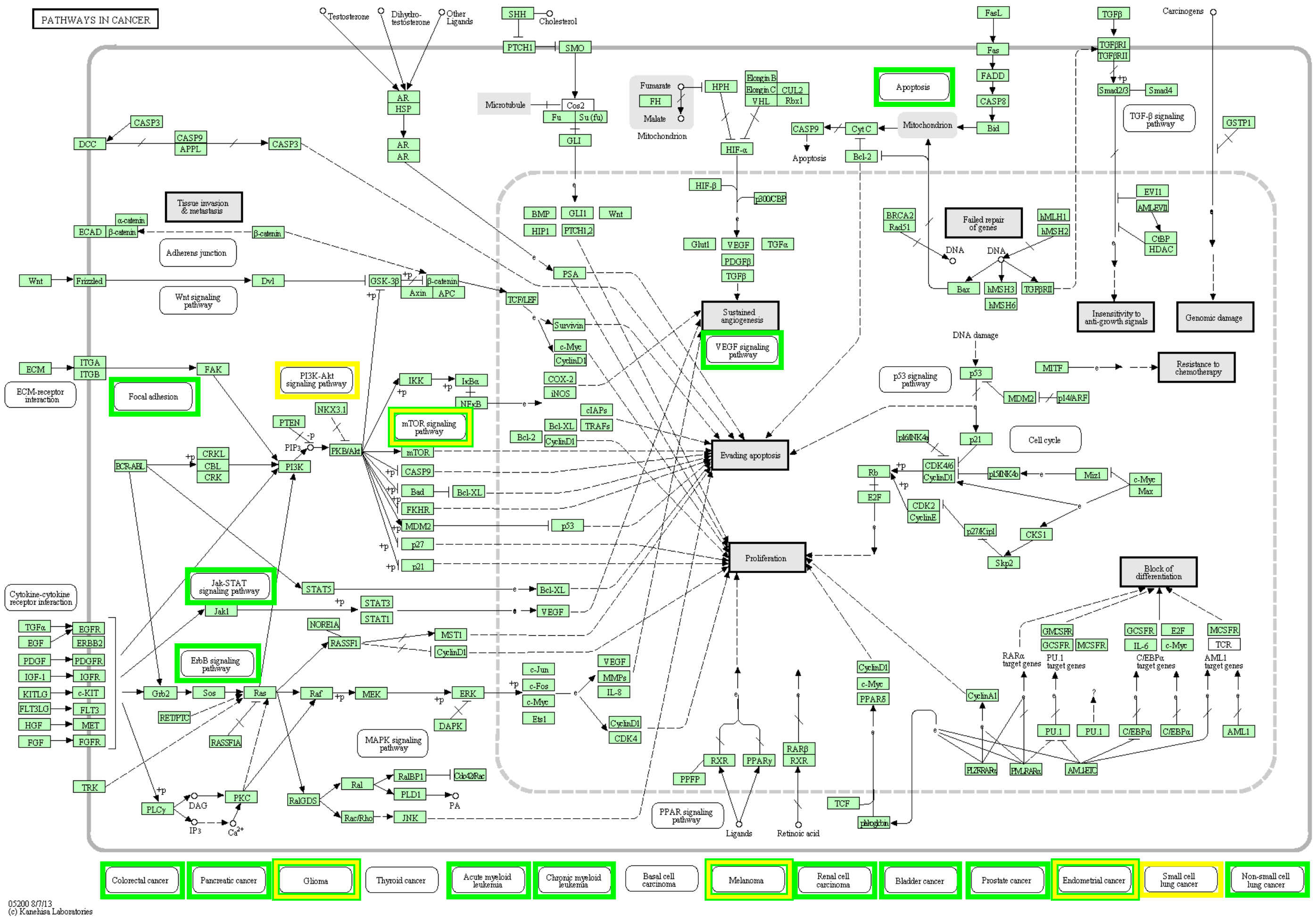


Figure 3.6 Over-represented pathways from clusters 4 and 6 that are cancer-specific and cancer-related based on KEGG pathways in cancer diagram (figure legend continues on the next page).

Figure 3.6 Over-represented pathway names from clusters 4 and 6 are in yellow and green boxes, respectively. Pathways that are presented in both clusters are in boxes with green outer-layer and yellow inner-layer. *Ras* and *HIF-1α* are also present in the pathways in cancer but do not clearly labeled as separate pathways. *Ras* is downstream of ErbB signaling pathway, whereas *HIF-1α* is upstream of VEGF signaling pathway. Note that microRNAs in cancer, viral carcinogenesis, TNF signaling pathway, and chemical carcinogenesis do not present in the figure. The figure was retrieved from www.kegg.jp on 24 August 2014.

3.3.4 Experimentally validated miRNAs in the three cancer types identified in each miRNA cluster

Experimentally validated miRNAs associated with the three cancer types highlighted in the over-represented pathways were identified and marked in the tables with miRNAs in clusters 4 and 6. Three cancer types were identified to be regulated by both miRNA in clusters 4 and 6, i.e. lung cancer (small-cell in cluster 4 and non small-cell in cluster 6), melanoma, and glioma. At least five review articles with the sets of miRNAs that are involved in these cancer types were examined to identify which miRNAs in the clusters 4 and 6 are implicated in these cancer types. Lists of miRNA families in clusters 4 and 6 that are implicated in the three types of cancer are shown in **Table 3.7** and **Table 3.8**. These results show that the approach in this study can correctly identify groups of miRNAs that are implicated in these diseases. However, there are miRNAs that are involved in these three cancer types but not presented in these two clusters, which are let-7 family, miR-10a/10b, miR-34a, miR-21, miR-210 and miR-221/222. It is possible that these miRNAs regulate different types of target genes compared to the miRNAs in clusters 4 and 6.

3.4 Discussion

The analysis presented here can help narrow down the scope of the miRNA network that work together as the same units of transcriptional regulation. It is widely known that multiple miRNAs, especially miRNAs in the same cluster, can target mRNAs that are functionally related [see review by Peter (2010) and Hausser & Zavolan

(2014)], therefore this analysis built up from the sets of common predicted target genes between miRNA family pairs, and then used that information to classify miRNA families into clusters based on the distances between them.

Table 3.7 miRNA families of cluster 4 that are found in three cancer types.

miRNA family in cluster 4	Lung cancer					Melanoma						Glioma				
	Kang & Lee (2014)	Guz et al. (2014)	Gao et al. (2014)	Zhang et al. (2014)	Ulivi & Zoli (2014)	Aftab et al. (2014)	Sun et al. (2014)	Luo et al. (2014)	Bennett et al. (2013)	Völler et al. (2013)	Segura et al. (2012)	Brower et al. (2014)	Zhao et al. (2014)	Palumbo et al. (2014)	Zhi et al. (2013)	Hermansen et al. (2013)
miR-23abc/23b-3p	x			x	x							x	x			x
miR-25/32/92abc/363/363-3p/367									x	x		x		x	x	x
miR-26ab/1297/4465	x		x			x			x							x
miR-30abcdef/30abe-5p/384-5p					x	x		x		x	x			x	x	x
miR-101/101ab		x	x	x				x				x		x		x
miR-132/212/212-3p												x				x
miR-137/137ab		x				x	x	x	x	x	x	x	x			
miR-141/200a		x		x		x		x		x	x					
miR-144																
miR-181abcd/4262			x									x		x		x
miR-194		x														
miR-200bc/429/548a		x	x		x	x	x	x		x	x					x
miR-203						x	x	x	x	x						
miR-217																

Some of the findings in this chapter also support the results of the previous chapter on miR-22. Two pathways, which are glycine, serine and threonine metabolism and fatty acid biosynthesis pathways, were identified by KEGG pathway enrichment analysis (**Table 3.6**). These two pathways have two metabolic genes, i.e. *MTHFD2* and *ELVOL6*, which are under the regulation of miR-22, a member in cluster 6. Another miRNA that targets *MTHFD2* is miR-9 (Selcuklu et al. 2012), which was also classified to be a member in miRNA cluster 6.

Table 3.8 miRNA families of cluster 6 that are found in three cancer types.

miRNA families in cluster 6	Lung cancer					Melanoma						Glioma				
	Kang & Lee (2014)	Guz et al. (2014)	Gao et al. (2014)	Zhang et al. (2014)	Ulivi & Zoli (2014)	Aftab et al. (2014)	Sun et al. (2014)	Luo et al. (2014)	Bennett et al. (2013)	Völler et al. (2013)	Segura et al. (2012)	Brower et al. (2014)	Zhao et al. (2014)	Palumbo et al. (2014)	Zhi et al. (2013)	Hermansen et al. (2013)
miR-1ab/206/613	x				x											
miR-7/7ab			x						x			x				x
miR-9/9ab			x			x	x		x	x	x	x	x		x	x
miR-15abc/16/16abc/195*	x	x				x	x	x	x		x	x		x		x
miR-17/17-5p/20ab/20b-5p**	x			x	x				x			x	x	x	x	x
miR-22/22-3p									x							x
miR-24/24ab/24-3p																
miR-27abc/27a-3p														x		
miR-29abcd						x	x	x		x	x	x				x
miR-93/93a/105/106a/291a-3p$^{\#}$	x			x	x				x			x	x	x	x	x
miR-124/124ab/506								x				x	x		x	x
miR-125a-5p/125b-5p/351/670/4319				x	x			x	x			x	x	x	x	x
miR-128/128ab					x							x	x		x	x
miR-129-5p/129ab-5p		x														x
miR-133abc					x							x				
miR-135ab/135a-5p												x		x	x	x
miR-138/138ab												x				x
miR-143/1721/4770					x								x			
miR-145						x	x				x	x	x	x		
miR-150/5127		x			x				x							
miR-199ab-5p						x			x	x	x			x	x	
miR-204/204b/211						x	x	x	x	x	x		x	x		x
miR-214/761/3619-5p						x		x	x	x	x					x
miR-218/218a												x	x	x		x
miR-338/338-3p																x
miR-503																

Full family names
* miR-15abc/16/16abc/195/322/424/497/1907
** miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d
$^{\#}$ miR-93/93a/105/106a/291a-3p/294/295/302abcde/372/373/428/519a/520be/520acd-3p/1378/1420ac

Cluster 6 also has very high within-cluster similarity, and also has the highest number of predicted metabolic target genes of miRNAs (**Table 3.3** and **Table 3.6**). This cluster also contains two polycistronic miRNA clusters, which are miR-24-27a cluster and miR-17-92 cluster. It is widely known that miR-17-92 cluster has the same seed sequence. Hence, the predicted target gene pool of this miRNA cluster would be very similar. On the other hand, miR-24 and miR-27 families have very different seed sequences (miR-24 family: GGCUCAG, miR-27 family: UCACAGU), and it is interesting that these two miRNA families would have similar pool of predicted metabolic target genes and be clustered together. These two polycistronic miRNA clusters were also known to regulate several metabolic pathways (Dumortier et al. 2013).

One drawback of this analysis is that the list of metabolic genes used in this study might not be a complete one. Therefore, pathways that contain the missing metabolic genes from the list would not be identified by this approach. This problem will be resolved once more data, i.e. metabolic genes list, are discovered. Another drawback is that the analysis did not have any experimental validation to back up the findings. A simple correlation calculation might be useful to a certain extent, but might be misleading as well since not all the target genes of miRNAs have high anti-correlation between them. As for the cancer-specific pathways such as melanoma or glioma, etc., miRNAs and mRNA profiles from large-scale studies can be a good starting point for validating the findings. Resources such as The Cancer Genome Atlas (TCGA) database, which has miRNA and mRNA profiling data on these cancer, or Cancer Cell Line Encyclopedia (CCLE) database, which has mRNA expression where some of those mRNAs can be used as miRNA surrogate measurement, can be used to screen for the possibilities of associations.

The conditions under which the hypergeometric test was applied in the last two steps is violating the assumption of variable independence and might exaggerate the hypergeometric p-values. Although a p-value adjustment procedure was carried out, the effect of p-value over- or underestimation might still persist. One remedy for this is to set more stringent p-value cutoff, such as from 0.05 to 0.01 or to 0.001, or implanting more stringent multiple hypothesis testing procedure, such as the

Bonferroni adjustment. Another way to address the p-value inaccuracy is to perform permutation test, by reshuffle the names of the target genes in each miRNA family and then repeat the analysis until the desired statistically significant level is reached.

There is another study on miRNA clustering, called miRConnect, which uses sums of Pearson's correlation coefficients between miRNA and mRNA expression profiles as the criteria for miRNA clustering (Hua et al. 2011). Although the clusters in miRConnect study are correctly identified (according to the next chapter of this thesis), the way they classified the miRNA clusters are biased by using the correlation coefficients between the miRNA and mRNA, as it was shown in the previous chapter that the correlation between miRNA and its target genes might not always be high (i.e. miR-22 and ACLY). Therefore, some of the members in their clusters will be wrongly classified into the clusters (as will be shown later in the next chapter). Also, some of the miRNAs in the miRConnect study can be grouped by their seed sequences, which made some of the miRNAs in the same clusters redundant in terms of the target genes. The approach in this study does not suffer from the problem previously mentioned because the miRNAs were already grouped in terms of their seed sequences. The use of probability of conserved targeting (or P_{CT} score) for the calculation of the clusters in this study is less biased than correlation coefficients between miRNA and mRNA, which sometimes proved to be misleading. However, the P_{CT} score in itself does not include all the information that are necessary for the miRNA-mRNA regulation, such as minimum free energy of secondary structure of miRNA-mRNA duplex (Rehmsmeier et al. 2004; Kertesz et al. 2007), or a combination between the seed sequence complementarity and free energy of the miRNA-mRNA duplex (Krek et al. 2005). Integration of these metrics into the

Jaccard distance calculation might improve the accuracy of the miRNA family cluster classification.

One important limitation was that this study used only highly conserved predicted target genes in the TargetScan database. The latest TargetScan database disregards three features that might be important in targeting of genes by miRNAs, which are site accessibility, minimum free energy of miRNA:mRNA duplex, and target site abundance. By including these features into the algorithm, predicted target genes in the TargetScan database might be more sensitive and yield more true positive predicted target genes (while sacrificing the precision of the prediction in the process). Increasing sensitivity of the database used can be achieved through union between different databases, which would in a way take into account the features that different algorithms did not use in the individual algorithms. However, a review of several miRNA prediction software found that the benefit of union or intersect the databases for sensitivity did not yield good return of the precision (Alexiou et al. 2009). Therefore, a novel approach that intrinsically combines probability of conserved targeting of TargetScan database and other features from other miRNA prediction algorithms could be immensely beneficial to this analysis in the future.

One improvement that can be implemented in this study is to include more pathways or gene sets information from different sources or databases in order to expand the pathways or gene sets that might be over-represented by the metabolic gene sets used in this study. Another possible tweak is the list of genes used as the constraints for clustering. This study used metabolic gene list as the constraints for the clustering process. If different sets of genes are applied as the constraints, along with pathways that are linked with the gene list, for example list of genes and pathways involved in cholesterol metabolism, this approach could be used to identify groups of miRNAs that linked to the cholesterol metabolism and the diseases linked to these processes. The scientific community is also constantly updating validated target genes of miRNAs. Therefore, manually curating the validated metabolic target genes in the analysis might add more accuracy of cluster identification and the enriched pathway in each cluster.

Chapter 4 : Over-Representation of Correlation Analysis (ORCA)

4.1 Introduction

High-throughput, high-dimensional biological data analysis is becoming a routine task for most biological and biomedical researchers. Several layers of measurements, comprising gene (either non-coding or protein-coding) expression, protein profiles, and metabolite concentrations are overwhelmingly generated in massive quantity, sometimes simultaneously. The rationale behind the taking of these measurements together is to better understanding complex diseases, such as cancer or metabolic disorders like diabetes, where these biomolecules are often altered in concert. In the past, two conventional analyses, correlation analysis and pathway analysis, were considered the standard steps in analysing coordinated responses in the biomolecular profiles of those complex diseases. Correlation analysis (and its multivariate extensions such as principal component analysis, or PCA, and canonical correlation analysis) focuses on the estimation of quantitative explanatory power of one variable, which could be any gene, miRNA, or protein measurement, over another variable to infer an association between those two variables quantitatively. By contrast, pathway analysis has the objective of exploiting prior knowledge on the interactions or related sets of biomolecules previously identified and then test specific hypotheses whether or not the relationship between a particular set of biomolecules (either a specific pathway or a gene set) and a given experimental condition exists. Conventional pathway analysis uses the over-representation analysis (ORA) to determine that a relationship between the sets in a specific group of variables exists. The first implementation of ORA in biological data analysis was by Tavazoie et al. (1999), where transcriptional regulatory sub-networks in yeast were identified using data from gene expression microarray profiling. Since the first biological implementation, ORA has been used to perform pathway analysis based on pathway information such as those from Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto 2000) or from consolidated pathway database (Cavill et al. 2011), or to infer biological categories (molecular functions, cellular compartments, etc.) that were deemed important from two different experimental conditions in Gene Ontology (GO) (Zeeberg et al. 2003; Beissbarth & Speed 2004).

One problem with pathway analysis using the ORA is that after the variables have been determined as differentially expressed between different conditions the measurements of those variables will be disregarded. A novel method introduced here called Over-Representation of Correlation Analysis (ORCA) provides a new way to test for a high prevalence of informative associations between two sets of variables, whether the variables are mRNAs, miRNAs, proteins, or metabolites. The set of variables could be pathways, gene sets, or any groupings that are deemed biologically important. This study also provides a means to determine the threshold of correlation coefficient for each data set based on Shannon's information entropy principle. The work in this chapter has recently been published (Pomyen et al. 2014) and the manuscript is included as an appendix to this thesis (**Appendix A2**).

4.2 Materials and methods

4.2.1 Over-representation of Correlation analysis method outline

ORCA combines two conventional statistical analyses together, which are correlation analysis and ORA. For this method development, the correlation metric used was Spearman's rank correlation. The reason for using Spearman's rank correlation is to avoid the assumption of a normal distribution in the data set. The ORA is performed after the correlation coefficient calculation by using the hypergeometric test to calculate the probability of getting the number of correlation coefficients that are higher than expected by chance in a certain group. There are two prerequisites for the data sets in order to apply ORCA: 1) the variables must be divided into sets according to relevant criteria, such as KEGG pathways or GO annotations where the variables are mRNA genes, or groups resulting from unsupervised classification techniques such as hierarchical clustering; and 2) the number of data points in each variable must be adequate in order to accurately calculate the correlation coefficient. The schematics of the method is shown in **Figure 4.1**. The method starts by calculating the correlation matrix of all *n* variables, resulting in an *n* x *n* matrix of correlation coefficients between every possible pair of variables. The variables in the correlation matrix are then sorted into groups or sets according to relevant criteria, such as pathway information or GO annotation. There are two types of correlation coefficients in the correlation matrix: the first type is the within-group correlation, which is the correlation coefficients between variables within the same group or set; and the second type is the between-group correlation, which is the correlation coefficients between variables from different groups. The correlation coefficients are labels as above or below a certain correlation coefficient threshold, which can be empirically chosen or calculated by the threshold selection method, which will be explained in detail in section 4.2.2. **Table 4.1** illustrates four categories of the correlation coefficients in the ORCA. Finally, to determine if the association between groups is statistically significant, ORCA is applied. The p-value of obtaining a certain number of correlation coefficients that passes the threshold and also presents in the between-group correlations can be calculated by a cumulative hypergeometric distribution:

$$p(n > k) = 1 - \sum_{i=0}^{k} \frac{\binom{M}{i}\binom{N-M}{X-i}}{\binom{N}{X}} \qquad (1)$$

where $\binom{t}{u} = \frac{t!}{u!(t-u)!}$ is the binomial coefficient, *M* is the number of between-group correlation coefficients, *N* is the number of all correlation coefficients in the correlation matrix (excluding all the within-group correlation), *X* is the number of all between-group correlation coefficients that pass the threshold, and *k* is the number of correlation coefficients in the between-group pair of interest that pass the threshold. The within-group correlation is excluded because the correlation coefficients between the variables in the same group are usually very high compared to the between-group correlation. This high correlation will lead to an underestimation of the p-values of the between-group correlations. The p-values of the within-group correlation, therefore, are separately calculated using the same equation, with the inclusion of the within-group correlations in the variables *N*, *M*, and *X*.

4.2.2 Threshold selection

A correlation threshold selection method was developed in order to objectively select a correlation coefficient threshold that provides the highest amount of information from a given data set. This threshold selection method is based on Shannon's entropy (Shannon 1948), where the correlation coefficient threshold that provides the most contrast in a data set yields the highest amount of information. In order to calculate an entropy-based score for correlation coefficient threshold, ORCA is performed on every group-pairs in a given data set on every correlation threshold from 0.01 to 1.00 with an increment of 0.01. Therefore, there are 100 p-value lists for a given data set. Each p-value list is then subjected to the following equation:

$$H(X) = -\sum_{i=1}^{J} p(x_i) \ln p(x_i) \tag{2}$$

where *H*(*X*) is a Shannon's entropy-like score, $p(x_i)$ is a p-value of a within- or between group association between a group-pair *i* calculated by hypergeometric distribution, J is the number of all within- and between-group pairs (where the number of groups of the data set is n then $J = n(n\text{-}1)/2 + n$), and ln is the natural logarithm (i.e. $\log_e$). The correlation coefficient threshold that yields the highest Shannon's entropy-like score is used as the threshold for ORCA.

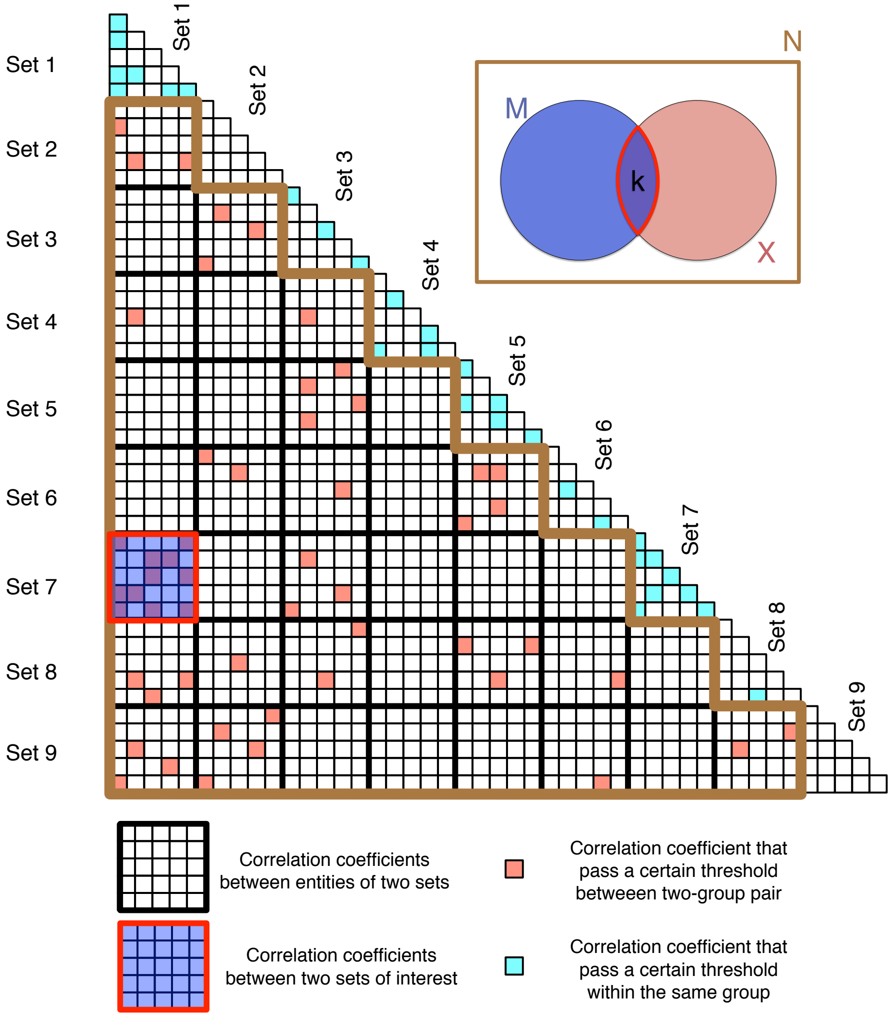

Set 1
Set 2
Set 3
Set 4
Set 5
Set 6
Set 7
Set 8
Set 9
N
M
k
X
Correlation coefficients between entities of two sets
Correlation coefficient that pass a certain threshold betweeen two-group pair
Correlation coefficients between two sets of interest
Correlation coefficient that pass a certain threshold within the same group

Figure 4.1 Concept of Over-Representation of Correlation Analysis. A lower half of an example correlation matrix is shown here with two types of correlation coefficients, labeled with different colours in the figure. The white, turquoise and pink boxes represent correlation coefficients between variables in the data set. The boundaries between groups are bold black lines. The Venn diagram labeled with variable names is corresponding to the equation 1. The colours in the Venn diagram are also corresponding to the types of correlation coefficients in the example correlation matrix. The variables are further explained in the Table 4.1.

Table 4.1 Contingency table with variables correspond to the equation 1 and figure 4.1.

		Correlation Coefficients in sets 1 & 7		
		Yes	No	Total
Correlation coefficients that pass a certain threshold	Yes	k	X – k	X
	No	M – k	N – (M + X – k)	N – k
	Total	M	N – M	N

The variables in rows represent the number of correlation coefficients that are pass/not pass a certain correlation coefficient threshold. The variables in columns represent the number of correlation coefficients that are in/out of the set-pair 1 & 7. According to equation 1, *M* is the number of correlation coefficients in the between set-pair 1 & 7, *X* is the number of all correlation coefficients between every set-pairs that pass the certain threshold, *k* is the number of correlation coefficients in the between set-pair 1 & 7 that also pass the certain threshold, and *N* is the number of all correlation coefficients of all between set-pairs (excluding the correlation coefficients between variables within the same sets). The calculation of hypergeometric p-values in the within set, the number of all correlation coefficients, including all the correlation coefficients between variables within the same sets, are added into *X* and *N*.

4.2.3 Permutation analysis

Because the variables in the correlation matrix of a given data set are not usually independent, permutation analysis was used to calculate the null distribution of the p-values obtained by the hypergeometric distribution for a given data set. To achieve this, the exclusive group membership for each variable was permuted, however, the group structure, i.e. the number of groups and the number of members in the groups, along with the correlation coefficients between the variables was retained. One million permutations were performed for each data set. The empirical p-values for each within- and between-groups were calculated by using this equation (Davison & Hinkley 1997):

$$P = \frac{(r+1)}{(n+1)} \tag{3}$$

where *p* is the empirical p-value of each within- or between-group, *r* is the number of times in which the hypergeometric p-values from permutation test are equal or less than the actual p-values from ORCA on the original data matrix, and *n* is the number of permutation used in the test (which is one million permutations).

4.2.4 Data sets

4.2.4.1 Drug sensitivity data set

The drug sensitivity profiles of 58 cell lines from National Cancer Institute cell line panel (the NCI-60 panel) were used (Scherf et al. 2000). The values represent the sensitivity to each drug in each cell line is called the GI_{50}, which is the concentration of a drug that can inhibit the growth of a cell line to 50% of the original growth rate of the cell line without the drug. Therefore, each drug has 58 GI_{50} values for every cell lines in the panel. The drugs can be grouped by using their mechanisms of actions, i.e. the molecular targets of the drugs. In total, 116 chemotherapeutic drugs with nine different mechanisms of actions were chosen from the Developmental Therapeutics Program (DTP) database. The compound names, NSC (National Service center) numbers and their mechanisms of actions, were shown in **Table 4.2**. Overall of the drug sensitivity data in the form of summed of GI_{50} for each compound is shown in **Figure 4.2**.

4.2.4.2 miRNA data sets

Two independent miRNA data sets were used in this study. The data sets are the baseline miRNA profiles of the cell lines from NCI-60 cell panel. The two data sets were from Liu *et al.* (2010), which used G4470B design ID 019118 miRNA microarray chips from Agilent Technologies and Søkilde *et al.* (2011), which used LNA (Locked Nucleic Acid)-enhanced miRCURY Dx 9.2 microarray platform from Exiqon. The overlapped number of miRNAs between the two data sets is 124 miRNAs, and the list is shown in **Table 4.3**. The groupings of the miRNAs used in this study were the miRNA clusters from miRConnect (Hua et al. 2011), which divided the miRNAs into 13 clusters according to the correlation between mRNA and miRNA profiles. Briefly, the summed correlation coefficients between mRNAs and each miRNA were ranked and the top 2000 genes on each direction (i.e. positive and negative) were selected to create a gene list for each miRNA. The number of overlapped genes of every miRNA pair was calculated, resulting in a 136 x 136 matrix. Finally, the overlapped gene list matrix was used for hierarchical clustering. Overall of the two miRNA data sets in the form of summed expression for each miRNA are shown in **Figure 4.3** and **Figure 4.4** for Liu *et al.* (2010) and Søkilde *et al.* (2011), respectively.

Table 4.2 NSC numbers and compound names of 116 chemotherapeutic drugs by their mechanisms of actions.

NSC	Chemical Name	Mechanisms of actions
56410	Porfiromycin	1) Alkylating agents
26980	Mytomycin C	
132313	Dianhydrogalactitol	
363812	Tetraplatin	
73754	Fluorodopan	
6396	Tris(aziridinyl)phosphine sulfide	
329680	Hepsulfam	
344007	Piperazine mustard	
135758	Piperazinedione	
241240	Carboplatin	
762	Nitrogen mustard hydrochloride	
119875	cis-Diamminedichloroplatinum(II) (cisplatin)	
34462	Uracil mustard	
8806	Melphalan	
271674	Diaminocyclohexyl-Pt-II	
25154	Pipobroman	
182986	Diaziridinylbenzoquinone	
296934	Teroxirone	
3088	Chlorambucil	
167780	Asaley	
9706	Triethylenemelamine	
172112	Spiromustine	
750	Busulfan	
348948	Cyclodisone	
102627	Yoshi-864	
256927	Iproplatin	
95441	Semustine (MeCCNU)	
353451	Mitozolamide	
338947	Clomesone	
409962	Carmustine (BCNU)	
79037	Lomustine (CCNU)	
178248	Chlorozotocin	
95466	PCNU	
139105	Baker's-soluble-antifolate	2) Antifols
184692	Aminopterin-derivative	
740	Methotrexate	
132483	Aminopterin	
623017	an-antifol	
633713	an-antifol	
134033	Aminopterin-derivative	
174121	Methotrexate-derivative	
352122	Trimetrexate	
366140	Pyrazoloacridine	3) DNA binder
354646	Morpholino-adriamycin	
142982	Hycanthone methanesulfonate	
268242	N-N-Dibenzyl-daunomycin	
357704	Cyanomorpholinodoxorubicin	
102816	Azacytidine	4) DNA incorporation
71851	alpha-2'-Deoxythioguanosine	
264880	5-6-Dihydro-5-azacytidine	
752	Thioguanine	
71261	beta-2'-Deoxythioguanosine	
145668	Cyclocytidine	5) DNA synthesis inhibitors
303812	Aphidicolin-glycinate	
63878	Cytarabine hydrochloride	
27640	5-fluorodeoxyuridine	
755	Thiopurine (6MP)	
19893	5-fluorouracil	
148958	Ftorafur	
32065	Hydroxyurea	6) Ribonucleotide reductase inhibitors
1895	Guanazole	
51143	Pyrazoloimidazole	
143095	Pyrazofurin	7) RNA synthesis inhibitors
153353	L-Alanosine	
163501	Acivicin	
224131	N-phosphonoacetyl-L-aspartic-acid	

126771	Dichloroallyl-lawsone	
368390	DUP785 (brequinar)	
629971	Camptothecin,9-NH2 (RS)	**8) Topoisomerase Inhibitors**
176323	Camptothecin,9-MeO	
94600	Camptothecin	
603071	Camptothecin,9-NH2 (S)	
606172	Camptothecin,11-formyl (RS)	
606497	Camptothecin,20-ester (S)	
606985	Camptothecin,20-ester (S)	
618939	Camptothecin,20-ester (S)	
249910	Camptothecin,7-Cl	
610456	Camptothecin,20-ester (S)	
606173	Camptothecin,11-HOMe (RS)	
107124	Camptothecin,10-OH	
337766	Bisantrene	
123127	Doxorubicin	
301739	Mitoxantrone	
269148	Menogaril	
267469	Deoxydoxorubicin	
249992	Amsacrine	
355644	Anthrapyrazole-derivative	
122819	Teniposide	
82151	Daunorubicin hydrochloride	
141540	Etoposide	
349174	Oxanthrazole (piroxantrone)	
308847	Amonafide	
164011	Zorubicin (Rubidazone)	
95678	3-Hydropicolinaldehyde-thiosemicarbazone	**9) Unknown**
49842	Vinblastine-sulfate	
658831	Taxol analog	
107392	5-Hydroxypicolinaldehyde-thiosemicarbazone	
153858	Maytansine	
757	Colchicine	
673188	Taxol analog	
671867	Taxol analog	
664402	Taxol analog	
661746	Taxol analog	
33410	Colchicine-derivative	
376128	Dolastatin-10	
673187	Taxol analog	
664404	Taxol analog	
671870	Taxol analog	
67574	Vincristine-sulfate	
83265	Trityl-cysteine	
666608	Taxol analog	
600222	Taxol analog	
609395	Halichondrin B	
125973	Taxol	
118994	Inosine-glycodialdehyde	
656178	Taxol analog	

Table 4.3 miRNA cluster assignment from miRConnect study on 124 miRNAs.

microRNA	microRNA Cluster
hsa-miR-135a	Cluster I
hsa-miR-135b	
hsa-miR-7	
hsa-miR-192	
hsa-miR-194	
hsa-miR-429	
hsa-miR-200a	
hsa-miR-200b	
hsa-miR-141	
hsa-miR-200c	
hsa-miR-203	
hsa-miR-375	
hsa-miR-335	Cluster II
hsa-miR-328	
hsa-miR-95	
hsa-miR-425*	
hsa-miR-374a	
hsa-miR-98	
hsa-miR-340	
hsa-miR-330-3p	
hsa-miR-33a	
hsa-miR-96	
hsa-miR-15b	Cluster III
hsa-miR-339-5p	
hsa-miR-26a	
hsa-miR-148b	
hsa-miR-196a	
hsa-miR-331-3p	Cluster IV
hsa-miR-182	
hsa-miR-183	
hsa-let-7d	
hsa-miR-345	
hsa-miR-107	
hsa-miR-103	
hsa-miR-301a	
hsa-miR-423-3p	Cluster V
hsa-miR-142-3p	
hsa-miR-17*	
hsa-miR-18a	
hsa-miR-106a	
hsa-miR-17	
hsa-miR-19a	
hsa-miR-19b	
hsa-miR-20a	
hsa-miR-92a	
hsa-miR-32	
hsa-miR-191	
hsa-miR-93	
hsa-miR-106b	
hsa-miR-25	
hsa-miR-126	
hsa-miR-130b	
hsa-miR-186	
hsa-miR-378	
hsa-miR-378*	
hsa-miR-197	
hsa-let-7f	
hsa-miR-361-5p	
hsa-miR-342-3p	
hsa-miR-181c	
hsa-miR-181a	
hsa-miR-181b	
hsa-miR-15a	
hsa-miR-16	
hsa-miR-30e	Cluster VI
hsa-let-7g	
hsa-miR-29c	

hsa-miR-101	
hsa-miR-148a	
hsa-miR-195	
hsa-miR-296-5p	**Cluster VII**
hsa-miR-188-5p	
hsa-miR-34a	
hsa-miR-30d	
hsa-miR-29a	
hsa-miR-29b	
hsa-miR-204	
hsa-miR-105	
hsa-miR-146a	
hsa-miR-324-5p	
hsa-miR-140-5p	**Cluster VIII**
hsa-miR-145	
hsa-miR-152	
hsa-miR-28-5p	
hsa-miR-132	**Cluster IX**
hsa-miR-9	
hsa-let-7i	**Cluster X**
hsa-miR-99b	
hsa-let-7e	
hsa-miR-125a-5p	
hsa-miR-21	
hsa-miR-23b	
hsa-miR-27b	
hsa-miR-196b	**Cluster XI**
hsa-miR-212	
hsa-miR-10b	
hsa-miR-30a	
hsa-miR-30b	
hsa-miR-30e*	
hsa-miR-30a*	
hsa-miR-30c	
hsa-miR-221	**Cluster XII**
hsa-miR-222	
hsa-miR-130a	
hsa-miR-149	
hsa-miR-10a	
hsa-miR-151-3p	
hsa-miR-31	
hsa-miR-125b	**Cluster XIII**
hsa-miR-100	
hsa-miR-99a	
hsa-miR-22	
hsa-miR-24	
hsa-let-7c	
hsa-miR-137	
hsa-miR-23a	
hsa-miR-27a	
hsa-miR-218	
hsa-let-7a	
hsa-let-7b	
hsa-miR-155	
hsa-miR-210	
hsa-miR-193a-3p	
hsa-miR-424	

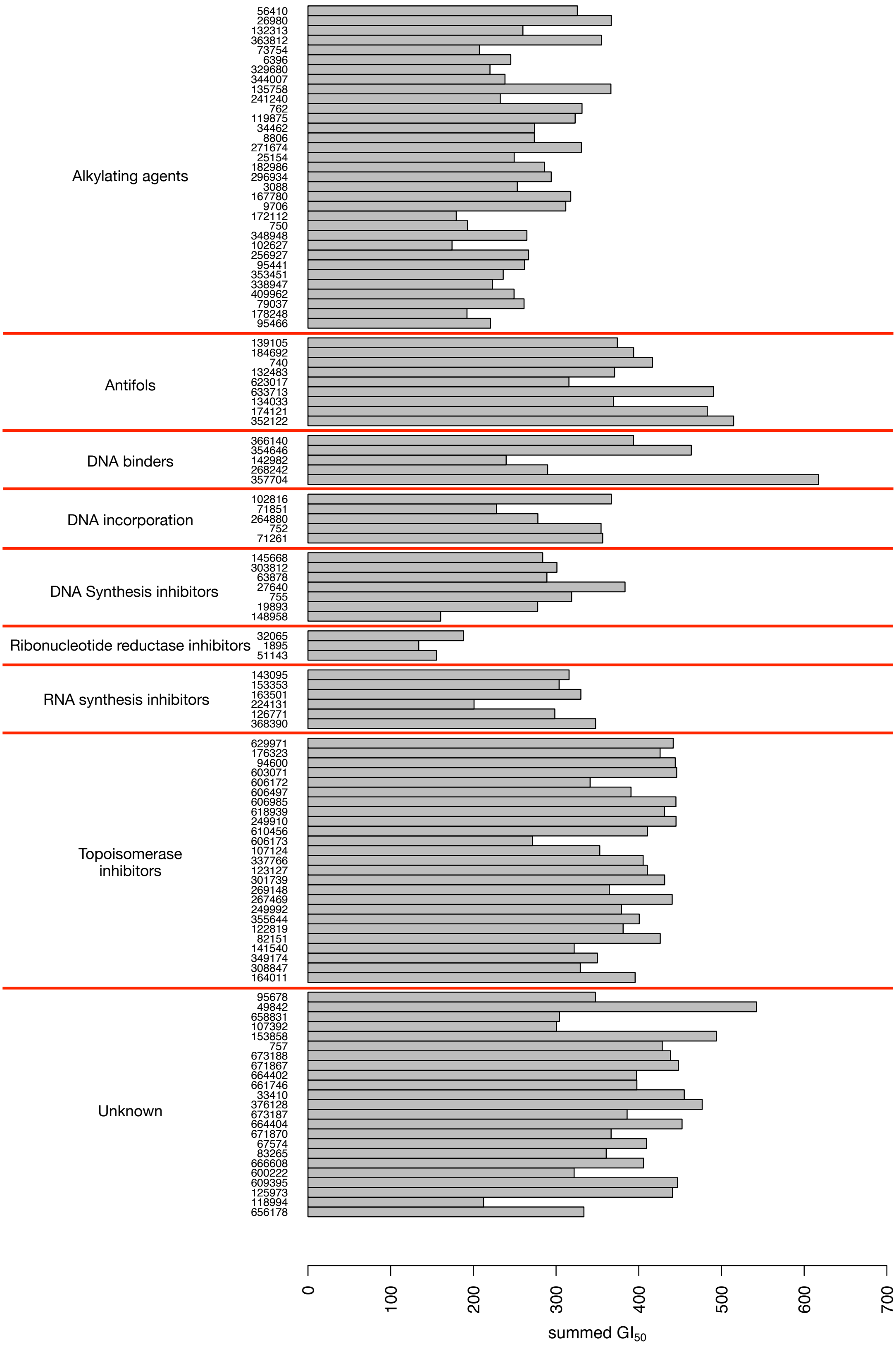


Figure 4.2 Summed GI_{50} of each drug compound. The drugs are divided into groups according to their mechanisms of actions. The labels of drugs are the NSC numbers, which can be traced back to the names of compounds in **Table 4.2**.

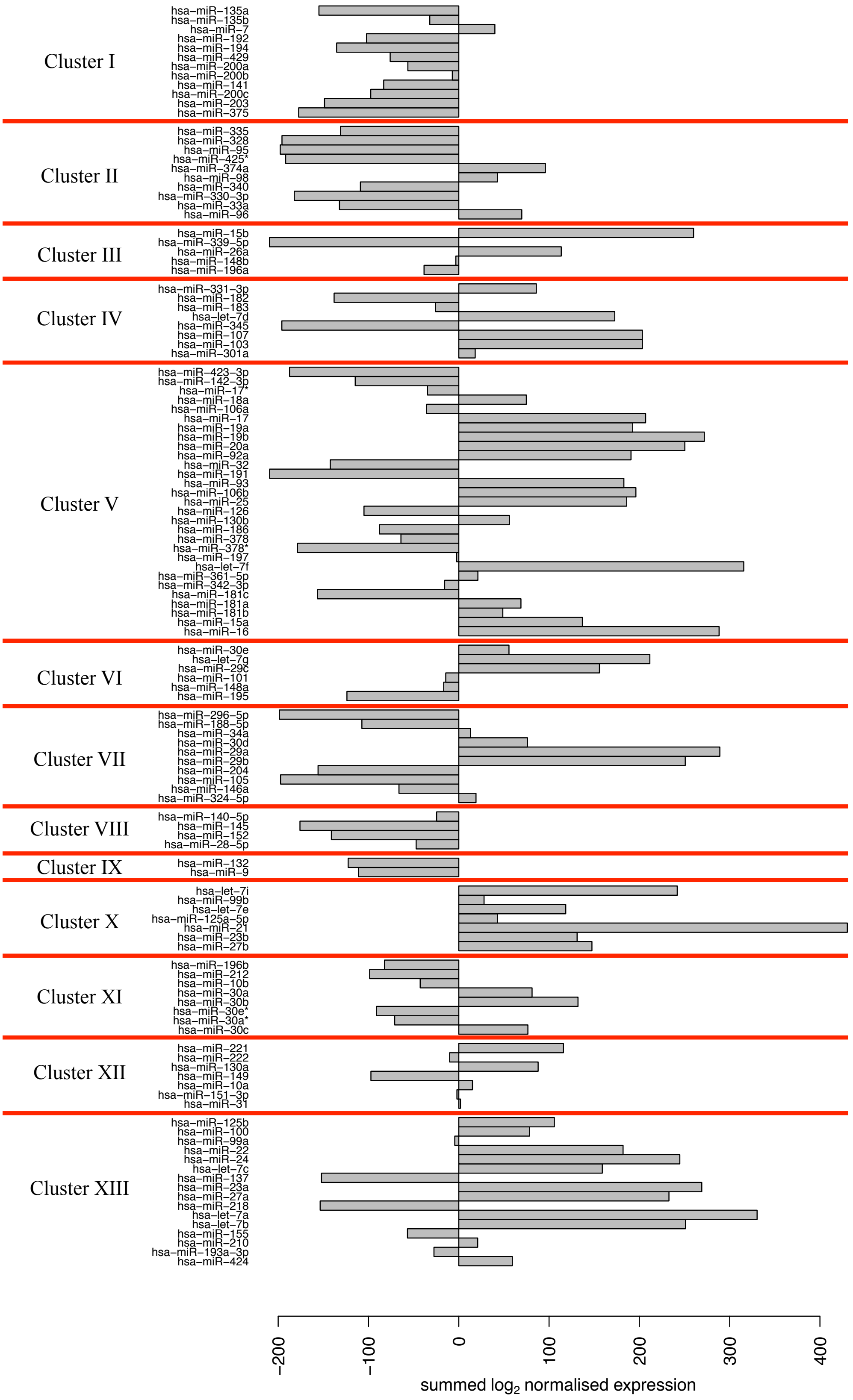


Figure 4.3 Summed $\log_2$ normlised expression of each miRNA in Lie et al. (2010) data set. The miRNAs are divided into groups according to miRNA clusters from miRConnect study. As the expression data was log scaled, some miRNAs that have very low expression (i.e. less than 1) might have negative values.

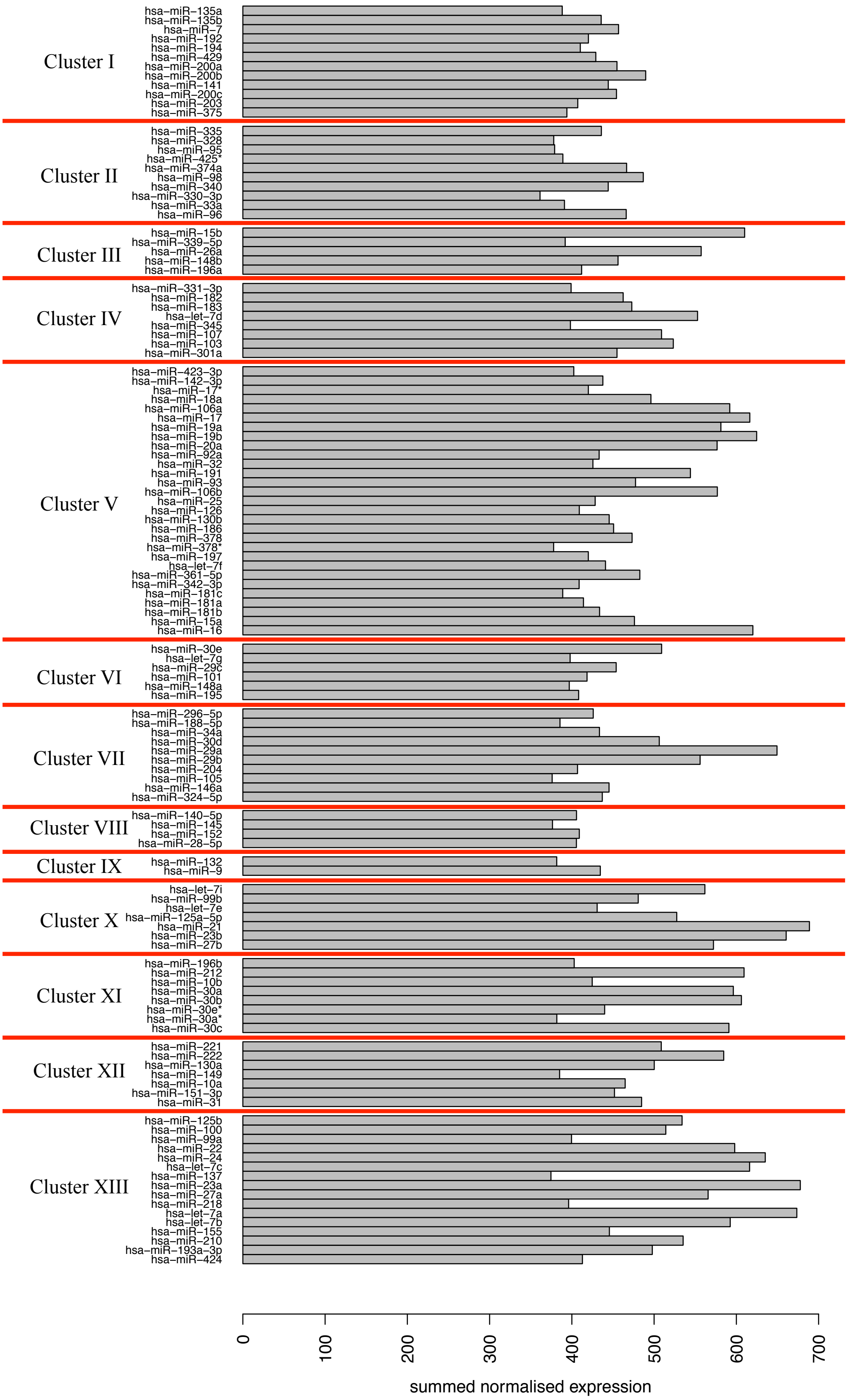


Figure 4.4 Summed normalised expression of each miRNA in Søkilde et al. (2011) data set. The miRNAs are divided into groups according to miRNA clusters from miRConnect study.

4.3 Results

4.3.1 ORCA identifies an association between the sensitivity in cancer cell lines to two chemotherapeutic drug groups: alkylating agents and topoisomerase inhibitors

To evaluate ORCA, the method was tested on varied data sets, including drug sensitivity data and miRNA profiles for the NCI-60 cell line panel. The objective of ORCA on the drug sensitivity data set was to determine if there are similarity of sensitivity pattern across the cell line panel between any pair of drug groups. From the correlation matrix between all 116 chemotherapeutic drugs, there are drug groups that have high similarity pattern between them. ORCA was used to determine if any two drug-groups pair has correlation coefficients that are above a correlation threshold more than expected by chance. Applying the threshold selection method in section 4.2.2 identified the correlation coefficient threshold of the drug sensitivity data set to be 0.79, i.e. a correlation coefficient threshold of 0.79 yields the highest information content according to the Shannon's entropy-like score. **Figure 4.5** illustrates Shannon's entropy-like score from different correlation coefficient cutoffs. **Figure 4.6** illustrates the correlation matrix between every drug from drug sensitivity data set. Correlation coefficients that passed the 0.79 correlation cutoff highlighted in red circles with the boundaries of the drug groups in yellow lines. After ORCA was applied at the correlation coefficient threshold of 0.79, a p-value table for obtaining the number of correlation coefficients above the threshold by chance between every possible group pairs was generated. **Table 4.4**, shows the hypergeometric p-values and FDR-adjusted p-values between all the group pairs.

From a closer inspection, the sensitivity pattern across NCI-60 cell panel between alkylating agents and topoisomerase inhibitors (drug group 1 and 8 according to

Table 4.2 and **Figure 4.6**, respectively) was very similar, judging by the number of correlation coefficients that passed the threshold of 0.79 between the two drug groups.

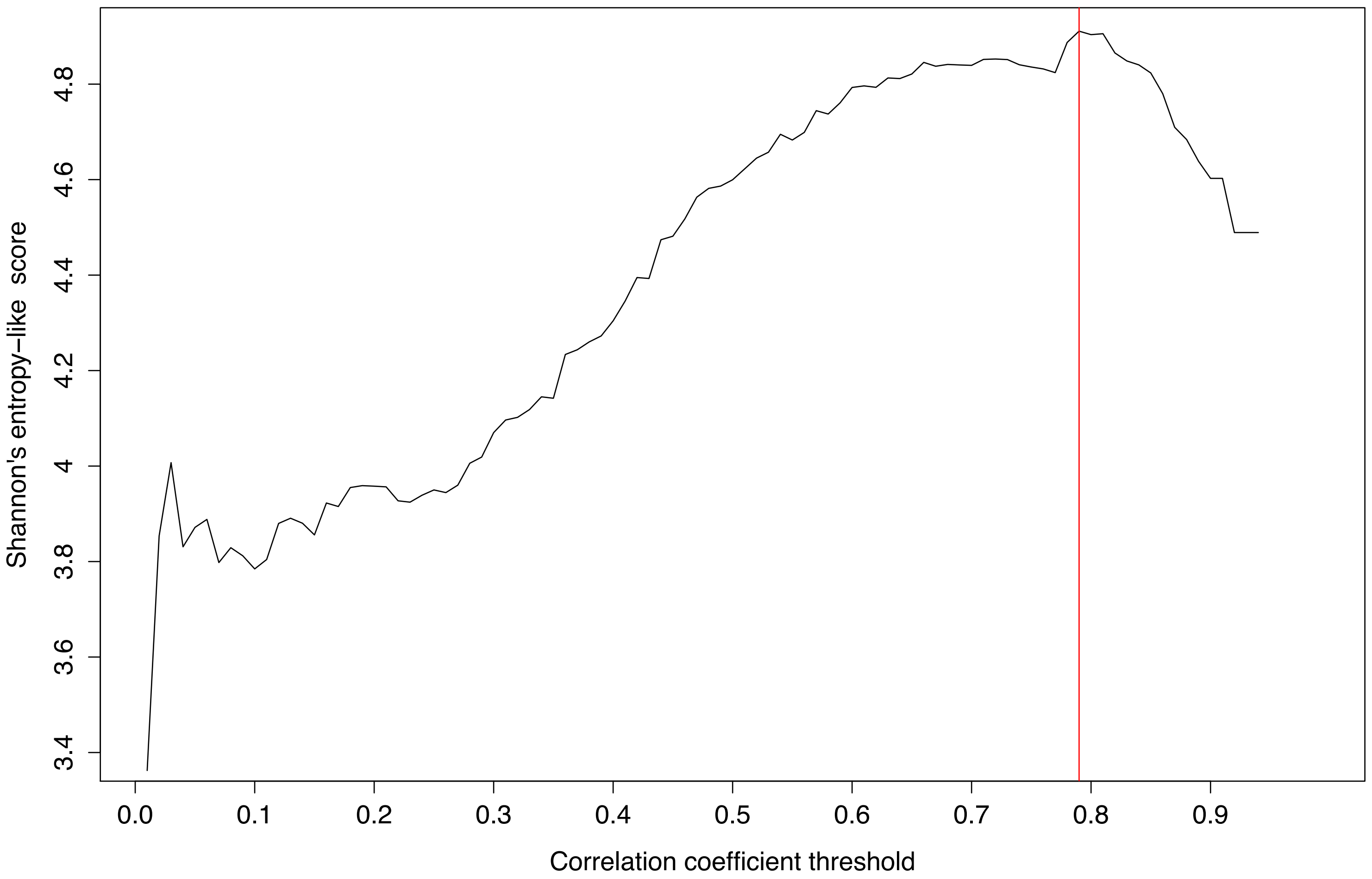


Figure 4.5 Shannon's entropy-like scores from different correlation coefficient thresholds of the drug sensitivity data. Shannon’s entropy-like score was calculated for each correlation coefficient cutoff (from 0 to 1 with an increment of 0.01). The red vertical line marks the correlation coefficient threshold that yields the highest Shannon’s entropy-like score, which was 0.79 for the drug sensitivity data set.

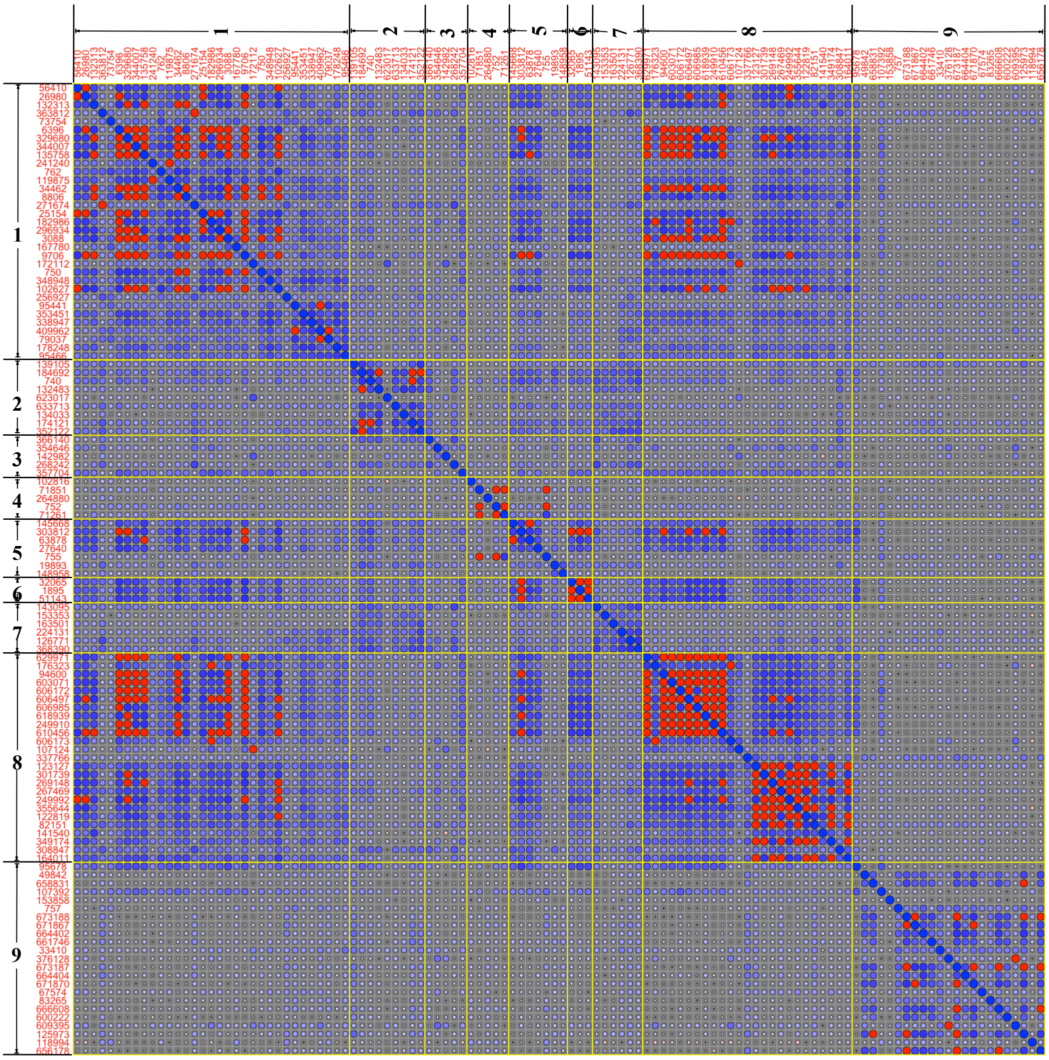


Figure 4.6 Correlation matrix between all drugs in the drug sensitivity data set. Each circle represents a correlation coefficient between two drugs. The sizes of the circles reflect the strength of the correlation coefficients. Blue circles represent positive correlation coefficients, whereas pink circles represent negative correlation coefficients. The circles marked in red are the correlation coefficients that pass the correlation coefficient threshold of 0.79. Each drug is labeled using its NSC number according to **Table 4.2** . The drug group names are as follow: 1) Alkylating agents; 2) Antifols; 3) DNA binders; 4) DNA incorporation; 5) DNA synthesis inhibitors; 6) Ribonucleotide reductase inhibitors; 7) RNA synthesis inhibitors; 8) Topoisomerase inhibitors; 9) Unknown.

From **Table 4.4A,** the FDR-adjusted hypergeometric p-value for obtaining the number of correlation coefficients of the drugs between alkylating agents and topoisomerase inhibitors was 1 × 10^{-57}, and the empirical p-value from the one-million permutation test was 1 × 10^{-6}, which suggests that it is highly unlikely that this amount of correlation coefficients that passed the threshold of 0.79 was by chance. The high correlation between these two drug groups may results from the fact that the compounds in these two drug groups can cause single- and double-stranded DNA breaks in rapidly dividing cells, in which the cells have to rely on a common set of DNA repair pathways for cell survival (Bargonetti et al. 2010; Rudolf et al. 2011). Another drug group pair that shows the association that is higher than expected is the group pair between DNA synthesis inhibitors (group 5) and ribonucleotide reductase inhibitors (group 6), although with a lesser degree of statistical significance than the group pair between alkylating agents and topoisomerase inhibitors (see **Table 4.4A**).

Table 4.4 P-value table after ORCA was applied to the drug sensitivity data set.

A) Empirical p-value

Drug Group	1	2	3	4	5	6	7	8	9	Drug Group
1	■	0.94	0.82	0.82	0.29	0.70	0.87	1E-6	0.99	1
2	1.00	■	0.59	0.59	0.66	0.58	0.60	0.92	0.91	2
3	1.00	1.00	■	0.58	0.68	0.42	0.64	0.78	0.79	3
4	1.00	1.00	1.00	■	0.11	0.42	0.64	0.78	0.79	4
5	1.00	1.00	1.00	0.33	■	0.01	0.62	0.24	0.86	5
6	1.00	1.00	1.00	1.00	0.01	■	0.47	0.64	0.62	6
7	1.00	1.00	1.00	1.00	1.00	1.00	■	0.81	0.81	7
8	1E-57	1.00	1.00	1.00	1.00	1.00	1.00	■	0.99	8
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	■	9
Drug Group	1	2	3	4	5	6	7	8	9	Drug Group

B) FDR-adjusted p-value

Drug Group	1	2	3	4	5	6	7	8	9
Empirical p-value	5E-5	0.03	0.31	0.007	0.21	1E-6	0.11	1E-6	0.01
FDR-adjusted p-value	9E-21	0.01	0.41	0.001	0.31	0	0.56	2E-37	0.26

A) Benjamini-Hochberg FDR adjusted hypergeometric p-values (lower half) and empirical p-values from permutation test (upper half). B) P-values from within-group correlation coefficients (from diagonal of the matrix). The p-values in grey-shaded boxes are the group pairs that have the p-values less than 0.05.

4.3.2 Using ORCA to verify miRNA clusters defined by the miRConnect study

In this part, the baseline expression of 124 miRNAs from the two data sets, i.e. Liu et al. (2010) and Søkilde et al. (2011), were divided into 13 clusters according to the miRConnect study (Hua et al. 2011). The two data sets have the same basic data structure, i.e. equal number of variables (miRNAs) and number of data points in each variable from the same number of cell lines used in the experiments. However, the information content might be different between the two data sets, due to systematic variations, for example, from different types of microarrays used in the measurement of miRNA expression. If a single correlation coefficients threshold is used in ORCA for both of the data sets, where the amount of information between the two data sets are different, the results would not be comparable. The threshold selection method based on Shannon's entropy in the section 4.2.2 is an objective means to select the correlation coefficient cutoff that is suitable for individual data sets. Therefore, after applied ORCA to the two miRNA data sets, the correlation coefficient thresholds for the two data sets were determined to be 0.29 and 0.61 for Liu et al. (2010) and Søkilde et al. (2011) data sets, respectively. **Figure 4.7** shows the Shannon's entropy-like scores from different correlation coefficient cutoffs of the Liu et al. (2010) and Søkilde et al. (2011) miRNA data sets. **Figure 4.8** shows the correlation matrices from the two miRNA data sets with cluster boundaries. **Table 4.5** and **Table 4.6** show the p-values after ORCA were applied on the two miRNA data sets.

From the **Table 4.5** and **Table 4.6**, several clusters proposed by Hua et al. (2011) contained significant within-cluster overrepresentation of correlations identified by ORCA at FDR-adjusted p-value of 0.05 in both data sets, which were clusters I, IV, V, X and XIII. As for association between any two clusters, ORCA identified one cluster pair that had overrepresentation of correlations, which was cluster pair X/XIII. The miRNAs within the clusters I, IV, V, X and XIII contain overrepresentation of correlations as determined by ORCA support the hypothesis that the miRNAs in these clusters might be controlled by the same transcription factors, located at the same chromosomal regions, or involved in the same processes or pathways. The association between cluster pair X/XIII, which was determined to be statistically significant by ORCA, could be explained by the fact that several of the miRNAs in these two clusters have the same seed sequences from the same miRNA families. According to miRBase database release 21 in June 2014 (Kozomara & Griffiths-Jones 2014), the miRNA families presented in these two miRNA clusters are let-7 family (let-7e, and let-7i in cluster X and let-7a, let-7b, let-7c in cluster XIII), miR-23 family (miR-23b in cluster X and miR-23a in cluster XIII), miR-27 family (mir-27b in cluster X and miR-27a in cluster XIII), miR-125 family (miR-125a-5p in cluster X and miR-125b in cluster XIII), and miR-99/100 family (miR-99b in cluster X, and miR-99a, miR-100 in cluster XIII). This observation has not been reported in the original miRConnect study. The result suggests that this cluster pair X/XIII might be considered as a single superfamily of miRNAs, or a single cluster. Other miRNA cluster pairs identified by ORCA as statistically significant, miRNA cluster pairs II/IV and V/VI, were difficult to rationalised than cluster pair X/XIII because within-cluster overrepresentation of correlation in clusters II and VI were not statistically significant. This result suggests that there might be possible misclassification in the original clustering of these miRNAs.

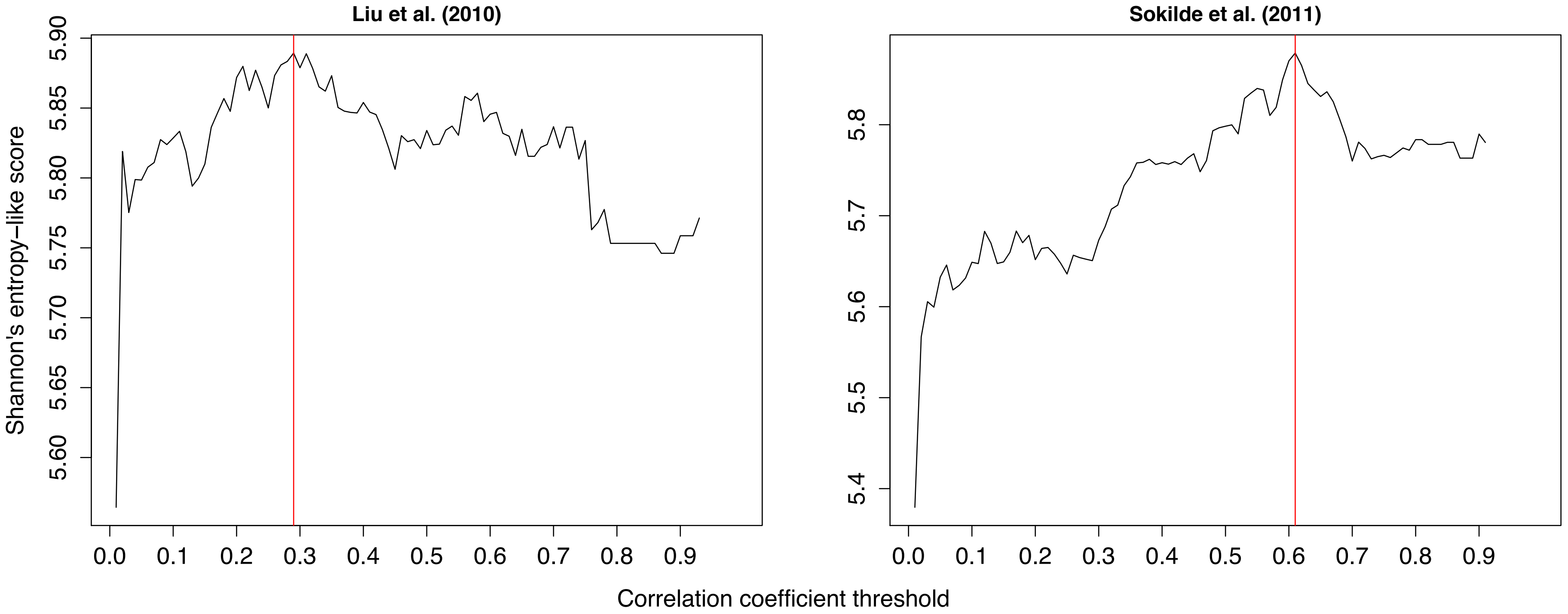
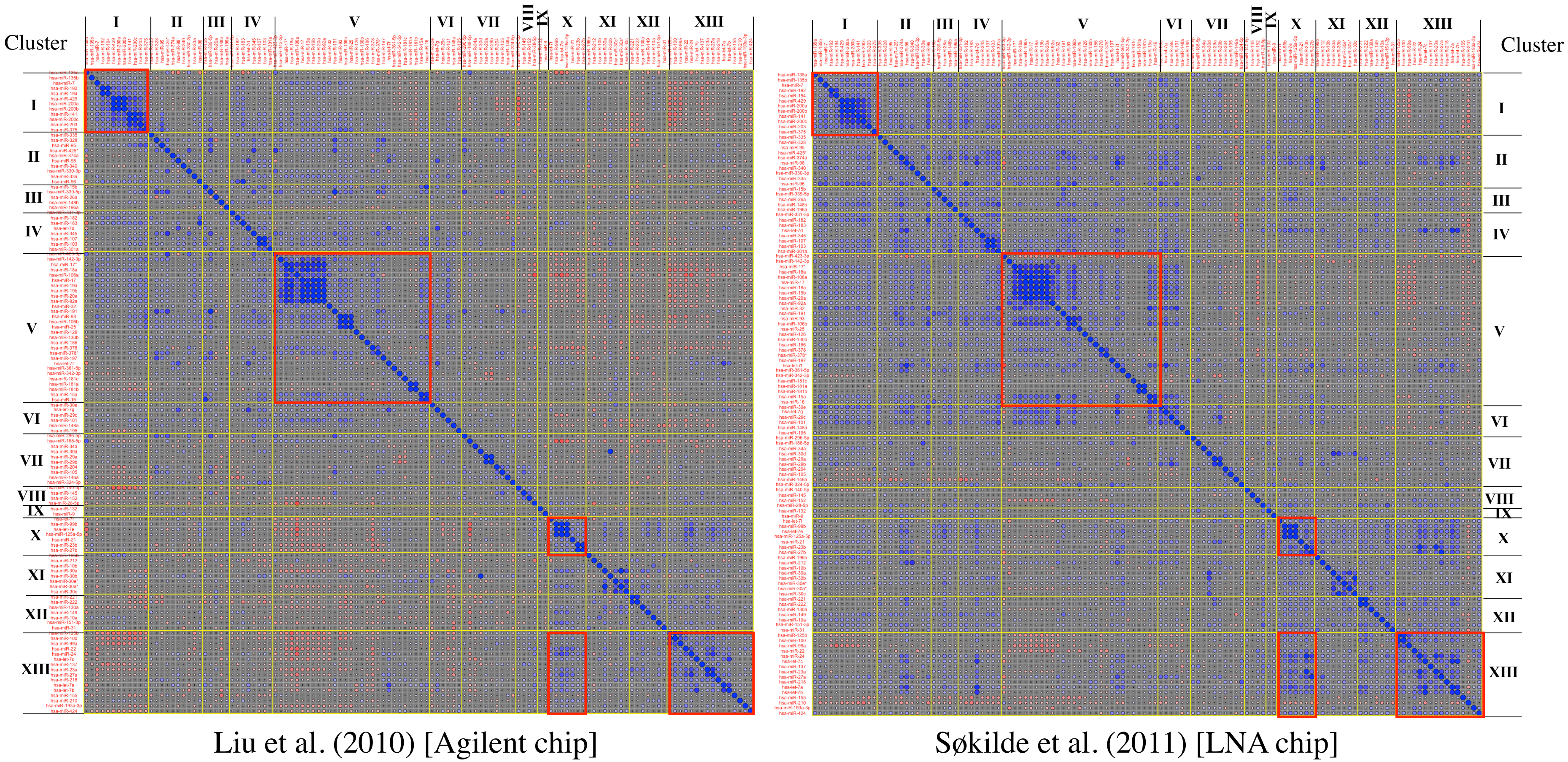


Figure 4.7 Shannon's entropy-like scores from different correlation coefficient thresholds of the two miRNA data sets. Shannon's entropy-like score was calculated for each correlation coefficient cutoff (from 0 to 1 with an increment of 0.01). The red vertical lines mark the correlation coefficient threshold that yields the highest Shannon's entropy-like scores, which were 0.29 (left) and 0.61 (right) for Liu et al. (2010) and Søkilde et al. (2011), respectively. These results demonstrate that different sources of data have different amount of information, in which a single correlation coefficient threshold will not be able to yield the highest amount of data from different sets of data.



Liu et al. (2010) [Agilent chip]

Søkilde et al. (2011) [LNA chip]

Figure 4.8 Correlation matrices from the two miRNA data sets. The circles represent correlation coefficients between two miRNAs. Blue and pink circles are positive and negative correlation coefficients, respectively. The size and the colour shade of the circle signify the strength of the correlation. The clusters are divided by yellow lines, and the red boxes indicate the cluster pairs that have FDR-adjusted hypergeometric p-values and empirical p-values less than 0.05 in both data sets.

Table 4.5 P-value table after ORCA was applied to the Liu et al. (2010) miRNA data set.

A)

Empirical p-value

Cluster	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**XIII**	Cluster
I		0.17	0.94	0.005	0.003	0.74	0.94	0.20	0.95	0.99	0.94	0.35	0.007	**I**
II	0.33		0.80	0.04	0.02	0.71	0.54	0.75	0.67	0.86	0.99	0.94	0.99	**II**
III	1.00	1.00		0.18	0.06	0.55	0.61	0.20	0.68	0.99	0.48	0.90	0.88	**III**
IV	4E-4	0.027	0.44		0.01	0.35	0.77	0.85	0.87	0.88	0.96	0.64	0.99	**IV**
V	7E-6	8E-4	0.025	4E-4		0.04	0.96	0.65	0.99	0.008	0.46	0.97	0.08	**V**
VI	1.00	1.00	1.00	0.85	6E-3		0.98	0.97	0.93	0.66	0.26	0.67	0.48	**VI**
VII	1.00	1.00	1.00	1.00	1.00	1.00		0.26	0.82	0.91	0.99	0.54	0.99	**VII**
VIII	0.44	1.00	0.44	1.00	1.00	1.00	0.69		0.85	0.63	0.97	0.99	0.61	**VIII**
IX	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		0.83	0.71	0.83	0.99	**IX**
X	1.00	1.00	1.00	1.00	1E-4	1.00	1.00	1.00	1.00		0.73	0.28	3E-6	**X**
XI	1.00	1.00	1.00	1.00	1.00	0.67	1.00	1.00	1.00	1.00		0.02	0.46	**XI**
XII	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.7	8E-3		0.11	**XII**
XIII	3E-4	1.00	1.00	1.00	0.031	1.00	1.00	1.00	1.00	9E-14	1.00	0.16		**XIII**
Cluster	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**XIII**	Cluster

FDR-adjusted p-value

B)

Cluster	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**XIII**
Empirical p-value	3E-6	0.15	0.21	0.006	1E-5	0.45	0.57	0.05	0.24	1E-4	0.33	0.23	4E-5
FDR-adjusted p-value	4E-15	0.67	0.83	4E-3	4E-15	1.00	1.00	0.22	0.83	4E-7	1.00	0.87	3E-11

A) Benjamini-Hochberg FDR adjusted hypergeometric p-values (lower half) and empirical p-values from permutation test (upper half). B) P-values from within-cluster correlation coefficients (from diagonal of the matrix). The p-values in grey-shaded boxes are the clusters and cluster pairs that have the FDR-adjusted and empirical p-values less than 0.05. P-values in red are the clusters and cluster pair that passed the threshold of in both miRNA data sets (also correspond to red boxes on the left panel in the figure 4.8). P-values in blue correspond to the figure 4.9.

Table 4.6 P-value table after ORCA was applied to the Søkilde et al. (2011) miRNA data set.

Empirical p-value

A)

Cluster	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**XIII**	Cluster
I		0.02	0.51	0.55	0.84	0.23	0.91	0.72	0.47	0.87	0.92	0.89	0.96	**I**
II	0.016		0.12	0.003	0.15	0.005	0.38	0.66	0.42	0.11	0.27	0.48	0.28	**II**
III	0.61	0.32		0.024	0.37	0.25	0.74	0.44	0.26	0.31	0.67	0.63	0.86	**III**
IV	0.79	1E-3	**0.049**		0.27	0.12	0.88	0.59	0.37	0.50	0.82	0.78	0.17	**IV**
V	0.95	0.28	0.61	0.52		1E-6	0.64	0.88	0.70	0.87	0.88	0.95	0.99	**V**
VI	0.52	1E-3	0.43	0.32	6E-3		0.16	0.50	0.30	0.69	0.73	0.69	0.91	**VI**
VII	0.95	0.61	0.83	0.92	5E-14	0.43		0.66	0.43	0.48	0.15	0.85	0.82	**VII**
VIII	0.83	0.79	0.61	0.75	0.85	0.68	0.79		0.22	0.55	0.59	0.23	0.79	**VIII**
IX	0.68	0.61	0.54	0.61	0.95	0.56	0.61	0.51		0.33	0.37	0.33	0.55	**IX**
X	0.92	0.26	0.51	0.61	0.86	0.79	0.68	0.72	0.59		0.50	0.74	6E-5	**X**
XI	0.93	0.55	0.79	0.86	0.95	0.83	0.32	0.75	0.61	0.61		0.78	0.71	**XI**
XII	0.92	0.68	0.77	0.85	0.97	0.79	0.89	0.41	0.59	0.83	0.85		0.07	**XII**
XIII	0.99	0.55	0.92	0.41	0.99	0.93	0.92	0.86	0.75	6E-8	0.86	0.13		**XIII**
Cluster	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**XIII**	Cluster

FDR-adjusted p-value

B)

Cluster	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**XIII**
Empirical p-value	2E-5	0.19	0.26	1E-4	3E-6	0.006	0.38	0.17	0.03	6E-4	0.001	0.14	0.003
FDR-adjusted p-value	2E-10	0.57	0.61	4E-7	2E-17	0.013	0.77	0.55	0.22	8E-5	0.54	3E-4	3E-11

A) Benjamini-Hochberg FDR adjusted hypergeometric p-values (lower half) and empirical p-values from permutation test (upper half). B) P-values from within-cluster correlation coefficients (from diagonal of the matrix). The p-values in grey-shaded boxes are the clusters and cluster pairs that have the FDR-adjusted and empirical p-values less than 0.05. P-values in red are the clusters and cluster pair that passed the threshold and in both miRNA data sets (also correspond to red boxes on the right panel in the figure 4.8).

In a follow-up study by the authors of the miRConnect study, clusters I and V was found to be functionally antagonistic to miRNAs in cluster XIII, in which they observed that clusters I and V had the opposite effect on the same set of mRNAs compared to cluster XIII (Hua et al. 2013). ORCA also identified significant overrepresentation of correlations (which were predominantly negative correlation coefficients) between cluster pairs I/XIII and V/XIII in one of microarray data sets [Liu et al. (2010)]. **Figure 4.9** shows the correlation coefficients between the cluster pairs I/XIII and V/XIII. ORCA can detect functional antagonism by using only microRNA expression data set alone unlike the follow-up study, which highlighted the potential of this method.

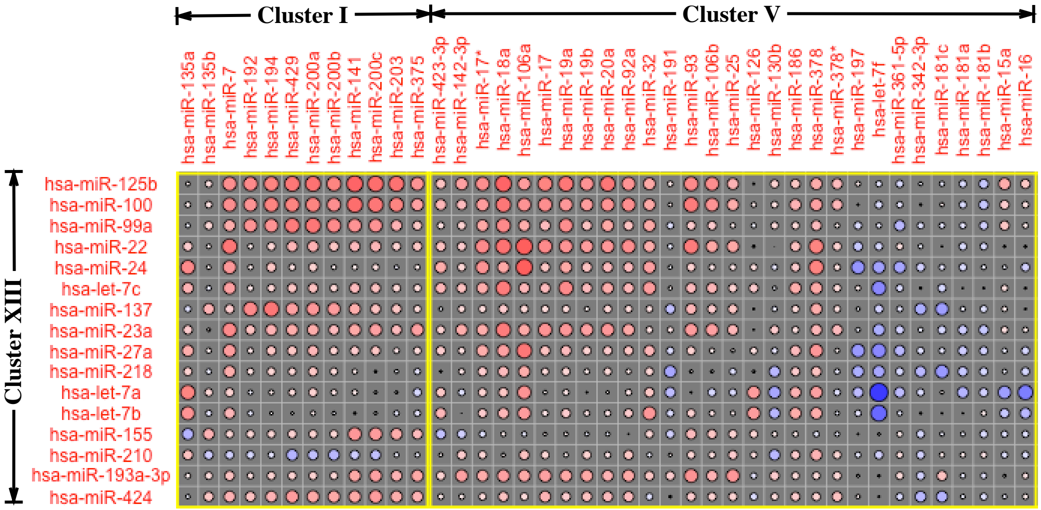


Figure 4.9 Correlation coefficients between cluster pairs I/XIII and V/XIII. Most of the correlation coefficients between these cluster pairs were predominantly negative correlation, which suggests that miRNAs in clusters I and V might be working against miRNAs in cluster XIII.

4.4 Discussion

ORCA has a potential to be used as a pathway analysis tool as well, but the research question that need to be asked will be different from existing pathway or gene set analysis methods. Currently, most pathway analysis methods identify pathways that are significantly enriched or depleted through the genes that deemed differentially expressed in a certain biological condition. However, ORCA could identify associated pathways through correlations between the genes within the pathways. Although the examples presented in this chapter were not pathway analysis, ORCA could be applied to any type of pathway or gene set in order to determine the associated pathways. OCRA could address some limitations of existing pathway analysis methods outlined by Khatri et al. (2012). First, pathway analysis tools using ORA do not take into account the actual levels of variables (such as gene expression or metabolite levels). ORCA takes these values into account through correlation coefficient calculations, albeit indirectly, which in effect ORCA does not weigh the variables equally. Second, ORCA does not assume that the variables are independent, which usually is an important assumption of typical pathway analysis tools that implementing ORA. Third, classical ORA only use variables, such as mRNAs or miRNAs, which are deemed differentially expressed in a certain condition. ORCA, on the other hand, uses all the variables in the calculation. Fourth, ORA-based pathway analysis tools usually apply multiple hypothesis testing for p-value correction, which has an assumption on the independence of the pathways tested. However, ORCA has the opposite assumption and seeks to find the association between two groups of variables.

ORCA is based on the hypergeometric distribution, therefore it could be argued that this method violates the assumption of independence of each data point because the use of correlation coefficient in the p-value calculation, which may lead to inaccurate p-value calculations. A simulation study by Goeman & Bühlmann (2007) have shown that when hypergeometric test was used in correlated data, the p-values calculated from the test were found to be underestimated. There are, however, means to correct the underestimated p-values by using the multiple comparison procedures, such as Bonferroni correction or Benjamini-Hochberg FDR correction (which was used in this study). The alpha level of the p-value can also be selected from a table of nominal alpha level for correlated data from (Goeman & Bühlmann 2007) to match the correlation coefficient threshold used in the analysis. Note that in the simulation study previously mentioned, correlated data have the same correlation coefficient in every data points. Wilcoxon rank-sum test could also be used instead of hypergeometric test when the comparison is between the numbers of correlation coefficients in one group pair against all other group pairs.

ORCA can be applied as a pathway analysis tool, but the information from the analysis will be related specifically to the pathway interactions. For example, ORCA can calculate the correlation coefficients between the mRNAs or metabolites from pathways, and then identify the pathway pairs with higher correlation coefficients that pass a certain threshold (determined by either the threshold selection method or by other means) than expected by chance.

This version of ORCA requires the data sets to have exclusive group memberships for the variables, i.e. each variable can belong to only one group. Therefore, at present ORCA cannot be used with pathway data where variables are overlapped in two or more pathways. This is a subject for the future work.

In conclusion, ORCA is a novel method that combines correlation analysis with overrepresentation analysis, which has potential to reveal associations between sets of variables that might not be uncovered by conventional statistical methods in a wide variety of biological data sets. ORCA has clear application in "-omics" data analysis, however, it will be profitable in any circumstance where an association network can be constructed between variables that can be classified into meaningful sets.

Chapter 5 : Conclusions and future works

5.1 Conclusions

This thesis presented novel utilisation of various statistical methods for exploring the roles of miRNAs on metabolism. In the second chapter a specific miRNA, miR-22 was studied from several independent breast cancer data sets with patient survival data. Spearman's Rank correlation analysis was used to assess the potential of the miR-22 host gene expression, *MIR22HG*, as a surrogate marker for mature miR-22 expression. Correlation and survival analyses were also used to evaluate the association between miR-22 and the three metabolic target genes, *MTHFD2*, *ELOVL6*, and *ACLY*, previously identified in the Keun group laboratory, *in vivo*. Correlation between miR-22 and the two target genes, *MTHFD2* and *ELOVL6,* were found to be statistically significant. Survival analyses on the relapse-free of breast cancer patient data were performed using miR-22 and the three metabolic target genes as the predictors. It was found that the expression of miR-22, *MTHFD2*, and *ELOVL6*, affected the survival outcome of the breast cancer patients. High expression of miR-22 was beneficial to the survival outcome of the breast cancer patients, while the high expression of the metabolic target genes, i.e. *MTHFD2* and *ELOVL6*, exhibited the opposite effects. The second chapter concluded with the effect modification analysis of miR-22 on the three metabolic target genes using Cochran-Armitage trend test. The survival outcomes of breast cancer patients with low expression of *MTHFD2* and *EVOLV6* were modified to be better by the high expression of miR-22 than the target genes alone.

The third chapter was the global analysis on the predicted metabolic target genes of miRNA families from the TargetScan miRNA target prediction database. This chapter applied weighted Jaccard index to overlapped predicted metabolic target genes of every miRNA family pair to create a weighted Jaccard distance matrix for hierarchical clustering analysis. The miRNA families were then classified into six different clusters, and the predicted metabolic target genes in these clusters were analysed by pathway enrichment analysis. Three clusters yielded several enriched pathways in metabolism, cancer-specific and cancer-related pathways.

The fourth chapter described a novel statistical method that was created by combining two classical statistical methods, called over-representation of correlation analysis (ORCA). This new statistical analysis was invented to identify association between two sets of variables through correlation coefficients between the variables in the two sets. The method was applied to two problems: one problem was to identify associations between chemotherapeutic drugs by using drug sensitivity data on NCI-60 cancer cell line panel; another problem was to verify miRNA clusters previously identified by an independent study. **In conclusion, combinations of classical statistical methods can be used to reveal novel insights and provide a better understanding of miRNA biology from publicly available data sets.**

5.2 Future work

The main limitation in the miR-22 analysis was the usage of *MIR22HG* expression as the surrogate marker for mature miR-22 expression. Although the high degree and statistically significant results of positive correlation between the host gene and mature species were observed, the actual measurement of active miRNA species is definitely better than a surrogate. This could be achieved when the high-throughput mature miRNA data sets are more widely available. Another drawback was lying in the survival analysis. Breast cancer survival outcomes are heavily affected by the hormone receptors status in the tumours. The survival analysis in this thesis did not stratify the patients by any hormone receptor status in order to gain the highest number of samples possible. Again, with more data available, complete with hormone receptors status, the reanalysis can be performed to determine the roles of the hormone receptors.

In global analysis of predicted metabolic target genes of miRNA families, the main drawback is that only one predicted target genes database was used, in which only two main features of miRNA target prediction were used. To increase the sensitivity of the analysis, a new way to integrate other main features for miRNA target gene prediction is needed in order to widen the predicted metabolic target genes. For a specificity issue, experimentally validated targets could be used for reanalysis to weed out spuriously enriched pathways. Another extension of the work would be to include the gene expression of specific pathways or conditions to confirm the findings.

For the chapter on ORCA, this method had been applied to the cluster verification problem. It should be used to verify the findings in the global predicted target genes analysis. Another extension of the method that could be implemented is to use other association metric other than the correlation, such as Jaccard index or other distance metrics. One big limitation of the method is the exclusive groupings requirement in the data used in the analysis. When the method can accept overlapped variables between two groups, this method could be applied more widely on the pathway level data, in which the variables usually overlapped between the sets.

References

Alexiou, P. et al., 2009. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics (Oxford, England)*, 25(23), pp.3049–55.

Ameres, S.L. & Zamore, P.D., 2013. Diversifying microRNA sequence and function. *Nature reviews. Molecular cell biology*, 14(8), pp.475–88.

Armitage, P., 1955. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3), pp.375–386.

Bargonetti, J., Champeil, E. & Tomasz, M., 2010. Differential toxicity of DNA adducts of mitomycin C. *Journal of nucleic acids*, 2010.

Bartel, D., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *cell*, 116, pp.281–297.

Baskerville, S. & Bartel, D.P., 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, 11(3), pp.241–247.

Beissbarth, T. & Speed, T.P., 2004. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics (Oxford, England)*, 20(9), pp.1464–5.

Benjamini, Y. & Hochberg, Y.B., 1995. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society , Series B*, 57(1), pp.289–300.

Biasiolo, M. et al., 2011. Impact of host genes and strand selection on miRNA and miRNA* expression. *PloS one*, 6(8), p.e23854.

Bioinformatics and research computing, 2012. TargetScan. Available at: http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_61 [Accessed October 7, 2013].

Blower, P.E. et al., 2007. MicroRNA expression profiles for the NCI-60 cancer cell panel. *Molecular cancer therapeutics*, 6(5), pp.1483–91.

Borchert, G.M., Lanier, W. & Davidson, B.L., 2006. RNA polymerase III transcribes human microRNAs. *Nature structural & molecular biology*, 13(12), pp.1097–101.

Boross, G., Orosz, K. & Farkas, I.J., 2009. Human microRNAs co-silence in well-separated groups and have different predicted essentialities. *Bioinformatics (Oxford, England)*, 25(8), pp.1063–9.

Bullard, J.H. et al., 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11, p.94.

Bushman, B.J. & Wang, M.C., 1995. A procedure for combining sample correlation coefficients and vote counts to obtain an estimate and a confidence interval for the population correlation coefficient. *Psychological Bulletin*, 117(3), pp.530–546.

Cairns, R. a, Harris, I.S. & Mak, T.W., 2011. Regulation of cancer cell metabolism. *Nature reviews. Cancer*, 11(2), pp.85–95.

Carthew, R.W. & Sontheimer, E.J., 2009. Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136(4), pp.642–55.

Cavill, R. et al., 2011. Consensus-Phenotype Integration of Transcriptomic and Metabolomic Data Implies a Role for Metabolism in the Chemosensitivity of Tumour Cells G. Tucker-Kellogg, ed. *PLoS Computational Biology*, 7(3), p.e1001113.

Chang, T.-C. et al., 2008. Widespread microRNA repression by Myc contributes to tumorigenesis. *Nature genetics*, 40(1), pp.43–50.

Chen, B. et al., 2012. Roles of microRNA on cancer cell metabolism. *Journal of translational medicine*, 10(1), p.228.

Cochran, W.G., 1954. The Combination of Estimates from Different Experiments. *Biometrics*, 10(1), pp.101–129.

Dang, C. V, 2013. MYC, metabolism, cell growth, and tumorigenesis. *Cold Spring Harbor perspectives in medicine*, 3(8).

Davis, B.N. & Hata, A., 2009. Regulation of MicroRNA Biogenesis: A miRiad of mechanisms. *Cell communication and signaling : CCS*, 7, p.18.

Davison, A.C. & Hinkley, D.V., 1997. *Bootstrap methods and their application* 9th ed., New York: Cambridge University Press.

Dillies, M.-A. et al., 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics*, 14(6), pp.671–83.

Dória, M.L. et al., 2014. Fatty acid and phospholipid biosynthetic pathways are regulated throughout mammary epithelial cell differentiation and correlate to breast cancer survival. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, pp.1–18.

Duarte, N.C. et al., 2007. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6), pp.1777–82.

Dumortier, O., Hinault, C. & Van Obberghen, E., 2013. MicroRNAs and metabolism crosstalk in energy homeostasis. *Cell metabolism*, 18(3), pp.312–24.

Elton, T.S. et al., 2013. Regulation of the MIR155 host gene in physiological and pathological processes. *Gene*, 532(1), pp.1–12.

Esquela-Kerscher, A. & Slack, F.J., 2006. Oncomirs - microRNAs with a role in cancer. *Nature reviews. Cancer*, 6(4), pp.259–69.

Eulalio, A., Huntzinger, E. & Izaurralde, E., 2008. Getting to the root of miRNA-mediated gene silencing. *Cell*, 132(1), pp.9–14.

Farazi, T. a et al., 2011. MicroRNA sequence and expression analysis in breast tumors by deep sequencing. *Cancer research*, 71(13), pp.4443–53.

Friedman, R.C. et al., 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research*, 19(1), pp.92–105.

Garmire, L.X. & Subramaniam, S., 2012. Evaluation of normalization methods in mammalian microRNA-Seq data. *RNA (New York, N.Y.)*, 18(6), pp.1279–88.

Garzon, R., Calin, G. a & Croce, C.M., 2009. MicroRNAs in Cancer. *Annual review of medicine*, 60, pp.167–79.

Gaur, A. et al., 2007. Characterization of microRNA expression levels and their biological correlates in human cancer cell lines. *Cancer research*, 67(6), pp.2456–68.

Gennarino, V.A. et al., 2009. MicroRNA target prediction by expression analysis of host genes. *Genome research*, 19(3), pp.481–90.

Goeman, J.J. & Bühlmann, P., 2007. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics (Oxford, England)*, 23(8), pp.980–7.

Guo, M.-M. et al., 2013. miR-22 is down-regulated in gastric cancer, and its overexpression inhibits cell migration and invasion via targeting transcription factor Sp1. *Medical oncology (Northwood, London, England)*, 30(2), p.542.

Ha, M. & Kim, V.N., 2014. Regulation of microRNA biogenesis. *Nature reviews. Molecular cell biology*, 15(8), pp.509–524.

Hammell, M., 2010. Computational methods to identify miRNA targets. *Seminars in cell & developmental biology*, 21(7), pp.738–44.

Hanahan, D. & Weinberg, R. a, 2011. Hallmarks of cancer: the next generation. *Cell*, 144(5), pp.646–74.

Hanahan, D. & Weinberg, R.A., 2000. The Hallmarks of Cancer. , 100, pp.57–70.

Hanai, J.-I. et al., 2013. ATP citrate lyase knockdown impacts cancer stem cells in vitro. *Cell death & disease*, 4(6), p.e696.

Hausser, J. & Zavolan, M., 2014. Identification and consequences of miRNA-target interactions - beyond repression of gene expression. *Nature reviews. Genetics*, 15(9), pp.599–612.

He, L. & Hannon, G.J., 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nature reviews. Genetics*, 5(7), pp.522–31.

Heiden, M.G. Vander, Cantley, L.C. & Thompson, C.B., 2009. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science (New York, N.Y.)*, 324(5930), pp.1029–33.

Hinske, L.C.G. et al., 2010. A potential role for intragenic miRNAs on their hosts' interactome. *BMC genomics*, 11, p.533.

Hua, Y. et al., 2013. *miRConnect 2.0: identification of oncogenic, antagonistic miRNA families in three human cancers.*,

Hua, Y. et al., 2011. miRConnect: Identifying Effector Genes of miRNAs and miRNA Families in Cancer Cells L. Zhang, ed. *PLoS ONE*, 6(10), p.e26521.

Huang, Z.-P. et al., 2013. MicroRNA-22 regulates cardiac hypertrophy and remodeling in response to stress. *Circulation research*, 112(9), pp.1234–43.

John, B. et al., 2004. Human MicroRNA targets. *PLoS biology*, 2(11), p.e363.

Jovicic, A. et al., 2013. MicroRNA-22 (miR-22) overexpression is neuroprotective via general anti-apoptotic effects and may also target specific Huntington's disease-related mechanisms. *PloS one*, 8(1), p.e54222.

Kanehisa, M. & Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), pp.27–30.

Karolchik, D. et al., 2014. The UCSC Genome Browser database: 2014 update. *Nucleic acids research*, 42(Database issue), pp.D764–70.

Kendall, M. & Gibbons, J.D., 1990. *Rank correlation methods*, Oxford University Press.

Kertesz, M. et al., 2007. The role of site accessibility in microRNA target recognition. *Nature genetics*, 39(10), pp.1278–84.

Khatri, P., Sirota, M. & Butte, A.J., 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2), p.e1002375.

Kim, V.N., 2005. MicroRNA biogenesis: coordinated cropping and dicing. *Nature reviews. Molecular cell biology*, 6(5), pp.376–85.

Kim, V.N., Han, J. & Siomi, M.C., 2009. Biogenesis of small RNAs in animals. *Nature reviews. Molecular cell biology*, 10(2), pp.126–39.

Kleinbaum, D.G. & Klein, M., 2005a. Kaplan–Meier Survival Curves and the Log–Rank Test. In *Survival analysis : a self-learning text*. Springer, pp. 45–82.

Kleinbaum, D.G. & Klein, M., 2005b. The Cox Proportional Hazards Model and Its Characteristics. In *Survival analysis : a self-learning text*. Springer, pp. 83–130.

Kong, L.-M. et al., 2014. A regulatory loop involving miR-22, Sp1, and c-Myc modulates CD147 expression in breast cancer invasion and metastasis. *Cancer research*, 74(14), pp.3764–78.

Koppenol, W.H., Bounds, P.L. & Dang, C. V, 2011. Otto Warburg's contributions to current concepts of cancer metabolism. *Nature reviews. Cancer*, 11(5), pp.325–37.

Kozomara, A. & Griffiths-Jones, S., 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, 42(Database issue), pp.D68–73.

Krek, A. et al., 2005. Combinatorial microRNA target predictions. *Nature genetics*, 37(5), pp.495–500.

Lagos-Quintana, M. et al., 2002. Identification of tissue-specific microRNAs from mouse. *Current biology : CB*, 12(9), pp.735–9.

Langfelder, P., Zhang, B. & Horvath, S., 2008. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics (Oxford, England)*, 24(5), pp.719–20.

Larose, D.T., 2005. *Discovering knowledge in data: an introduction to data mining*, John Wiley & Sons.

Lee, R.C., Feinbaum, R.L. & Ambros, V., 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5), pp.843–54.

Lehtinen, L. et al., 2013. High-throughput RNAi screening for novel modulators of vimentin expression identifies MTHFD2 as a regulator of breast cancer cell migration and invasion. *Oncotarget*, 4(1), pp.48–63.

Lenkala, D. et al., 2014. The impact of microRNA expression on cellular proliferation. *Human genetics*, 133(7), pp.931–8.

Leonard, T., 2000. *A course in categorical data analysis*, Chapman & Hall/CRC.

Di Leva, G. et al., 2010. MicroRNA cluster 221-222 and estrogen receptor alpha interactions in breast cancer. *Journal of the National Cancer Institute*, 102(10), pp.706–21.

Levandowsky, M. & Winter, D., 1971. Distance between sets. *Nature*, 234, pp.34–35.

Li, J. et al., 2010. An inhibitory effect of miR-22 on cell migration and invasion in ovarian cancer. *Gynecologic oncology*, 119(3), pp.543–8.

Li, X. et al., 2014. miR-22 targets the 3' UTR of HMGB1 and inhibits the HMGB1-associated autophagy in osteosarcoma cells during chemotherapy. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, 35(6), pp.6021–8.

Li, Y. et al., 2014. Mirsynergy: detecting synergistic miRNA regulatory modules by overlapping neighbourhood expansion. *Bioinformatics (Oxford, England)*, (2011), pp.1–9.

Ling, B. et al., 2012. Tumor suppressor miR-22 suppresses lung cancer cell progression through post-transcriptional regulation of ErbB3. *Journal of cancer research and clinical oncology*, 138(8), pp.1355–61.

Liu, F. et al., 2014. Increased MTHFD2 expression is associated with poor prognosis in breast cancer. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, 35(9), pp.8685–90.

Liu, H. et al., 2010. mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Molecular cancer therapeutics*, 9(5), pp.1080–91.

Liu, J., 2008. Control of protein synthesis and mRNA degradation by microRNAs. *Current opinion in cell biology*, 20(2), pp.214–21.

Malumbres, M., 2013. miRNAs and cancer: an epigenetics view. *Molecular aspects of medicine*, 34(4), pp.863–74.

Martinez-Sanchez, A. & Murphy, C.L., 2013. MicroRNA Target Identification-Experimental Approaches. *Biology*, 2(1), pp.189–205.

Melamed, Z. et al., 2013. Alternative splicing regulates biogenesis of miRNAs located across exon-intron junctions. *Molecular cell*, 50(6), pp.869–81.

Melo, S. a & Esteller, M., 2011. Dysregulation of microRNAs in cancer: playing with fire. *FEBS letters*, 585(13), pp.2087–99.

Migita, T. et al., 2014. Inhibition of ATP citrate lyase induces triglyceride accumulation with altered fatty acid composition in cancer cells. *International journal of cancer. Journal international du cancer*, 135(1), pp.37–47.

Miller, D.M. et al., 2012. c-Myc and cancer metabolism. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 18(20), pp.5546–53.

Miranda, K.C. et al., 2006. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, 126(6), pp.1203–17.

Mogilyansky, E. & Rigoutsos, I., 2013. The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell death and differentiation*, 20(12), pp.1603–14.

Morozova, N. et al., 2012. Kinetic signatures of microRNA modes of action. *RNA (New York, N.Y.)*, 18(9), pp.1635–55.

Mortazavi, A., Williams, B. & McCue, K., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5, pp.621–628.

Nam, J.-W. et al., 2014. Global analyses of the effect of different cellular contexts on microRNA targeting. *Molecular cell*, 53(6), pp.1031–43.

National Cancer Institute, 2014. Molecular Target Data. Available at: https://wiki.nci.nih.gov/display/NCIDTPdata/Molecular+Target+Data [Accessed April 24, 2014].

Nilsson, R. et al., 2014. Metabolic enzyme expression highlights a key role for MTHFD2 and the mitochondrial folate pathway in cancer. *Nature communications*, 5, p.3128.

Pandey, D.P. & Picard, D., 2009. miR-22 inhibits estrogen signaling by directly targeting the estrogen receptor alpha mRNA. *Molecular and cellular biology*, 29(13), pp.3783–90.

Pasquinelli, a E. et al., 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808), pp.86–9.

Patel, J.B. et al., 2011. Control of EVI-1 oncogene expression in metastatic breast cancer cells through microRNA miR-22. *Oncogene*, 30(11), pp.1290–301.

Pawelek, P.D. & MacKenzie, R.E., 1998. Methenyltetrahydrofolate cyclohydrolase is rate limiting for the enzymatic conversion of 10-formyltetrahydrofolate to 5,10-methylenetetrahydrofolate in bifunctional dehydrogenase-cyclohydrolase enzymes. *Biochemistry*, 37(4), pp.1109–15.

Peter, M.E., 2010. Targeting of mRNAs by multiple miRNAs: the next step. *Oncogene*, 29(15), pp.2161–4.

Peterson, S.M. et al., 2014. Common features of microRNA target prediction tools. *Frontiers in genetics*, 5(February), p.23.

Pillai, R.S., Bhattacharyya, S.N. & Filipowicz, W., 2007. Repression of protein synthesis by miRNAs: how many mechanisms? *Trends in cell biology*, 17(3), pp.118–26.

Polioudakis, D. et al., 2013. A Myc-microRNA network promotes exit from quiescence by suppressing the interferon response and cell-cycle arrest genes. *Nucleic acids research*, 41(4), pp.2239–54.

Pomyen, Y. et al., 2014. Over-representation of correlation analysis (ORCA): a method for identifying associations between variable sets. *Bioinformatics*, pp.1–7.

Psathas, J.N. & Thomas-Tikhonenko, A., 2014. MYC and the art of microRNA maintenance. *Cold Spring Harbor perspectives in medicine*, 4(8).

Rajaram, S. & Oono, Y., 2010. NeatMap--non-clustering heat map alternatives in R. *BMC bioinformatics*, 11(1), p.45.

Reczko, M. et al., 2012. Functional microRNA targets in protein coding sequences. *Bioinformatics (Oxford, England)*, 28(6), pp.771–6.

Rehmsmeier, M. et al., 2004. Fast and effective prediction of microRNA / target duplexes. , pp.1507–1517.

Reinhart, B.J. et al., 2000. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772), pp.901–6.

Resendis-Antonio, O. et al., 2015. Modeling metabolism: A window toward a comprehensive interpretation of networks in cancer. *Seminars in cancer biology*, 30, pp.79–87.

Ritchie, W. & Rasko, J.E.J., 2014. Refining microRNA target predictions: sorting the wheat from the chaff. *Biochemical and biophysical research communications*, 445(4), pp.780–4.

Robinson, M.D. & Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3), p.R25.

Rodriguez, A. et al., 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Research*, 14, pp.1902–1910.

Rottiers, V. & Näär, A.M., 2012. MicroRNAs in metabolism and metabolic disorders. *Nature reviews. Molecular cell biology*, 13(4), pp.239–50.

Rudolf, E., Rudolf, K. & Cervinka, M., 2011. Camptothecin induces p53-dependent and -independent apoptogenic signaling in melanoma cells. *Apoptosis : an international journal on programmed cell death*, 16(11), pp.1165–76.

Ruike, Y. et al., 2008. Global correlation analysis for micro-RNA and mRNA expression profiles in human cell lines. *Journal of human genetics*, 53(6), pp.515–23.

Saito, T. & Saetrom, P., 2010. MicroRNAs--targeting and target prediction. *New biotechnology*, 27(3), pp.243–9.

Sarkies, P. & Miska, E. a., 2014. Small RNAs break out: the molecular cell biology of mobile small RNAs. *Nature Reviews Molecular Cell Biology*, 15(8), pp.525–535.

Scherf, U. et al., 2000. A gene expression database for the molecular pharmacology of cancer. *nature genetics*, 24(3), pp.236–44.

Selcuklu, S.D. et al., 2012. MicroRNA-9 inhibition of cell proliferation and identification of novel miR-9 targets by transcriptome profiling in breast cancer cells. *The Journal of biological chemistry*, 287(35), pp.29516–28.

Sempere, L.F. et al., 2004. Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome biology*, 5(3), p.R13.

Sethi, P. & Alagiriswamy, S., 2010. Association rule based similarity measures for the clustering of gene expression data. *The open medical informatics journal*, 4, pp.63–73.

Seyfried, T.N. et al., 2014. Cancer as a metabolic disease: implications for novel therapeutics. *Carcinogenesis*, 35(3), pp.515–27.

Shankavaram, U.T. et al., 2007. Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Molecular cancer therapeutics*, 6(3), pp.820–32.

Shannon, C., 1948. A mathematical theory of communination. *The bell system technical journal*, 27(3), p.379.

Shiau, a. L. et al., 2014. Elovl6 overexpression promotes liver carcinogenesis. *European Journal of Cancer*, 50, p.S75.

Sikand, K., Slane, S.D. & Shukla, G.C., 2009. Intrinsic expression of host genes and intronic miRNAs in prostate carcinoma cells. *Cancer Cell International*, 9(1), p.21.

Slezak-Prochazka, I. et al., 2013. Cellular localization and processing of primary transcripts of exonic microRNAs. *PloS one*, 8(9), p.e76647.

Søkilde, R. et al., 2011. Global microRNA analysis of the NCI-60 cancer cell panel. *Molecular cancer therapeutics*, 10(3), pp.375–84.

Song, S.J., Poliseno, L., et al., 2013. MicroRNA-antagonism regulates breast cancer stemness and metastasis via TET-family-dependent chromatin remodeling. *Cell*, 154(2), pp.311–24.

Song, S.J., Ito, K., et al., 2013. The oncogenic microRNA miR-22 targets the TET2 tumor suppressor to promote hematopoietic stem cell self-renewal and transformation. *Cell stem cell*, 13(1), pp.87–101.

Sun, W. et al., 2010. microRNA: a master regulator of cellular processes for bioengineering systems. *Annual review of biomedical engineering*, 12, pp.1–27.

Sun, Z. & Zhu, Y., 2012. Systematic comparison of RNA-Seq normalization methods using measurement error models. *Bioinformatics (Oxford, England)*, 28(20), pp.2584–91.

Tavazoie, S. et al., 1999. Systematic determination of genetic network architecture. *Nature genetics*, 22(3), pp.281–5.

Tenenbaum, D., 2013. KEGGREST: Client-side REST access to KEGG.

The Cancer Genome Atlas Network, 2012. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), pp.61–70.

The ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74.

Thomson, D.W., Bracken, C.P. & Goodall, G.J., 2011. Experimental strategies for microRNA target identification. *Nucleic acids research*, 39(16), pp.6845–53.

Tibiche, C. & Wang, E., 2008. MicroRNA Regulatory Patterns on the Human Metabolic Network. *The Open Systems Biology Journal*, 1, pp.1–8.

Ting, Y. et al., 2010. Differentiation-associated miR-22 represses Max expression and inhibits cell cycle progression. *Biochemical and biophysical research communications*, 394(3), pp.606–11.

Tomasetti, M. et al., 2014. MicroRNA regulation of cancer metabolism: role in tumour suppression. *Mitochondrion*, 19, Part A, pp.29–38.

Veronese, a et al., 2014. Allele-specific loss and transcription of the miR-15a/16-1 cluster in chronic lymphocytic leukemia. *Leukemia*, (April), pp.1–10.

Wang, J. et al., 2011. Microarray profiling of monocytic differentiation reveals miRNA-mRNA intrinsic correlation. *Journal of cellular biochemistry*, 112(9), pp.2443–53.

Wang, W. et al., 2013. Integrative network-based Bayesian analysis of diverse genomics data. *BMC bioinformatics*, 14 Suppl 1(Suppl 13), p.S8.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), pp.236–244.

Ward, P.S. & Thompson, C.B., 2012. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer cell*, 21(3), pp.297–308.

Whitlock, M.C., 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher’s approach. *Journal of evolutionary biology*, 18(5), pp.1368–73.

Wightman, B., Ha, I. & Ruvkun, G., 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5), pp.855–62.

Witkos, T.M., Koscianska, E. & Krzyzosiak, W.J., 2011. Practical Aspects of microRNA Target Prediction. *Current molecular medicine*, 11(2), pp.93–109.

Xiong, J. et al., 2010. An estrogen receptor alpha suppressor, microRNA-22, is downregulated in estrogen receptor alpha-positive human breast cancer cell lines and clinical samples. *The FEBS journal*, 277(7), pp.1684–94.

Xiong, J., Du, Q. & Liang, Z., 2010. Tumor-suppressive microRNA-22 inhibits the transcription of E-box-containing c-Myc target genes by silencing c-Myc binding protein. *Oncogene*, 29(35), pp.4980–8.

Xu, D. et al., 2011. miR-22 represses cancer progression by inducing cellular senescence. *The Journal of cell biology*, 193(2), pp.409–24.

Xu, J. et al., 2011. MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. *Nucleic acids research*, 39(3), pp.825–36.

Yoon, S. & De Micheli, G., 2005. Prediction of regulatory modules comprising microRNAs and target genes. *Bioinformatics (Oxford, England)*, 21 Suppl 2, pp.ii93–100.

Yoshimoto, N. et al., 2011. Distinct expressions of microRNAs that directly target estrogen receptor α in human breast cancer. *Breast cancer research and treatment*, 130(1), pp.331–9.

Zeeberg, B.R. et al., 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome biology*, 4(4), p.R28.

Zhang, G. et al., 2012. Clinical significance of miR-22 expression in patients with colorectal cancer. *Medical oncology (Northwood, London, England)*, 29(5), pp.3108–12.

Zhang, J. et al., 2010. microRNA-22, downregulated in hepatocellular carcinoma and correlated with prognosis, suppresses cell proliferation and tumourigenicity. *British journal of cancer*, 103(8), pp.1215–20.

Zhang, Z. et al., 2009. MicroRNA miR-210 modulates cellular response to hypoxia through the MYC antagonist MNT. *Cell cycle (Georgetown, Tex.)*, 8(17), pp.2756–68.

Zhu, Y. et al., 2012. MicroRNA-26a/b and their host genes cooperate to inhibit the G1/S transition by activating the pRb protein. *Nucleic acids research*, 40(10), pp.4615–25.

Appendices

Appendix A1

Table A1.1 List of miRNA in cluster 1.

miRNA in cluster1	miR-451
	miR-551a
	miR-187
	miR-210

Table A1.2 List of miRNA in cluster 2.

miRNA in cluster 2	miR-99ab/100
	miR-191
	miR-184
	miR-122/122a/1352
	miR-126-3p

Table A1.2 List of miRNA in cluster 3.

miRNA in cluster 3	miR-499-5p
	miR-208ab/208ab-3p
	miR-18ab/4735-3p
	miR-383
	miR-193/193b/193a-3p
	miR-196abc
	miR-10abc/10a-5p
	miR-33a-3p/365/365-3p
	miR-455-5p
	miR-190/190ab
	miR-223
	miR-155
	miR-153
	miR-33ab/33-5p
	miR-375
	miR-142-3p
	miR-21/590-5p
	miR-139-5p
	miR-221/222/222ab/1928
	miR-219-5p/508/508-3p/4782-3p
	miR-216a
	miR-216b/216b-5p

Table A1.3 List of miRNA in cluster 4.

miRNA in cluster 4	miR-23abc/23b-3p
	miR-25/32/92abc/363/363-3p/367
	miR-26ab/1297/4465
	miR-30abcdef/30abe-5p/384-5p
	miR-101/101ab
	miR-132/212/212-3p
	miR-137/137ab
	miR-141/200a
	miR-144
	miR-181abcd/4262
	miR-194
	miR-200bc/429/548a
	miR-203
	miR-217

Table A1.4 List of miRNA in cluster 5.

miRNA in cluster 5	miR-96/507/1271
	miR-182
	miR-148ab-3p/152
	miR-130ac/301ab/301b/301b-3p/454/721/4295/3666
	miR-19ab
	miR-490-3p
	miR-146ac/146b-5p
	miR-34ac/34bc-5p/449abc/449c-5p
	miR-103a/107/107ab
	let-7/98/4458/4500
	miR-425/425-5p/489
	miR-31
	miR-183
	miR-205/205ab
	miR-192/215
	miR-140/140-5p/876-3p/1244

Table A1.5 List of miRNA in cluster 6.

miRNA in cluster 6	miR-1ab/206/613
	miR-7/7ab
	miR-9/9ab
	miR-15abc/16/16abc/195/322/424/497/1907
	miR-17/17-5p/20ab/20b-5p/93/106ab/427/518a-3p/519d
	miR-22/22-3p
	miR-24/24ab/24-3p
	miR-27abc/27a-3p
	miR-29abcd
	miR-93/93a/105/106a/291a-3p/294/295/302abcde/372/373/428/519a/520be/520acd-3p/1378/1420ac
	miR-124/124ab/506
	miR-125a-5p/125b-5p/351/670/4319
	miR-128/128ab
	miR-129-5p/129ab-5p
	miR-133abc
	miR-135ab/135a-5p
	miR-138/138ab
	miR-143/1721/4770
	miR-145
	miR-150/5127
	miR-199ab-5p
	miR-204/204b/211
	miR-214/761/3619-5p
	miR-218/218a
	miR-338/338-3p
	miR-503

Table A1.6 Over-represented pathways in cluster 1.

Pathway name	Number of occurrences	p-value	Adjusted p-value
Purine metabolism	1	0.35	0.35
Rap1 signaling pathway	1	0.16	0.32
cGMP-PKG signaling pathway	1	0.23	0.32
cAMP signaling pathway	2	0.04	0.32
Chemokine signaling pathway	1	0.31	0.34
Adrenergic signaling in cardiomyocytes	1	0.11	0.32
Platelet activation	1	0.10	0.32
Glutamatergic synapse	1	0.21	0.32
Cholinergic synapse	1	0.29	0.34
Inflammatory mediator regulation of TRP channels	1	0.27	0.34
Progesterone-mediated oocyte maturation	1	0.33	0.34
Estrogen signaling pathway	1	0.23	0.32
Oxytocin signaling pathway	1	0.17	0.32
Pancreatic secretion	1	0.17	0.32
Bile secretion	1	0.20	0.32
Morphine addiction	1	0.21	0.32
Chagas disease (American trypanosomiasis)	1	0.19	0.32
Amoebiasis	1	0.23	0.32
HTLV-I infection	1	0.19	0.32

Table A1.7 Over-represented pathways in cluster 2.

Pathway name	Number of occurrences	p.value	Adjusted p-value
Purine metabolism	1	0.46	0.46
Arginine and proline metabolism	1	0.23	0.43
Glycosaminoglycan biosynthesis - heparan sulfate / heparin	1	0.31	0.43
Glycerolipid metabolism	1	0.37	0.43
Glycerophospholipid metabolism	2	0.10	0.43
Glycosphingolipid biosynthesis - ganglio series	1	0.15	0.43
Drug metabolism - other enzymes	1	0.14	0.43
Calcium signaling pathway	1	0.29	0.43
cGMP-PKG signaling pathway	1	0.39	0.43
cAMP signaling pathway	1	0.37	0.43
Phosphatidylinositol signaling system	1	0.44	0.46
Adrenergic signaling in cardiomyocytes	2	0.06	0.42
Vascular smooth muscle contraction	1	0.24	0.43
Gap junction	1	0.24	0.43
Platelet activation	1	0.28	0.43
Circadian entrainment	2	0.04	0.42
Insulin secretion	1	0.27	0.43
Thyroid hormone synthesis	1	0.31	0.43
Oxytocin signaling pathway	1	0.26	0.43
Salivary secretion	3	0.01	0.29
Gastric acid secretion	1	0.31	0.43
Pancreatic secretion	2	0.07	0.42
Bile secretion	1	0.36	0.43

Table A1.8 Over-represented pathways in cluster 3.

pathway name	Number of occurrences	p-value	Adjusted p-value
Purine metabolism	199	2.48E-12	4.76E-10
Glycosphingolipid biosynthesis - lacto and neolacto series	104	9.36E-07	8.99E-05
Morphine addiction	132	6.60E-06	0.0004
Valine, leucine and isoleucine biosynthesis	61	3.68E-05	0.002
Pantothenate and CoA biosynthesis	70	0.0001	0.005
Glycosphingolipid biosynthesis - globo series	92	0.0002	0.006
Pancreatic secretion	120	0.0002	0.006
GABAergic synapse	104	0.0004	0.01
cAMP signaling pathway	120	0.0006	0.01
Salivary secretion	127	0.0006	0.01
Lysine degradation	68	0.0007	0.01
Phosphatidylinositol signaling system	141	0.002	0.02
Valine, leucine and isoleucine degradation	102	0.002	0.03
alpha-Linolenic acid metabolism	31	0.002	0.03
Retinol metabolism	37	0.003	0.04
Nicotinate and nicotinamide metabolism	41	0.006	0.07
Other types of O-glycan biosynthesis	33	0.007	0.08
Alanine, aspartate and glutamate metabolism	68	0.007	0.08
Gap junction	70	0.01	0.13
Biotin metabolism	13	0.02	0.19
Glutamatergic synapse	104	0.02	0.20
Peroxisome	46	0.03	0.25
Cardiac muscle contraction	29	0.03	0.25
Fatty acid degradation	73	0.04	0.31
cGMP-PKG signaling pathway	112	0.04	0.31
Circadian entrainment	78	0.04	0.32
Protein digestion and absorption	37	0.04	0.32
Bile secretion	102	0.05	0.34
Chagas disease (American trypanosomiasis)	40	0.05	0.34
Glycosaminoglycan biosynthesis - heparan sulfate / heparin	86	0.06	0.39
FoxO signaling pathway	26	0.07	0.42
Amoebiasis	38	0.07	0.42
Pentose and glucuronate interconversions	9	0.08	0.47
PPAR signaling pathway	53	0.09	0.47
Inflammatory mediator regulation of TRP channels	67	0.09	0.47
Estrogen signaling pathway	66	0.09	0.47
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	33	0.09	0.49
Glycerophospholipid metabolism	115	0.10	0.49
Progesterone-mediated oocyte maturation	66	0.10	0.49
Thyroid hormone synthesis	83	0.10	0.49
Citrate cycle (TCA cycle)	12	0.12	0.54
Aldosterone-regulated sodium reabsorption	38	0.12	0.55
Insulin secretion	69	0.13	0.58
Propanoate metabolism	23	0.15	0.64
Cholinergic synapse	63	0.15	0.64
Oocyte meiosis	27	0.16	0.68
Nitrogen metabolism	15	0.17	0.71
GnRH signaling pathway	48	0.21	0.83
Melanogenesis	42	0.21	0.83
Adrenergic signaling in cardiomyocytes	81	0.22	0.84

Table A1.9 Over-represented pathways in cluster 3.

pathway name	Number of occurrences	p-value	Adjusted p-value
Proximal tubule bicarbonate reclamation	32	0.26	1.00
Glycolysis / Gluconeogenesis	21	1.00	1.00
Pentose phosphate pathway	15	0.74	1.00
Fructose and mannose metabolism	18	0.89	1.00
Galactose metabolism	4	1.00	1.00
Ascorbate and aldarate metabolism	2	0.97	1.00
Fatty acid biosynthesis	3	1.00	1.00
Steroid biosynthesis	2	0.83	1.00
Primary bile acid biosynthesis	1	0.99	1.00
Steroid hormone biosynthesis	14	0.89	1.00
Oxidative phosphorylation	17	0.99	1.00
Caffeine metabolism	3	0.99	1.00
Pyrimidine metabolism	64	0.49	1.00
Glycine, serine and threonine metabolism	6	0.98	1.00
Cysteine and methionine metabolism	23	1.00	1.00
Lysine biosynthesis	2	1.00	1.00
Arginine and proline metabolism	41	0.90	1.00
Histidine metabolism	4	1.00	1.00
Tyrosine metabolism	2	1.00	1.00
Tryptophan metabolism	7	1.00	1.00
Phenylalanine, tyrosine and tryptophan biosynthesis	1	1.00	1.00
beta-Alanine metabolism	19	0.91	1.00
Taurine and hypotaurine metabolism	1	0.86	1.00
Selenocompound metabolism	16	0.84	1.00
Cyanoamino acid metabolism	1	0.38	1.00
D-Glutamine and D-glutamate metabolism	15	0.64	1.00
Glutathione metabolism	4	0.78	1.00
Starch and sucrose metabolism	21	0.57	1.00
N-Glycan biosynthesis	51	0.76	1.00
Other glycan degradation	1	0.67	1.00
Mucin type O-Glycan biosynthesis	67	0.56	1.00
Amino sugar and nucleotide sugar metabolism	53	0.34	1.00
Butirosin and neomycin biosynthesis	1	1.00	1.00
Glycosaminoglycan degradation	6	0.99	1.00
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	20	1.00	1.00
Glycosaminoglycan biosynthesis - keratan sulfate	11	1.00	1.00
Glycerolipid metabolism	92	0.36	1.00
Inositol phosphate metabolism	92	0.46	1.00
Ether lipid metabolism	36	0.88	1.00
Arachidonic acid metabolism	19	0.96	1.00
Linoleic acid metabolism	5	0.92	1.00
Sphingolipid metabolism	41	0.70	1.00
Glycosphingolipid biosynthesis - ganglio series	20	0.99	1.00
Pyruvate metabolism	14	0.97	1.00
Glyoxylate and dicarboxylate metabolism	3	0.31	1.00
Butanoate metabolism	13	0.95	1.00
One carbon pool by folate	8	1.00	1.00
Thiamine metabolism	5	0.56	1.00
Riboflavin metabolism	2	0.73	1.00
Vitamin B6 metabolism	10	0.83	1.00
Folate biosynthesis	3	0.29	1.00
Porphyrin and chlorophyll metabolism	5	0.85	1.00
Terpenoid backbone biosynthesis	2	0.47	1.00

Table A1.9 Over-represented pathways in cluster 3.

pathway name	Number of occurrences	p-value	Adjusted p-value
Metabolism of xenobiotics by cytochrome P450	3	0.98	1.00
Drug metabolism - cytochrome P450	3	0.98	1.00
Drug metabolism - other enzymes	6	1.00	1.00
Biosynthesis of unsaturated fatty acids	13	0.88	1.00
ABC transporters	1	1.00	1.00
Protein export	1	0.83	1.00
ErbB signaling pathway	9	0.94	1.00
Ras signaling pathway	4	1.00	1.00
Rap1 signaling pathway	24	1.00	1.00
Calcium signaling pathway	48	0.99	1.00
Chemokine signaling pathway	25	1.00	1.00
HIF-1 signaling pathway	24	0.81	1.00
Lysosome	7	0.99	1.00
Phagosome	7	0.99	1.00
mTOR signaling pathway	13	0.98	1.00
AMPK signaling pathway	35	0.71	1.00
Apoptosis	6	0.98	1.00
Vascular smooth muscle contraction	56	0.38	1.00
VEGF signaling pathway	10	1.00	1.00
Osteoclast differentiation	5	0.66	1.00
Focal adhesion	2	0.98	1.00
Platelet activation	64	0.44	1.00
Toll-like receptor signaling pathway	5	0.95	1.00
Jak-STAT signaling pathway	4	0.49	1.00
Natural killer cell mediated cytotoxicity	6	0.64	1.00
T cell receptor signaling pathway	8	0.86	1.00
B cell receptor signaling pathway	7	0.99	1.00
Fc epsilon RI signaling pathway	10	0.98	1.00
Fc gamma R-mediated phagocytosis	19	0.87	1.00
TNF signaling pathway	5	0.98	1.00
Leukocyte transendothelial migration	8	0.61	1.00
Long-term potentiation	12	0.91	1.00
Synaptic vesicle cycle	10	1.00	1.00
Neurotrophin signaling pathway	8	0.60	1.00
Retrograde endocannabinoid signaling	45	0.60	1.00
Serotonergic synapse	8	0.88	1.00
Dopaminergic synapse	2	0.92	1.00
Long-term depression	20	0.99	1.00
Regulation of actin cytoskeleton	5	1.00	1.00
Insulin signaling pathway	12	0.99	1.00
Ovarian steroidogenesis	47	0.80	1.00
Prolactin signaling pathway	6	1.00	1.00
Thyroid hormone signaling pathway	42	0.67	1.00
Adipocytokine signaling pathway	7	1.00	1.00
Oxytocin signaling pathway	47	0.93	1.00
Type II diabetes mellitus	6	1.00	1.00
Non-alcoholic fatty liver disease (NAFLD)	16	0.80	1.00
Endocrine and other factor-regulated calcium reabsorption	42	0.29	1.00
Vasopressin-regulated water reabsorption	4	0.77	1.00
Collecting duct acid secretion	5	1.00	1.00
Gastric acid secretion	74	0.36	1.00
Carbohydrate digestion and absorption	42	0.80	1.00
Fat digestion and absorption	6	1.00	1.00
Vitamin digestion and absorption	7	1.00	1.00

Table A1.9 Over-represented pathways in cluster 3.

pathway name	Number of occurrences	p-value	Adjusted p-value
Mineral absorption	33	0.94	1.00
Alzheimer's disease	14	0.99	1.00
Parkinson's disease	7	1.00	1.00
Amyotrophic lateral sclerosis (ALS)	36	0.40	1.00
Huntington's disease	9	0.98	1.00
Cocaine addiction	1	0.99	1.00
Amphetamine addiction	1	0.99	1.00
Alcoholism	1	0.64	1.00
Bacterial invasion of epithelial cells	6	0.99	1.00
Vibrio cholerae infection	11	1.00	1.00
Epithelial cell signaling in Helicobacter pylori infection	7	1.00	1.00
Toxoplasmosis	5	0.86	1.00
Hepatitis C	5	0.66	1.00
Hepatitis B	7	0.36	1.00
Measles	5	0.66	1.00
Influenza A	1	0.90	1.00
HTLV-I infection	14	1.00	1.00
Epstein-Barr virus infection	10	0.83	1.00
Pathways in cancer	2	0.97	1.00
Viral carcinogenesis	1	0.85	1.00
Chemical carcinogenesis	3	1.00	1.00
MicroRNAs in cancer	2	0.99	1.00
Colorectal cancer	6	0.99	1.00
Renal cell carcinoma	7	1.00	1.00
Pancreatic cancer	12	0.92	1.00
Endometrial cancer	13	0.98	1.00
Glioma	16	0.97	1.00
Prostate cancer	13	0.86	1.00
Melanoma	13	0.97	1.00
Chronic myeloid leukemia	6	0.99	1.00
Acute myeloid leukemia	6	0.99	1.00
Small cell lung cancer	13	0.96	1.00
Non-small cell lung cancer	9	0.98	1.00
Rheumatoid arthritis	6	0.91	1.00
Dilated cardiomyopathy	31	0.61	1.00

Table A1.9 Over-represented pathways in cluster 4.

Pathway name	Number of occurrences	p-value	Adjusted p-value
MicroRNAs in cancer	28	2.91E-11	5.74E-09
Pyruvate metabolism	48	1.55E-07	1.02E-05
Vibrio cholerae infection	57	1.50E-07	1.02E-05
Citrate cycle (TCA cycle)	25	3.41E-07	1.68E-05
Endometrial cancer	44	3.21E-06	0.0001
Protein export	10	5.14E-06	0.0002
Synaptic vesicle cycle	47	1.59E-05	0.0004
Peroxisome	58	3.02E-05	0.0007
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	46	6.31E-05	0.001
mTOR signaling pathway	40	4.95E-05	0.001
Collecting duct acid secretion	36	5.45E-05	0.001
Protein digestion and absorption	48	6.53E-05	0.001
Melanoma	37	0.0002	0.003
Tyrosine metabolism	25	0.0003	0.004
Focal adhesion	15	0.0003	0.004
Glutathione metabolism	14	0.0006	0.007
Starch and sucrose metabolism	36	0.001	0.01
Small cell lung cancer	34	0.001	0.01
Vitamin digestion and absorption	28	0.001	0.01
Chemical carcinogenesis	20	0.001	0.01
Glioma	39	0.001	0.01
Adipocytokine signaling pathway	33	0.002	0.02
Pathways in cancer	13	0.002	0.02
Valine, leucine and isoleucine biosynthesis	51	0.003	0.02
PI3K-Akt signaling pathway	9	0.003	0.02
Other glycan degradation	5	0.004	0.03
Phagosome	25	0.004	0.03
Proximal tubule bicarbonate reclamation	42	0.004	0.03
Primary bile acid biosynthesis	11	0.005	0.03
Fat digestion and absorption	34	0.005	0.03
Oxidative phosphorylation	42	0.005	0.03
Glycolysis / Gluconeogenesis	51	0.005	0.03
FoxO signaling pathway	30	0.006	0.04
Serotonergic synapse	20	0.007	0.04
Phenylalanine metabolism	12	0.007	0.04
Histidine metabolism	22	0.009	0.05
Hepatitis B	12	0.01	0.06
VEGF signaling pathway	33	0.01	0.06
D-Glutamine and D-glutamate metabolism	25	0.01	0.07
Pantothenate and CoA biosynthesis	57	0.02	0.08
Parkinson's disease	26	0.02	0.08
Epithelial cell signaling in Helicobacter pylori infection	25	0.02	0.08
Regulation of actin cytoskeleton	25	0.02	0.08
Fatty acid degradation	73	0.02	0.09
Ether lipid metabolism	55	0.02	0.10
Steroid hormone biosynthesis	27	0.02	0.11
Nitrogen metabolism	18	0.03	0.12
Renal cell carcinoma	23	0.03	0.12
Taurine and hypotaurine metabolism	5	0.04	0.14
Glycosphingolipid biosynthesis - ganglio series	41	0.04	0.14
Riboflavin metabolism	6	0.04	0.14

Table A1.10 Over-represented pathways in cluster 4.

Pathway name	Number of occurrences	p-value	Adjusted p-value
TNF signaling pathway	16	0.04	0.17
Propanoate metabolism	25	0.05	0.18
Cocaine addiction	9	0.05	0.20
Drug metabolism - cytochrome P450	12	0.06	0.20
Phenylalanine, tyrosine and tryptophan biosynthesis	11	0.06	0.22
Rheumatoid arthritis	14	0.07	0.25
HIF-1 signaling pathway	35	0.07	0.25
Glycine, serine and threonine metabolism	17	0.09	0.29
Porphyrin and chlorophyll metabolism	11	0.09	0.29
Metabolism of xenobiotics by cytochrome P450	11	0.09	0.29
Tryptophan metabolism	31	0.10	0.31
ABC transporters	15	0.10	0.31
Viral carcinogenesis	4	0.10	0.32
p53 signaling pathway	1	0.11	0.33
PPAR signaling pathway	50	0.12	0.33
Insulin signaling pathway	26	0.11	0.33
Colorectal cancer	17	0.11	0.33
Acute myeloid leukemia	17	0.12	0.34
Cysteine and methionine metabolism	51	0.13	0.38
Galactose metabolism	17	0.16	0.45
Non-small cell lung cancer	20	0.17	0.46
Lysine biosynthesis	16	0.19	0.52
Prostate cancer	20	0.20	0.52
Pentose and glucuronate interconversions	7	0.24	0.59
Ascorbate and aldarate metabolism	7	0.24	0.59
Valine, leucine and isoleucine degradation	79	0.25	0.59
N-Glycan biosynthesis	59	0.24	0.59
Apoptosis	14	0.23	0.59
Fc epsilon RI signaling pathway	20	0.24	0.59
Aldosterone-regulated sodium reabsorption	34	0.24	0.59
Chronic myeloid leukemia	15	0.23	0.59
Protein processing in endoplasmic reticulum	1	0.26	0.62
Huntington's disease	18	0.26	0.62
ErbB signaling pathway	16	0.28	0.64
Glyoxylate and dicarboxylate metabolism	3	0.29	0.66
Proteoglycans in cancer	4	0.29	0.66
Non-alcoholic fatty liver disease (NAFLD)	21	0.30	0.67
Carbohydrate digestion and absorption	49	0.31	0.69
Bacterial invasion of epithelial cells	14	0.31	0.69
NF-kappa B signaling pathway	1	0.34	0.73
Nicotine addiction	1	0.34	0.73
Alanine, aspartate and glutamate metabolism	51	0.35	0.74
Arginine and proline metabolism	50	0.36	0.74
B cell receptor signaling pathway	15	0.36	0.74
Steroid biosynthesis	4	0.37	0.74
beta-Alanine metabolism	26	0.37	0.74
Toll-like receptor signaling pathway	10	0.38	0.74
Prolactin signaling pathway	15	0.37	0.74
Influenza A	3	0.38	0.74
Pancreatic cancer	18	0.37	0.74

Table A1.10 Over-represented pathways in cluster 4.

Pathway name	Number of occurrences	p-value	Adjusted p-value
Inositol phosphate metabolism	90	0.39	0.76
Synthesis and degradation of ketone bodies	3	0.40	0.76
Amphetamine addiction	5	0.40	0.76
Retinol metabolism	23	0.44	0.83
AMPK signaling pathway	37	0.48	0.90
Biosynthesis of unsaturated fatty acids	17	0.50	0.91
Endocrine and other factor-regulated calcium reabsorption	37	0.51	0.93
Arachidonic acid metabolism	26	0.52	0.93
Thiamine metabolism	5	0.53	0.93
Fc gamma R-mediated phagocytosis	23	0.52	0.93
Folate biosynthesis	2	0.54	0.95
Lysine degradation	43	0.55	0.96
Vasopressin-regulated water reabsorption	5	0.57	0.98
Dopaminergic synapse	4	0.57	0.98
Glutamatergic synapse	80	0.59	1.00
Pentose phosphate pathway	6	1.00	1.00
Fructose and mannose metabolism	18	0.86	1.00
Fatty acid biosynthesis	8	0.89	1.00
Purine metabolism	89	0.99	1.00
Caffeine metabolism	1	1.00	1.00
Pyrimidine metabolism	45	0.99	1.00
Selenocompound metabolism	18	0.63	1.00
Mucin type O-Glycan biosynthesis	56	0.89	1.00
Other types of O-glycan biosynthesis	7	1.00	1.00
Amino sugar and nucleotide sugar metabolism	46	0.63	1.00
Butirosin and neomycin biosynthesis	2	0.99	1.00
Glycosaminoglycan degradation	10	0.80	1.00
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	31	0.76	1.00
Glycosaminoglycan biosynthesis - keratan sulfate	11	1.00	1.00
Glycosaminoglycan biosynthesis - heparan sulfate / heparin	64	0.77	1.00
Glycerolipid metabolism	73	0.92	1.00
Glycerophospholipid metabolism	84	0.94	1.00
Linoleic acid metabolism	5	0.90	1.00
alpha-Linolenic acid metabolism	6	1.00	1.00
Sphingolipid metabolism	36	0.86	1.00
Glycosphingolipid biosynthesis - lacto and neolacto series	57	0.72	1.00
Glycosphingolipid biosynthesis - globo series	45	0.98	1.00
Butanoate metabolism	6	1.00	1.00
One carbon pool by folate	14	0.99	1.00
Nicotinate and nicotinamide metabolism	23	0.75	1.00
Biotin metabolism	1	1.00	1.00
Drug metabolism - other enzymes	21	0.90	1.00
Ras signaling pathway	10	0.64	1.00
Rap1 signaling pathway	29	0.95	1.00
Calcium signaling pathway	31	1.00	1.00
cGMP-PKG signaling pathway	76	0.96	1.00
cAMP signaling pathway	66	0.99	1.00
Chemokine signaling pathway	28	0.96	1.00

Table A1.10 Over-represented pathways in cluster 4.

Pathway name	Number of occurrences	p-value	Adjusted p-value
Phosphatidylinositol signaling system	90	0.94	1.00
Oocyte meiosis	6	1.00	1.00
Lysosome	10	0.92	1.00
Cardiac muscle contraction	13	0.95	1.00
Adrenergic signaling in cardiomyocytes	52	0.99	1.00
Vascular smooth muscle contraction	36	0.99	1.00
Osteoclast differentiation	3	0.91	1.00
Gap junction	31	1.00	1.00
Platelet activation	46	0.98	1.00
Jak-STAT signaling pathway	3	0.68	1.00
Natural killer cell mediated cytotoxicity	3	0.95	1.00
T cell receptor signaling pathway	10	0.61	1.00
Leukocyte transendothelial migration	6	0.83	1.00
Circadian entrainment	35	1.00	1.00
Long-term potentiation	8	0.99	1.00
Neurotrophin signaling pathway	6	0.82	1.00
Retrograde endocannabinoid signaling	33	0.97	1.00
Cholinergic synapse	40	0.97	1.00
GABAergic synapse	62	0.88	1.00
Long-term depression	23	0.94	1.00
Phototransduction	1	0.69	1.00
Inflammatory mediator regulation of TRP channels	37	1.00	1.00
Insulin secretion	32	1.00	1.00
GnRH signaling pathway	17	1.00	1.00
Ovarian steroidogenesis	45	0.80	1.00
Progesterone-mediated oocyte maturation	41	0.97	1.00
Estrogen signaling pathway	34	1.00	1.00
Melanogenesis	17	1.00	1.00
Thyroid hormone synthesis	41	1.00	1.00
Thyroid hormone signaling pathway	36	0.87	1.00
Oxytocin signaling pathway	48	0.85	1.00
Type II diabetes mellitus	20	0.60	1.00
Salivary secretion	63	1.00	1.00
Gastric acid secretion	42	1.00	1.00
Pancreatic secretion	69	0.95	1.00
Bile secretion	65	0.98	1.00
Mineral absorption	26	0.99	1.00
Alzheimer's disease	21	0.73	1.00
Amyotrophic lateral sclerosis (ALS)	15	1.00	1.00
Morphine addiction	77	0.84	1.00
Chagas disease (American trypanosomiasis)	22	0.93	1.00
Toxoplasmosis	6	0.71	1.00
Amoebiasis	20	0.96	1.00
Hepatitis C	3	0.91	1.00
Measles	3	0.91	1.00
HTLV-I infection	26	0.68	1.00
Epstein-Barr virus infection	11	0.69	1.00
Dilated cardiomyopathy	13	1.00	1.00

Table A1.10 Over-represented pathways in cluster 5.

Pathway name	Number of occurrences	p-value	Adjusted p-value
Glycolysis / Gluconeogenesis	1	0.35	0.69
Fructose and mannose metabolism	1	0.24	0.69
Galactose metabolism	1	0.15	0.69
Fatty acid degradation	1	0.50	0.69
Oxidative phosphorylation	1	0.28	0.69
Purine metabolism	1	0.75	0.75
Pyrimidine metabolism	1	0.52	0.69
Alanine, aspartate and glutamate metabolism	1	0.44	0.69
Cysteine and methionine metabolism	1	0.41	0.69
Valine, leucine and isoleucine degradation	1	0.59	0.69
Valine, leucine and isoleucine biosynthesis	1	0.34	0.69
Lysine degradation	1	0.41	0.69
Arginine and proline metabolism	1	0.44	0.69
beta-Alanine metabolism	1	0.25	0.69
Starch and sucrose metabolism	1	0.22	0.69
Mucin type O-Glycan biosynthesis	1	0.55	0.69
Amino sugar and nucleotide sugar metabolism	1	0.44	0.69
Butirosin and neomycin biosynthesis	1	0.08	0.69
Glycosaminoglycan degradation	1	0.14	0.69
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	1	0.34	0.69
Glycosaminoglycan biosynthesis - heparan sulfate / heparin	1	0.57	0.69
Glycerolipid metabolism	1	0.65	0.69
Glycerophospholipid metabolism	1	0.70	0.71
Ether lipid metabolism	1	0.40	0.69
Arachidonic acid metabolism	1	0.27	0.69
Sphingolipid metabolism	1	0.40	0.69
Glycosphingolipid biosynthesis - lacto and neolacto series	1	0.52	0.69
Glycosphingolipid biosynthesis - globo series	1	0.52	0.69
Glycosphingolipid biosynthesis - ganglio series	1	0.31	0.69
One carbon pool by folate	1	0.26	0.69
Nicotinate and nicotinamide metabolism	1	0.27	0.69
Pantothenate and CoA biosynthesis	1	0.40	0.69
Retinol metabolism	1	0.23	0.69
Drug metabolism - other enzymes	1	0.28	0.69
cGMP-PKG signaling pathway	1	0.67	0.69
cAMP signaling pathway	1	0.65	0.69
HIF-1 signaling pathway	1	0.28	0.69
Lysosome	1	0.16	0.69
Adrenergic signaling in cardiomyocytes	1	0.58	0.69
Vascular smooth muscle contraction	1	0.46	0.69
VEGF signaling pathway	1	0.23	0.69
Gap junction	1	0.46	0.69
Platelet activation	1	0.52	0.69
Circadian entrainment	1	0.53	0.69
Glutamatergic synapse	1	0.63	0.69
GABAergic synapse	1	0.58	0.69

Table A1.11 Over-represented pathways in cluster 5.

Pathway name	Number of occurrences	p-value	Adjusted p-value
Long-term depression	1	0.31	0.69
Ovarian steroidogenesis	1	0.46	0.69
Thyroid hormone synthesis	1	0.57	0.69
Thyroid hormone signaling pathway	1	0.40	0.69
Oxytocin signaling pathway	1	0.49	0.69
Type II diabetes mellitus	1	0.22	0.69
Aldosterone-regulated sodium reabsorption	1	0.30	0.69
Proximal tubule bicarbonate reclamation	1	0.28	0.69
Salivary secretion	1	0.67	0.69
Gastric acid secretion	1	0.56	0.69
Pancreatic secretion	1	0.63	0.69
Carbohydrate digestion and absorption	1	0.42	0.69
Fat digestion and absorption	1	0.22	0.69
Bile secretion	1	0.64	0.69
Mineral absorption	1	0.39	0.69
Amyotrophic lateral sclerosis (ALS)	1	0.33	0.69
Morphine addiction	1	0.65	0.69
Vibrio cholerae infection	1	0.28	0.69
Pancreatic cancer	1	0.18	0.69

Table A1.11 Over-represented pathways in cluster 6.

Pathway name	Number of occurrences	p-value	Adjusted p-value
Type II diabetes mellitus	161	1.56E-16	3.22E-14
Butanoate metabolism	144	6.41E-15	6.60E-13
Insulin signaling pathway	145	2.91E-12	2.00E-10
Regulation of actin cytoskeleton	120	5.74E-12	2.96E-10
Prolactin signaling pathway	105	1.93E-11	7.94E-10
Bacterial invasion of epithelial cells	94	1.37E-10	4.71E-09
Chronic myeloid leukemia	94	1.98E-10	5.10E-09
Acute myeloid leukemia	96	1.89E-10	5.10E-09
Colorectal cancer	95	2.53E-10	5.80E-09
B cell receptor signaling pathway	100	5.59E-10	1.07E-08
Vitamin digestion and absorption	109	5.70E-10	1.07E-08
Renal cell carcinoma	105	1.37E-08	2.34E-07
Glycine, serine and threonine metabolism	86	4.12E-08	6.53E-07
Non-small cell lung cancer	106	1.16E-07	1.70E-06
Tyrosine metabolism	82	1.98E-07	2.72E-06
Apoptosis	80	2.20E-07	2.84E-06
Vitamin B6 metabolism	86	3.23E-07	3.91E-06
AMPK signaling pathway	205	1.02E-06	1.16E-05
Pancreatic cancer	105	1.13E-06	1.23E-05
Glycosaminoglycan biosynthesis - keratan sulfate	145	1.28E-06	1.32E-05
Proteoglycans in cancer	28	1.44E-06	1.42E-05
Fc epsilon RI signaling pathway	107	1.53E-06	1.43E-05
Epstein-Barr virus infection	82	3.96E-06	3.54E-05
Lysine biosynthesis	83	4.44E-06	3.81E-05
Caffeine metabolism	56	4.63E-06	3.82E-05
beta-Alanine metabolism	140	7.92E-06	6.27E-05
Aldosterone-regulated sodium reabsorption	168	8.64E-06	6.59E-05
Fatty acid biosynthesis	75	1.84E-05	0.00013541
Propanoate metabolism	105	2.48E-05	0.000175891
Pentose phosphate pathway	100	2.91E-05	0.000199533
Nicotine addiction	8	3.22E-05	0.000213921
Terpenoid backbone biosynthesis	17	3.50E-05	0.000225309
Tryptophan metabolism	137	4.03E-05	0.000246234
ErbB signaling pathway	84	4.18E-05	0.000246234
Viral carcinogenesis	19	4.10E-05	0.000246234
Phenylalanine, tyrosine and tryptophan biosynthesis	47	5.21E-05	0.00029824
Sphingolipid metabolism	220	8.43E-05	0.000469127
Non-alcoholic fatty liver disease (NAFLD)	107	8.69E-05	0.000471063
HIF-1 signaling pathway	148	9.99E-05	0.000527423
Influenza A	20	0.000318179	0.001638624
Ras signaling pathway	66	0.000480099	0.002412204
Jak-STAT signaling pathway	27	0.00056637	0.002777909
Toll-like receptor signaling pathway	55	0.000656828	0.003146664
Phenylalanine metabolism	36	0.001131254	0.00517863
HTLV-I infection	146	0.001115401	0.00517863
Histidine metabolism	73	0.001218231	0.005455557
VEGF signaling pathway	116	0.001260968	0.005526796
Mineral absorption	203	0.001347207	0.005781764
Fructose and mannose metabolism	119	0.001505351	0.006328617
Endometrial cancer	110	0.002226267	0.009172221
Carbohydrate digestion and absorption	222	0.002449239	0.009893004
Lysosome	81	0.002776239	0.010998178
mTOR signaling pathway	107	0.00341812	0.013285525

Table A1.12 Over-represented pathways in cluster 6.

Pathway name	Number of occurrences	p-value	Adjusted p-value
Butirosin and neomycin biosynthesis	44	0.003524148	0.013443973
Ascorbate and aldarate metabolism	33	0.003736629	0.013995374
Melanoma	103	0.003841056	0.014129599
Phototransduction	11	0.004646859	0.016793913
TNF signaling pathway	57	0.004897168	0.017393389
Lysine degradation	210	0.005475571	0.019118095
Ubiquinone and other terpenoid-quinone biosynthesis	7	0.00628012	0.021489529
Chemokine signaling pathway	185	0.006363404	0.021489529
ABC transporters	58	0.007543879	0.025065147
Galactose metabolism	70	0.009571577	0.030882798
Drug metabolism - other enzymes	134	0.009594656	0.030882798
T cell receptor signaling pathway	58	0.010157899	0.032192726
Folate biosynthesis	14	0.010814755	0.033755144
Fc gamma R-mediated phagocytosis	115	0.011164135	0.03432555
Prostate cancer	84	0.011900869	0.036052633
Glioma	116	0.012907729	0.038536118
Focal adhesion	32	0.01606941	0.047289978
Thyroid hormone signaling pathway	201	0.017511931	0.050809265
Natural killer cell mediated cytotoxicity	36	0.019701765	0.05636894
Fat digestion and absorption	104	0.020457311	0.057728849
Nicotinate and nicotinamide metabolism	125	0.025886533	0.07206251
Hepatitis C	31	0.026728798	0.072449111
Measles	31	0.026728798	0.072449111
Linoleic acid metabolism	43	0.02901022	0.07713219
Osteoclast differentiation	31	0.029205392	0.07713219
Toxoplasmosis	39	0.030264107	0.078916533
Glycosphingolipid biosynthesis - ganglio series	144	0.036050981	0.092831277
Phagosome	70	0.040111535	0.102012052
Leukocyte transendothelial migration	42	0.062715543	0.15755368
Porphyrin and chlorophyll metabolism	36	0.070215487	0.174269763
Amoebiasis	131	0.071958416	0.17646945
Glycolysis / Gluconeogenesis	159	0.076619093	0.185688626
alpha-Linolenic acid metabolism	80	0.084589609	0.201821539
Rap1 signaling pathway	173	0.08523531	0.201821539
FoxO signaling pathway	86	0.089434414	0.209357834
Parkinson's disease	78	0.091759038	0.212386087
Arginine and proline metabolism	210	0.097974921	0.224253708
Hepatitis B	29	0.10057449	0.227646204
Small cell lung cancer	89	0.101667237	0.227646204
Neurotrophin signaling pathway	40	0.102786542	0.227677716
Proximal tubule bicarbonate reclamation	123	0.125319565	0.274636493
Pyruvate metabolism	96	0.13476915	0.292236262
Tuberculosis	1	0.15314939	0.328633066
PI3K-Akt signaling pathway	16	0.156250325	0.331830588
Alanine, aspartate and glutamate metabolism	208	0.191666093	0.398820355
Chagas disease (American trypanosomiasis)	129	0.191386218	0.398820355
Sulfur metabolism	5	0.209649621	0.431878218
Glycosaminoglycan degradation	56	0.236120488	0.481592283
African trypanosomiasis	6	0.250705321	0.506326433
Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	149	0.258902437	0.517804875
Alzheimer's disease	102	0.274502233	0.543725578
Sulfur relay system	2	0.292228895	0.573325261
Mucin type O-Glycan biosynthesis	274	0.324411645	0.630460367

Table A1.12 Over-represented pathways in cluster 6.

Pathway name	Number of occurrences	p-value	Adjusted p-value
Selenocompound metabolism	82	0.333733786	0.642515514
Biotin metabolism	29	0.340716218	0.649884637
D-Arginine and D-ornithine metabolism	1	0.392678335	0.742126028
Metabolism of xenobiotics by cytochrome P450	30	0.397206094	0.743858686
Other types of O-glycan biosynthesis	84	0.405717482	0.752953165
Taste transduction	3	0.409516052	0.753217024
Collecting duct acid secretion	74	0.424899616	0.774595759
Amyotrophic lateral sclerosis (ALS)	136	0.451319791	0.81554278
Synthesis and degradation of ketone bodies	10	0.461165197	0.818965781
Cysteine and methionine metabolism	179	0.458705847	0.818965781
Biosynthesis of unsaturated fatty acids	69	0.46759626	0.823289141
D-Glutamine and D-glutamate metabolism	64	0.471760733	0.823582297
Leishmaniasis	1	0.48569122	0.840776398
Huntington's disease	62	0.522624161	0.897171477
Starch and sucrose metabolism	84	0.529215019	0.900977636
Drug metabolism - cytochrome P450	29	0.551000808	0.930378414
Adipocytokine signaling pathway	77	0.580407052	0.972063843
Citrate cycle (TCA cycle)	21	0.989016848	1
Pentose and glucuronate interconversions	19	0.691239621	1
Fatty acid elongation	1	0.735491808	1
Fatty acid degradation	205	0.978868298	1
Steroid biosynthesis	10	0.825068593	1
Primary bile acid biosynthesis	15	0.8082639	1
Steroid hormone biosynthesis	59	0.973214863	1
Oxidative phosphorylation	98	0.937508113	1
Purine metabolism	320	1	1
Pyrimidine metabolism	244	0.654217573	1
Valine, leucine and isoleucine degradation	235	0.999983818	1
Valine, leucine and isoleucine biosynthesis	88	0.999999774	1
Taurine and hypotaurine metabolism	6	0.78166927	1
Cyanoamino acid metabolism	1	0.863966596	1
Glutathione metabolism	15	0.935263098	1
N-Glycan biosynthesis	212	0.70812518	1
Other glycan degradation	3	0.824627683	1
Amino sugar and nucleotide sugar metabolism	187	0.759169198	1
Glycosaminoglycan biosynthesis - heparan sulfate / heparin	268	0.872180286	1
Glycerolipid metabolism	301	0.997575256	1
Inositol phosphate metabolism	304	0.99914439	1
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	84	0.977458871	1
Glycerophospholipid metabolism	315	0.999999402	1
Ether lipid metabolism	144	0.98637289	1
Arachidonic acid metabolism	100	0.751472878	1
Glycosphingolipid biosynthesis - lacto and neolacto series	181	0.999999646	1
Glycosphingolipid biosynthesis - globo series	189	0.999976096	1
Glyoxylate and dicarboxylate metabolism	2	0.997535887	1
One carbon pool by folate	91	0.834253522	1
Thiamine metabolism	18	0.689700478	1
Riboflavin metabolism	9	0.698728878	1
Pantothenate and CoA biosynthesis	140	0.997837279	1
Retinol metabolism	76	0.955025159	1
Nitrogen metabolism	43	0.643988329	1
Protein export	3	0.976598686	1

Table A1.12 Over-represented pathways in cluster 6.

Pathway name	Number of occurrences	p-value	Adjusted p-value
PPAR signaling pathway	164	0.738301332	1
Calcium signaling pathway	244	0.75496893	1
cGMP-PKG signaling pathway	292	0.999998931	1
cAMP signaling pathway	284	0.999944176	1
Phosphatidylinositol signaling system	319	1	1
Oocyte meiosis	73	0.952265731	1
Peroxisome	112	0.986158822	1
Cardiac muscle contraction	67	0.930870018	1
Adrenergic signaling in cardiomyocytes	280	0.768262888	1
Vascular smooth muscle contraction	143	0.999999955	1
Gap junction	134	0.999999999	1
Platelet activation	225	0.932542962	1
Circadian entrainment	182	0.999999745	1
Long-term potentiation	42	0.99965287	1
Synaptic vesicle cycle	78	0.992034958	1
Retrograde endocannabinoid signaling	152	0.99381721	1
Glutamatergic synapse	247	0.999999973	1
Cholinergic synapse	207	0.760697635	1
Serotonergic synapse	21	0.999989187	1
GABAergic synapse	234	0.999929233	1
Dopaminergic synapse	10	0.974207111	1
Long-term depression	85	0.999982444	1
Inflammatory mediator regulation of TRP channels	202	0.935563957	1
Insulin secretion	188	0.999789044	1
GnRH signaling pathway	118	0.999992243	1
Ovarian steroidogenesis	170	0.997814437	1
Progesterone-mediated oocyte maturation	196	0.966332987	1
Estrogen signaling pathway	199	0.929106075	1
Melanogenesis	107	0.999837625	1
Thyroid hormone synthesis	229	0.999857341	1
Oxytocin signaling pathway	209	0.881350998	1
Endocrine and other factor-regulated calcium reabsorption	139	0.858187332	1
Vasopressin-regulated water reabsorption	14	0.959785337	1
Salivary secretion	275	0.999999994	1
Gastric acid secretion	226	0.999768006	1
Pancreatic secretion	251	0.999999972	1
Protein digestion and absorption	97	0.883510573	1
Bile secretion	269	0.999993071	1
Cocaine addiction	13	0.967899923	1
Amphetamine addiction	10	0.981830901	1
Morphine addiction	246	1	1
Vibrio cholerae infection	87	0.99435115	1
Epithelial cell signaling in Helicobacter pylori infection	56	0.877884507	1
Pathways in cancer	18	0.812574211	1
Chemical carcinogenesis	31	0.921443354	1
MicroRNAs in cancer	14	0.997061505	1
Rheumatoid arthritis	33	0.782752456	1
Dilated cardiomyopathy	103	0.990994669	1

Appendix A2

BIOINFORMATICS ORIGINAL PAPER

2014, pages 1–7
doi:10.1093/bioinformatics/btu589

Data and text mining

Advance Access publication September 2, 2014

Over-representation of correlation analysis (ORCA): a method for identifying associations between variable sets

Yotsawat Pomyen[1,2], Marcelo Segura[1], Timothy M. D. Ebbels[1] and Hector C. Keun[1,*]

[1]Department of Surgery and Cancer, Section of Computational and Systems Medicine, Imperial College London, Exhibition Road, London SW7 2AZ, UK and [2]Translational Research Unit, Chulabhorn Research Institute, Bangkok 10210, Thailand

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Often during the analysis of biological data, it is of importance to interpret the correlation structure that exists between variables. Such correlations may reveal patterns of co-regulation that are indicative of biochemical pathways or common mechanisms of response to a related set of treatments. However, analyses of correlations are usually conducted by either subjective interpretation of the univariate covariance matrix or by applying multivariate modeling techniques, which do not take prior biological knowledge into account. Over-representation analysis (ORA) is a simple method for objectively deciding whether a set of variables of known or suspected biological relevance, such as a gene set or pathway, is more prevalent in a set of variables of interest than we expect by chance. However, ORA is usually applied to a set of variables differentiating a single experimental variable and does not take into account correlations.

Results: Over-representation of correlation analysis (ORCA) is a novel combination of ORA and correlation analysis that provides ameans to test whether more associations exist between two specific groups of variables than expected by chance. The method is exemplified by application to drug sensitivity and microRNA expression data from a panel of cancer cell lines (NCI60). ORCA highlighted a previously reported correlation between sensitivity to alkylating anticancer agents and topoisomerase inhibitors. We also used this approach to validate microRNA clusters predicted by mRNA correlations. These observations suggest that ORCA has the potential to reveal novel insights from these data, which are not readily apparent using classical ORA.

Availability and implementation: The R code of the method is available at https://github.com/ORCABioinfo/ORCAcode

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 10, 2014; revised on August 13, 2014; accepted on August 26, 2014

1 INTRODUCTION

High-throughput biological techniques such as gene expression microarray or high-throughput sequencing have become a routine for most research laboratories in life sciences. Thousands of biomolecules in different conditions, e.g. mRNAs, microRNAs, proteins or metabolites, can be measured simultaneously. In complex diseases, such as cancer, these biomolecules in a group (such as a pathway or a gene set) are often altered together. Two conventional, but contrasting, approaches for analyzing coordinated responses in molecular profile data are correlation analysis and pathway analysis. In correlation or covariance analysis (and its multivariate extensions such as principal components analysis and canonical correlation analysis), there is a focus on quantitatively estimating the explanatory power of one variable over another to infer some fundamental link between the observed biomolecules. In pathway analysis, however, the objective is to use prior knowledge about interacting or otherwise related sets of biomolecules and to test specific hypotheses about the relationship between a particular set of those biomolecules (the 'pathway') and a given experimental condition. A conventional approach for pathway analysis of high-throughput biological data is over-representation analysis (ORA). To perform ORA, firstly, the biomolecules, such as mRNA, proteins or microRNA, considered 'differentially expressed' in two or more conditions are identified. Secondly, the number of differentially expressed biomolecules in each pathway is determined. Finally, for each pathway, a probability value (*P*-value) of obtaining the number of differentially expressed biomolecules against the background list of all biomolecules measured is calculated using a hypergeometric distribution. The first implementation of ORA was by Tavazoie *et al.* (1999) where transcriptional regulatory subnetworks in yeast were identified by using mRNA microarray profiling. Most implementations of ORA are used to perform pathway analysis based on pathway information, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) or other pathway databases (Cavill *et al.*, 2011), or to infer important biological functions of two different conditions from biological categories, such as Gene Ontology (GO; Beissbarth and Speed, 2004; Zeeberg *et al.*, 2003).

We introduce a novel method called Over-Representation of Correlation Analysis (ORCA) that seeks to combine both conventional approaches for analysis of coordinated biological responses and provide a new means to test for an unexpectedly high prevalence of informative associations between two sets of variables, which could be mRNA, microRNA or proteins, that comprise pathways/groups deemed of importance *a priori.* The method first calculates the correlation coefficients between all the variables in the dataset, in which the variables can be divided into groups according to pre-defined criteria (exclusive groupings). Then, as an analogy to counting 'differentially' expressed

*To whom correspondence should be addressed.

Appendix A2

genes in ORA (which might be below or above a certain threshold of *P*-values or other metrics), ORCA defines the number of correlation coefficients that are ‘above’ a certain threshold within each group. Finally, the probability of association between any two-group pair is calculated in a similar fashion to the calculation of over-representation in the conventional ORA. This probability value is calculated from the number of correlation coefficients that pass a threshold against the background number of correlation coefficients of the two-group pairs and overall correlation coefficients by using the hypergeometric test. The correlation coefficient threshold can be empirically chosen or calculated by a Shannon’s entropy-based threshold selection. The method was applied to several biological datasets to demonstrate the concept and implications of ORCA in biological data analysis.

2 METHODS

To apply ORCA, the variables (such as genes, microRNA or proteins) of the dataset must be divided into sets according to relevant criteria, such as GO annotations, KEGG pathways or groups, resulting from unsupervised classification techniques such as hierarchical clustering. There should also be as many data points available for an accurate correlation coefficient as possible between two variables.

All statistical calculations and plots were made using R statistical packages version 2.12. The *P*-values were all adjusted by Benjamini–Hochberg false discovery rate (FDR) procedure (Benjamini and Hochberg, 2009).

2.1 Over-representation of correlation analysis

A schematic representation of ORCA is shown in Figure 1. Firstly, the method starts with calculation of the correlation coefficients between all *n* variables, yielding a symmetric correlation matrix ($n \times n$ matrix). We chose Spearman’s rank correlation coefficients in this study to avoid the assumption of a Normal distribution. Secondly, variables are sorted into sets or groups according to *a priori* information, such as pathways, GO (sets 1–6 in Fig. 1). Correlation coefficients between variables in the same group are called within-group correlations, whereas the ones between different groups are called between-group correlations. Thirdly, all correlation coefficients are labelled as above or below a certain threshold previously established by empirical means or a threshold selection method (see Section 2.2). The association between sets or groups is then quantified by the number of correlation coefficients between those two groups that are above the threshold.

To determine the significance of the association between groups, we introduce ORCA. The total number of correlation coefficients in the correlation matrix of the dataset (N) in the analysis can be categorized in two ways:

(i) correlation coefficients that are above the threshold (X of Venn diagram in Fig. 1)
(ii) correlation coefficients that link members of a particular pair of groups (M of Venn diagram in Fig. 1)

For any association between two groups, these criteria define four different categories of correlation coefficients (see contingency table in Fig. 1). The number of correlation coefficients that are both above the threshold and are members of a particular pair of groups (*k* in contingency table in Fig. 1) is the determinant of association between the pair of groups being calculated (we used association between sets 1 and 6 as an example in Fig. 1).

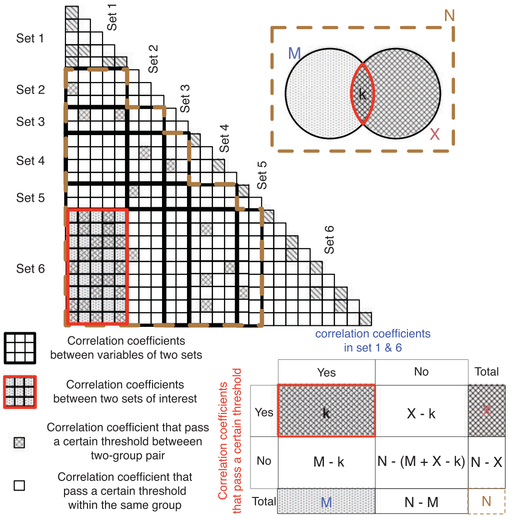

Set 1
Set 2
Set 3
Set 4
Set 5
Set 6
N
M
k
X
Correlation coefficients between variables of two sets
Correlation coefficients between two sets of interest
Correlation coefficient that pass a certain threshold between two-group pair
Correlation coefficient that pass a certain threshold within the same group
correlation coefficients in set 1 & 6
Correlation coefficients that pass a certain threshold
Yes
No
Total
k
X - k
X
M - k
N - (M + X - k)
N - X
M
N - M
N

Fig. 1. The concept of over-representation of correlation. A correlation matrix of correlation coefficients calculated from a dataset is divided into sets or groups. Correlation coefficients are represented by rectangles in each group pair. Non-filled and crosshatch-filled rectangles represent correlation coefficients that do not pass and do pass a certain threshold, respectively. The contingency table in the bottom right explains the variables required to calculate the probability value of having a certain number of correlation coefficients that pass the threshold in each group pair by chance alone via the hypergeometric distribution (Equation 1)

Finally, we calculate the *P*-value of obtaining, by chance, the number of correlation coefficients that pass the threshold and are present in each set of between-group correlations (Supplementary Fig. S1). This can be calculated by the hypergeometric distribution:

$$p(n>k)=1-\sum_{i=0}^{k}\frac{\binom{M}{i}\binom{N-M}{X-i}}{\binom{N}{X}} \qquad (1)$$

where $\binom{t}{u}=\frac{t!}{u!(t-u)!}$ is the binomial coefficient, *M* is the number of between-group correlation coefficients, *N* is the total number of correlation coefficients in the correlation matrix (excluding within-group correlations), *X* is the total number of between-group correlation coefficients that pass the threshold and *k* is the number of correlation coefficients in the between-group set of interest that pass the threshold.

The reason for excluding within-group correlation coefficients that pass the threshold is that these tend to have high correlation coefficients and will cause an underestimation of the *P*-value calculating from between-group correlation. The *P*-values for within-group correlations were separately calculated using the same equation but including within-group correlations in N, M and X.

2.2 Threshold selection

To select a threshold of correlation coefficient that yields the most information from the data, a selection method based on Shannon’s entropy (Shannon, 1948) was developed. Briefly, a score is calculated for a range

Appendix A2

of correlation coefficient thresholds based on the P-values from ORCA for all group pairs using the equation (2):

$$H(X) = -\sum_{i=1}^{J} p(x_i)\log_{e} p(x_i) \qquad (2)$$

where $H(X)$ is a Shannon entropy-like score, $P(x_i)$ is a P-value of a within- or between-group association calculated by hypergeometric test, J is the total number of within- and between-group pairs [if the total number of groups is n then $J = n(n-1)/2 + n$] and $\log_e$ is the natural logarithm. The correlation coefficient that yields the highest Shannon entropy-like score is selected for the ORCA threshold.

2.3 Permutation analysis

As the variables in the correlation matrix are not necessarily independent, permutation analysis was used to determine the null distribution of the P-values obtained by the hypergeometric test for a given data matrix. Specifically the group membership of each variable in the dataset was permuted, but the group structure (i.e. the number of groups and the number of variables within each group) together with the overall distribution of correlations was retained. For each dataset, one million permutations were performed, and the empirical P-values of each between and within groups were calculated by using the following equation (Davison and Hinkley, 1997):

$$P = \frac{(r+1)}{(n+1)} \qquad (3)$$

where P is the empirical P-value, r is the number of times that the hypergeometric P-values from permutation test are equal or less than the actual P-values from ORCA and n is the total number of permutation used in the test.

3 RESULTS

3.1 ORCA reveals an association between the sensitivity of tumour cells to alkylating agents and topoisomerase inhibitors

To demonstrate the concept and potential of ORCA, we examined publicly available phenotypic profiles from the National Cancer Institute cell line panel (NCI-60 panel). This dataset meets the basic requirements of ORCA: the variables can be divided into groups and each variable has multiple data points for correlation analysis. This dataset consists of drug sensitivity data associated with the NCI-60 cell panel: profiles of 58 cancer cell lines treated with a range of drug compounds (Scherf *et al.*, 2000). The effect of a drug on a cell line was represented by the concentration of the drug that lead to 50% growth inhibition (GI_{50}). We selected a collection of well-validated results for 116 chemotherapeutic drugs and divided them into nine groups according to their mechanisms of action, i.e. the molecular targets of the drugs. Our objective was to determine whether there was similarity between different groups of drugs in terms of the sensitivity pattern across the cell lines. The dataset is detailed in Supplementary data files S1 and S2 and Supplementary Figure S2.

Despite a wide range of different mechanisms of action, some of the drugs from different classes have an apparently high degree of similarity in the observed sensitivity profile, as observed in the correlation matrix of the GI_{50} data (Fig. 2). We used ORCA to determine whether any two drug classes, defined by independent modes of action, have more common correlations that are above an informative threshold than we would expect between them by chance. For this dataset, our threshold selection method identified a correlation coefficient of 0.79 to produce the highest Shannon entropy-like score, i.e. to be the most informative (Supplementary Fig. S3). Using this threshold, ORCA was then applied to generate a matrix of P-values for obtaining the observed number of high correlations, between every possible pair of groups of drugs (Table 1). The analysis revealed that the similarity in sensitivity profile across the NCI-60 panel between alkylating agents (group 1) and topoisomerase inhibitors (group 8) was much higher than expected by chance (q = 1e-6, Benjamini-Hochberg FDR). This finding has not previously been reported and could result in part by the fact that both sets of compounds are likely to lead to single and double-stranded DNA breaks in rapidly dividing cells, which leaves the cells reliant on a common set of DNA repair pathways for survival (Bargonetti *et al.*, 2010; Rudolf *et al.*, 2011). Another pair of drug classes exhibiting higher than expected association was DNA synthesis inhibitor (group 5) and ribonucleotide reductase inhibitor (group 6), albeit at a lesser extent than alkylating agents and topoisomerase inhibitors.

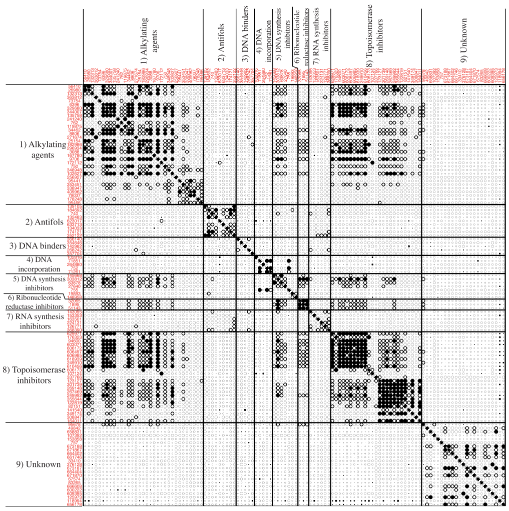


Fig. 2. Correlation matrix of drug sensitivity data from the NCI-60 cell panel. Circles and squares represent positive and negative correlation coefficients between two drugs, respectively. The sizes of circles and squares reflect the strength of correlation. The filled circles and squares are correlation coefficients that pass the threshold of 0.79. The lines divide drugs into their groups according to their mechanisms of actions. The number labels are NSC numbers (NCI's sample accession number) of the drugs. The drug names are given in Supplementary Material. Note the unusually high number of positive correlations between alkylating agents (group 1) and topoisomerase inhibitors (group 8)

3.2 Verification of microRNA cluster significance

The second and third datasets used in this study are microRNA expression profiles of NCI-60 cell panel from Liu *et al.* (2010)

Appendix A2

Table 1. (A) Benjamini–Hochberg FDR-adjusted (lower half) and empirical (upper half) *P*-value table for drug sensitivity dataset; (B) *P*-values from diagonal (within-group correlation coefficients)

A)

Drug Group	1	2	3	4	5	6	7	8	9	Drug Group
1		0.94	0.82	0.82	0.29	0.70	0.87	1E-6	0.99	**1**
2	1.00		0.59	0.59	0.66	0.58	0.60	0.92	0.91	**2**
3	1.00	1.00		0.58	0.68	0.42	0.64	0.78	0.79	**3**
4	1.00	1.00	1.00		0.11	0.42	0.64	0.78	0.79	**4**
5	1.00	1.00	1.00	0.33		0.01	0.62	0.24	0.86	**5**
6	1.00	1.00	1.00	1.00	0.01		0.47	0.64	0.62	**6**
7	1.00	1.00	1.00	1.00	1.00	1.00		0.81	0.81	**7**
8	1E-57	1.00	1.00	1.00	1.00	1.00	1.00		0.99	**8**
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		**9**
Drug Group	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	Drug Group

B)

Drug Group	1	2	3	4	5	6	7	8	9
Empirical *P*-value	5E-5	0.03	0.31	0.007	0.21	1E-6	0.11	1E-6	0.01
FDR-adjusted *P*-value	9E-21	0.01	0.41	0.001	0.31	0	0.56	2E-37	0.26

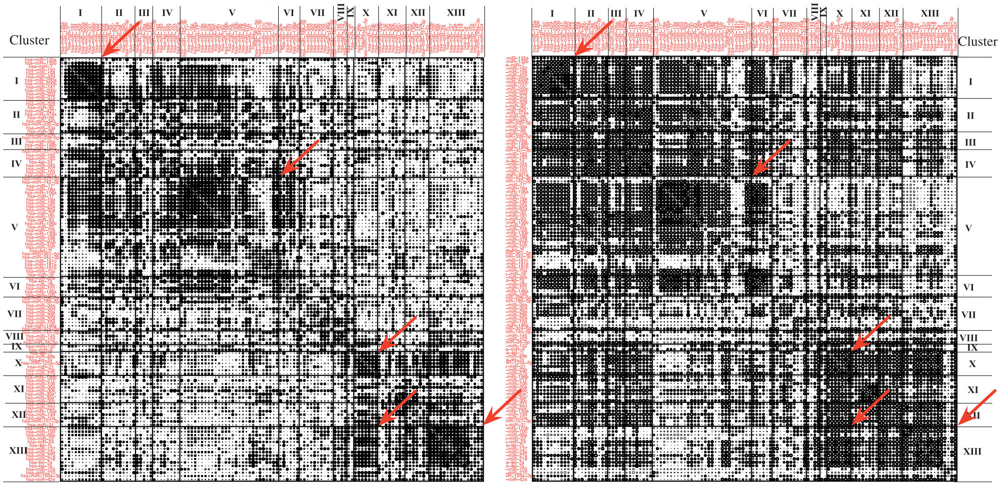
Notes: *P*-values in grey-shaded cells are the group pairs that have the number of correlation coefficients higher than expected that pass the threshold of 0.79 (FDR-adjusted and empirical *P*-value < 0.05).

and Søkilde *et al.* (2011). The datasets measured baseline microRNA expression of the 60 cancer cell lines. The primary objective for using these datasets in ORCA is to verify the groupings (clusters) generated in miRConnect study (Hua *et al.*, 2011). The secondary objective is to identify any interaction between these clusters.

The miRConnect study attempted to cluster microRNAs using a new correlation scheme (summed Pearson Correlation coefficient or sPCC) between baseline microRNA expression profiles and gene expression profiles from the same cell line panel. The study divided 136 microRNAs into 13 clusters according to the correlation patterns between microRNAs and gene expression of selected gene signatures. We focused on the microRNAs, which overlapped in both datasets, resulting in 124 microRNAs for this analysis.

Although the two datasets have the same structure, the data were generated using different microarray technologies, thus resulting in different microRNA profiles (see Supplementary files S3 and S4; Supplementary Figure S4 and S5). The threshold selection method determined correlation coefficient thresholds for Liu *et al.* (2010) and Søkilde *et al.* (2011) to be 0.29 and 0.61, respectively. Figure 3 shows correlation matrices of the two datasets and Tables 2 and 3 show FDR-adjusted *P*-values resulting from ORCA of the correlation matrices corresponding to the correlation matrices for Liu *et al.* (2010) and Søkilde *et al.* (2011) datasets, respectively.

At the thresholds identified previously for the two datasets, ORCA confirmed that several clusters as proposed by Hua *et al.* (2011) were determined to contain significant within-group overrepresentation of correlations at an FDR-adjusted significance level of 0.05 in both datasets, which are clusters I, IV, V, X



Liu et al. (2010) [Agilent chip]

Søkilde et al. (2011) [LNA chip]

Fig. 3. Correlation matrices of two microRNA datasets characterizing the NCI-60 cell panel. Circles and squares represent positive and negative correlation coefficients between two microRNAs, respectively. The sizes of circles and squares reflect the strength of correlation. The lines divide microRNAs into their clusters according to the miRConnect study. The arrows indicate the clusters and a cluster pair that are significant according to adjusted *P*-values from ORCA in both datasets. Note that the correlation thresholds for two datasets are different

Table 2. (A) Benjamini–Hochberg FDR-adjusted (lower half) and empirical (upper half) *P*-value table for Liu *et al*. (2010) microRNA dataset; (B) *P*-values from diagonal (within-group correlations)

A)

Cluster	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	Cluster
I	■	0.17	0.94	0.005	0.003	0.74	0.94	0.20	0.95	0.99	0.94	0.35	0.007	**I**
II	0.33	■	0.80	**0.04**	0.02	0.71	0.54	0.75	0.67	0.86	0.99	0.94	0.99	**II**
III	1.00	1.00	■	0.18	0.06	0.55	0.61	0.20	0.68	0.99	0.48	0.90	0.88	**III**
IV	4E-4	**0.027**	0.44	■	0.01	0.35	0.77	0.85	0.87	0.88	0.96	0.64	0.99	**IV**
V	7E-6	8E-4	0.025	4E-4	■	**0.04**	0.96	0.65	0.99	0.008	0.46	0.97	0.08	**V**
VI	1.00	1.00	1.00	0.85	**6E-3**	■	0.98	0.97	0.93	0.66	0.26	0.67	0.48	**VI**
VII	1.00	1.00	1.00	1.00	1.00	1.00	■	0.26	0.82	0.91	0.99	0.54	0.99	**VII**
VIII	0.44	1.00	0.44	1.00	1.00	1.00	0.69	■	0.85	0.63	0.97	0.99	0.61	**VIII**
IX	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	■	0.83	0.71	0.83	0.99	**IX**
X	1.00	1.00	1.00	1.00	1E-4	1.00	1.00	1.00	1.00	■	0.73	0.28	**3E-6**	**X**
XI	1.00	1.00	1.00	1.00	1.00	0.67	1.00	1.00	1.00	1.00	■	0.02	0.46	**XI**
XII	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.7	8E-3	■	0.11	**XII**
XIII	3E-4	1.00	1.00	1.00	0.031	1.00	1.00	1.00	1.00	**9E-14**	1.00	0.16	■	**XIII**
Cluster	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**XIII**	Cluster

B)

Cluster	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
Empirical p-value	**3E-6**	0.15	0.21	**0.006**	**1E-5**	0.45	0.57	0.05	0.24	**1E-4**	0.33	0.23	**4E-5**
FDR-adjusted p-value	**4E-15**	0.67	0.83	**4E-3**	**4E-15**	1.00	1.00	0.22	0.83	**4E-7**	1.00	0.87	**3E-11**

Notes: *P*-values in grey-shaded cells are the group pairs that have the number of correlation coefficients higher than expected that pass the threshold of 0.29 (FDR-adjusted and empirical *P*-value < 0.05). *P*-values in bold are the clusters and cluster pair that passed the threshold and overlapped with another microRNA dataset (also correspond to the arrows on left panel in Fig. 3). *P*-values in left-slanted cells correspond to the Figure 4.

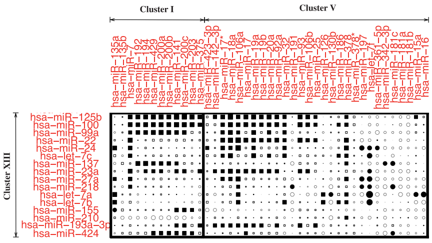


Fig. 4. Parts of the correlation matrix between the expression of selected miRNAs from the Liu et al. (2010) dataset (cluster pairs I/XIII and V/XIII). Circles and squares represent positive and negative correlation coefficients between two microRNAs, respectively. The sizes of circles and squares reflect the strength of correlation. Correlation threshold identified by threshold selection method for this dataset is 0.29. This result is consistent with the finding from Hua et al. (2013) that the microRNAs in cluster I and V are the antagonists of microRNAs in cluster XIII

Table 3. (A) Benjamini–Hochberg FDR-adjusted (lower half) and empirical *P*-value (upper half) table for Søkilde *et al*. (2011) microRNA dataset; (B) *P*-values from diagonal (within-group correlations)

A)

Cluster	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	Cluster
I	■	0.02	0.51	0.55	0.84	0.23	0.91	0.72	0.47	0.87	0.92	0.89	0.96	**I**
II	0.016	■	0.12	**0.003**	0.15	0.005	0.38	0.66	0.42	0.11	0.27	0.48	0.28	**II**
III	0.61	0.32	■	0.024	0.37	0.25	0.74	0.44	0.26	0.31	0.67	0.63	0.86	**III**
IV	0.79	**1E-3**	0.049	■	0.27	0.12	0.88	0.59	0.37	0.50	0.82	0.78	0.17	**IV**
V	0.95	0.28	0.61	0.52	■	**1E-6**	0.64	0.88	0.70	0.87	0.88	0.95	0.99	**V**
VI	0.52	1E-3	0.43	0.32	**6E-3**	■	0.16	0.50	0.30	0.69	0.73	0.69	0.91	**VI**
VII	0.95	0.61	0.83	0.92	5E-14	0.43	■	0.66	0.43	0.48	0.15	0.85	0.82	**VII**
VIII	0.83	0.79	0.61	0.75	0.85	0.68	0.79	■	0.22	0.55	0.59	0.23	0.79	**VIII**
IX	0.68	0.61	0.54	0.61	0.95	0.56	0.61	0.51	■	0.33	0.37	0.33	0.55	**IX**
X	0.92	0.26	0.51	0.61	0.86	0.79	0.68	0.72	0.59	■	0.50	0.74	**6E-5**	**X**
XI	0.93	0.55	0.79	0.86	0.95	0.83	0.32	0.75	0.61	0.61	■	0.78	0.71	**XI**
XII	0.92	0.68	0.77	0.85	0.97	0.79	0.89	0.41	0.59	0.83	0.85	■	0.07	**XII**
XIII	0.99	0.55	0.92	0.41	0.99	0.93	0.92	0.86	0.75	**6E-8**	0.86	0.13	■	**XIII**
Cluster	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**XIII**	Cluster

B)

Cluster	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
Empirical p-value	**2E-5**	0.19	0.26	**1E-4**	**3E-6**	0.006	0.38	0.17	0.03	**6E-4**	0.001	0.14	**0.003**
FDR-adjusted p-value	**2E-10**	0.57	0.61	**4E-7**	**2E-17**	0.013	0.77	0.55	0.22	**8E-5**	0.54	3E-4	**3E-11**

Notes: *P*-values in grey-shaded cells are the group pairs that have the number of correlation coefficients higher than expected that pass the threshold of 0.61 (FDR-adjusted and empirical *P*-value < 0.05). *P*-values in bold are the clusters and cluster pair that passed the threshold and overlapped with another microRNA dataset (also correspond to the arrows on right panel in Fig. 3).

and XIII. Considering associations between clusters, ORCA identified several cluster pairs where between-group correlations were overrepresented, but only one cluster pair had over-representation of correlations in both datasets: cluster pair X/XIII.

The finding that clusters I, IV, V, X and XIII contain over-representation of correlation as seen by ORCA supports the hypothesis that miRNAs within these clusters are controlled by the same transcription factors, located at the same chromosomal regions or involved in the same processes or pathways.

The significant association of cluster pair X/XIII could be explained by the fact that several of the miRNAs in the two clusters possess similar seed sequences, i.e. let-7 family, mir-23 family, mir-27 family, mir-125 family and mir-99/100 family (Lewis et al., 2005). This fact was not reported in the original study and might suggest that these two clusters be best considered as a single superfamily of miRNAs.

Other potentially related clusters we identified by ORCA, i.e. cluster pairs II/IV and V/VI, were less easy to be rationalized, as clusters 2 and 6 did not exhibit significant within-group over-representation of correlation suggest possible misclassification in the original clustering of these miRNA families.

Interestingly, a follow-up study Hua *et al.* (2013) observed that clusters I and V appeared to be functionally antagonistic to miRNAs in cluster XIII, i.e. had the opposite effect on the same set of mRNAs. ORCA was able to identify significant over-representation of (predominantly negative) correlations between clusters I and XIII and between clusters V and XIII in one of the microarray datasets (Fig. 4). This highlighted the potential of ORCA to detect such functional antagonism by analysis of miRNA co-expression alone.

4 DISCUSSION

ORCA can be used as pathway analysis tool, but the research question will be different from existing pathway or gene set analysis methods. Current methods identify pathways that are significantly enriched or depleted with respect to genes associated to a biological condition, while ORCA can identify pathways that are associated through correlations. Although the examples given here were not of typical pathways, ORCA can be applied to any type of pathway or gene set to determine pathways that are associated. In terms of pathway analysis, ORCA addresses some limitations of existing pathway analysis methods that were presented by Khatri *et al.* (2012). First, whereas ORA does not take the actual levels of variables (such as gene expression or metabolite levels) into consideration, ORCA can take these values into account, although indirectly, through the correlation coefficient calculation, which means that ORCA does not weigh the variables equally. Second, by using correlation, ORCA does not assume that the variables are independent, which is usually an important assumption in typical ORA implementations.

Third, classical ORA only uses variables, such as mRNAs or microRNAs, that are deemed most differentially expressed, while ORCA uses all the data in the calculation. Fourth, pathway analysis tools based on ORA use multiple testing corrections that assume independence of each pathway. ORCA, on the other hand, assumes the opposite and looks for the association between two sets of variables. Because our method is based on the hypergeometric test, it could be argued that ORCA violates the assumptions of the independence of each data point by using correlation coefficients as the source data. This may lead to inaccurate *P*-value calculations from the test. Goeman and Bühlmann (2007) have shown in simulated data that when hypergeometric test was used in correlated datasets, the calculated *P*-values will be underestimated. Possible remedies for the underestimated *P*-values could be multiple comparisons procedures, such as Bonferroni correction or Benjamini–Hochberg FDR correction (which was applied in this study). A table showing the nominal alpha level for correlated data from Goeman and Bühlmann (2007) can also be used to select a suitable alpha level according to a correlation threshold. A limitation of this approach is that in the simulation experiment mentioned above, all the data points have the same correlation coefficient. An alternative test that could be used instead of the hypergeometric test is the Wilcoxon rank-sum test, where the comparison is between the number of correlation coefficients in one group-pair and all other group-pairs.

ORCA can be used in pathway analysis in the same way as tools based on ORA. However, the information that will be derived from ORCA relates specifically to the pathway interactions. For pathway or gene set analysis, ORCA can calculate the correlation coefficients between genes or metabolites and then find the pathway-pairs that have more correlation coefficients that pass a certain threshold (determined by the threshold selection method or by other means) than expected.

Our version of ORCA requires group membership of variables to be mutually exclusive. Therefore, it cannot yet be used with pathway data where group members overlap, i.e. variables can not belong to more than one group. This is the subject of future work.

Recently gene co-function networks were used to identify cross-category association between different GO classes (CroGo, Peng *et al.*, 2013). However, CroGO does not explicitly include gene expression values into the analysis and, therefore, could miss actual association between GO categories in the real biological context. In this regard, ORCA can be a crucial downstream analysis to CroGO to highlight associations between GO categories that are of greatest importance, using gene expression to complement the associations identified by CroGO.

In conclusion, ORCA is a new method that combines analysis of correlation with ORA, and has the potential to reveal otherwise obscured associations between sets of variables, whether they are genes, proteins, metabolites or other molecular signals, in a wide variety of biological datasets. Although the method has clear application in '-omics' data analysis, ORCA can be profitable in any circumstance where an association network can be constructed between variables that can be classified into meaningful sets.

ACKNOWLEDGEMENTS

We would like to thank Dr Rachel Cavill for the discussions during the early development of this method.

Funding: Yotsawat Pomyen is funded by Ministry of Science and Technology, Royal Thai Government.

Conflict of Interest: none declared.

REFERENCES

Bargonetti,J. *et al.* (2010) Differential toxicity of DNA adducts of mitomycin C. *J. Nucleic Acids*, **2010**, 6.

Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Benjamini,Y. and Hochberg,Y.B. (2009) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Cavill,R. *et al.* (2011) Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. *PLoS Comput. Biol.*, **7**, e1001113.

Davison,A.C. and Hinkley,D.V. (1997) *Bootstrap Methods and Their Application.* 9th edn. Cambridge University Press, New York, NY.

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

Hua,Y. *et al.* (2013) miRConnect 2.0: identification of oncogenic, antagonistic miRNA families in three human cancers. *BMC Genomics*, **14**, 179.

Hua,Y. *et al.* (2011) miRConnect: identifying effector genes of miRNAs and miRNA families in cancer cells. *PLoS One*, **6**, e26521.

Khatri,P. *et al*. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.

Liu,H. *et al*. (2010) mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Mol. Cancer Ther.*, **9**, 1080–1091.

Peng,J. *et al*. (2013) Identifying cross-category relations in Gene Ontology and constructing genome-specific term association networks. *BMC Bioinformatics*, **14** (Suppl. 2), S15.

Rudolf,E. *et al*. (2011) Camptothecin induces p53-dependent and -independent apoptogenic signaling in melanoma cells. *Apoptosis*, **16**, 1165–1176.

Scherf,U. *et al*. (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, **24**, 236–244.

Shannon,C. (1948) A mathematical theory of communination. *Bell Syst. Tech. J.*, **27**, 379.

Søkilde,R. *et al*. (2011) Global microRNA analysis of the NCI-60 cancer cell panel. *Mol. Cancer Ther.*, **10**, 375–384.

Lewis, B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Tavazoie,S. *et al*. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Zeeberg,B.R. *et al*. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.