DEVELOPMENT OF A WEB APPLICATION/DATABASE FOR THE INTEGRATIVE ANALYSIS OF microRNA EXPRESSION PATTERNS

A THESIS SUBMITTED TO
THE DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS AND
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF
BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY

BY
KORAY DOĞAN KAYA
August, 2011

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy

Assist. Prof. Dr. Özlen Konu

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy

Prof. Dr. Volkan Atalay

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy

Prof. Dr. Mehmet Öztürk

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy

Assoc. Prof. Dr. Işık Yuluğ

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy

Assist Prof. Dr. Ayşe Elif Erson Bensan

Approved for the Graduate School of Engineering and Science

Director of Graduate School of Engineering and Science
Prof. Dr. Levent Onural

ABSTRACT

DEVELOPMENT OF A WEB APPLICATION/DATABASE FOR THE INTEGRATIVE ANALYSIS OF microRNA EXPRESSION PATTERNS

Koray Doğan Kaya

Ph.D. Thesis in Molecular Biology and Genetics

Advisor: Assist. Prof. Dr. Özlen Konu

August 2011, 125 pages

microRNAs, small non-coding RNA molecules with important roles in cellular machinery, target mRNAs for silencing by binding generally to their 3' UTR sequences via partial base complementation. Thus, microRNAs with similar sequences also might exhibit expression and/or functional similarities. In this study, a modular tool, mESAdb (http://konulab.fen.bilkent.edu.tr/mirna/), was developed allowing for multivariate analysis of sequences and expression of microRNAs from multiple taxa. Its framework comprises PHP, JavaScript, packages in the R language, and a database storing mature microRNA sequences along with microRNA targets and selected expression data sets for human, mouse and zebrafish. mESAdb allows for: (i) mining of microRNA expression data sets for subsets of microRNAs selected manually or by a sequence motif; (ii) pair-wise multivariate analysis of expression data sets within and between taxa; and (iii) association of microRNA subsets with annotation databases, *HuGE Navigator*, *KEGG* and *GO*. mESAdb also permits user specified dataset upload for these analyses. Herein, utility of mESAdb was illustrated using different datasets and case studies. First, it was shown that microRNAs carrying the embryonic stem cell specific seed sequence, 'AAGTGC', were able to discriminate between normal and tumor tissues from hepatocellular carcinoma patients using dataset GSE10694. Second, mRNA targets of a set of liver specific microRNAs were annotated with human diseases based on *HuGE Navigator*. Third, the similarity between mouse and human tissue specificity of a given set of

microRNAs was demonstrated. Forth, CHRNA5 targeting microRNAs were associated with estrogen receptor status in breast cancer using dataset GSE15885. Finally, a related tool under development for mRNA arrays planned for integration with mESAdb was presented.

Keywords: mESAdb, database, R, microRNA, sequence, expression data sets, data mining, multivariate analysis, annotation databases, HuGE, KEGG, GO, CHRNA5, estrogen receptor, hepatocellular carcinoma and breast cancer..

ÖZET

MikroRNA İFADE ÖRÜNTÜLERİNİN BÜTÜNLEŞTİRİCİ ANALİZİ İÇİN AĞ ARACI/VERİTABANI GELİŞTİRİLMESİ

Koray Doğan Kaya

Doktora Tezi, Moleküler Biyoloji ve Genetik

Danışman: Yrd. Doç. Dr. Özlen Konu

Ağustos 2011, 125 sayfa

mikroRNA'lar protein kodlamayan, küçük ve hücrelerdeki mekanizmalarda önemli rolleri olan RNA molekülleri olup genellikle mesajcı RNA'ların protein kodlamayan 3' bölgesine kısmi baz eşlemesi yoluyla bağlanır ve onların proteine çevrilmesine engel olur. Bu nedenle dizilerinde benzerlik gösteren mikroRNA'lar fonksiyonel ve/veya ifade düzeyi olarak da benzerlik gösterebilirler. Bu çalışmada, değişik taksonlardan gelen mikroRNA'ların dizi ve ifadelerinin çok değişkenli analizlerini yapmak için modüler bir araç/veritabanı olan mESAdb (http://konulab.fen.bilkent.edu.tr/mirna/) geliştirilmiştir. Omurgası, PHP, JavaScript, R paketleri ve ağırlıklı olarak, insan, fare ve zebrabalığı için, mikroRNA'ların olgun dizilerini, onların hedef genlerini ve seçilmiş mikrodizin veri setlerini depolayan bir veritabanından oluşur. mESAdb üç önemli kullanıma olanak verir: (i) dizi motifi veya seçilen mikroRNA'lar ile ifade veri madenciliği; (ii) taxonlar arası ikili veri setlerinin çok değişkenli analizi; (iii) mikroRNA gruplarının referans isimlendirme veri tabanları, örneğin *HuGE*, *KEGG* ve *GO*, ile ilişkilendirilmesi. mESAdb kullanıcıların özgün ifade veri setlerini yükleyip analiz etmelerine de izin vermektedir. Bu çalışmada, mESAdb kullanımı değişik veri setleri ve örnek durumlar ile anlatılmıştır. İlk olarak, embriyoya özgü kök hücre mikroRNA çekirdek dizisi AAGTGC'yi taşıyan mikroRNA'ların, GSE10694 veri seti kullanılarak, hepatosellüler karsinom hastalarından alınan tümörlü ve normal karaciğer dokularını ayrıştırabildiği gösterildi. İkinci olarak, karaciğere özgün bir gurup mikroRNA'nın hedef aldığı mRNA'lar HuGE Navigator veri tabanı esas alınarak insan hastalıklarına ait referans terimlerle ilişkilendirildi. Üçüncü olarak, seçilen bir gurup mikroRNA,

insan ve fare için doku iyeliği bakımından karşılaştırıldı. Dördüncü olarak, GSE15885 veri-seti kullanılarak, CHRNA5 genini hedef alan mikroRNA'lar östrojen duyargaç (ER) gen ifadesi bakımından farklılık gösteren meme kanseri örnekleri ile ilişkilendirildi. Son olarak, mESAdb ile benzer omurga kullanılarak mRNA dizin çalışmalarının analizi için tasarlanan ve yapımı devam eden bir başka çalışma tanıtıldı.

Anahtar Sözcükler: mESAdb, veritabanı, R, mikroRNA, dizi, gen ifade very setleri, very madenciliği, çok değişkenli analiz, referans isimlendirme veritabanları, HuGE, KEGG, GO, CHRNA5, östrojen duyargaçı, hepatosellüler karsinom and meme kanseri.

ACKNOWLEDGEMENT

First and foremost, I wish to thank my advisor Dr. Ozlen Konu for her invaluable guidance, patience and support during my studies. I have learned from her special ways of coping with the stress on the path leading to success in academic life. I always admired her positive attitude for every worst case and will go on taking it as a model throughout my life.

I have to acknowledge Prof. Dr. Mehmet Ozturk for believing and furthermore making me believe that I could manage to handle the scientific work in hand. He has always supported me in my academic life at Bilkent with his criticisms and suggestions for my works. I would also like to thank him for kindly agreeing to evaluate my PhD dissertation.

I am very pleased to extend my thanks to Dr. Aybar Acar for his contributions in the improvement of mESAdb, and his mentoring in my efforts to improve my coding abilities and other computer skills. He added too much to my scientific endowment although we worked together for a relatively short time.

I appreciate Assoc. Prof. Dr. Cengiz Yakıcıer for initiating my interests on microRNAs that ended up in the publication of mESAdb, and for his friendship and for motivating me through the intellectual conversations that we had.

I would like to thank Gökhan Karakülah, firstly for his contributions in creating the mESAdb and secondly for his great friendship. I am happy to be in collaboration with him for ongoing and further studies.

It is impossible not to thank Prof. Dr. Volkan Atalay, Assoc. Prof. Dr. Işık Yuluğ, and Assist. Prof. Dr. Ayşe Elif Erson Bensan for taking the time to read and evaluate my PhD dissertation.

Special thanks go to Dr. Rengül Çetin Atalay for sharing their resources in making our Server and Workstation facilities work properly, and to her students Dr. Sinan Saraç, Dr. Zerrin Işık and PhD candidate Tülin Erşahin for helping me have access to these facilities.

I would like to thank Assist. Prof. Dr. Uygar Tazebay for his friendship and for being a model for me by showing his scientific enthusiasm in every conversation

with me.

It is a pleasure for me to thank Gizem Ölmezer for the scientific discussions we made that helped a lot in shaping case studies for my thesis.

I am also thankful to one of my best friends so far, Hasan Colak, for his help in choosing heart touching thank words and also his invaluable friendship for many years.

I would like to thank Mr. Zihni Yalçın, for sharing all of his social power and his invaluable humanism for my success and to feel myself safe and happy in United Kingdom. Thanks for his life long friendship.

I would like to thank all Bilkent MBG family for providing me a nice environment in where I have felt very happy.

Last but not least, I want to thank my family for everything, especially my mother Selviye Kaya and my father Nami Kaya. They have done everything for my success with all power they have.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

Amy1	Amylase 1
API	Application Programming Interface
ARC	Annotation and Regulation of Co-Expression
B	Brain
Bl	Bladder
Br	Breast
CA	Correspondence Analysis
CHRNA5	Cholinergic Receptor, Nicotinic, Alfa Subunit 5
CIA	Co-Inertia Analysis
Co	Colon
CSC	Cancer Stem Cell
CSV	Comma Separated Values
DBMS	Database Management Systems
E.coli	*Escherichia coli*
EMBL-EBI	European Molecular Biology Laboratories - European Bioinformatics Institute
En	Endometrium
ER	Estrogen Receptor
GEO	Gene Expression Omnibus
GO	Gene ontology
GPL	Gene Expression Omnibus Platform
GRSN	Global Rank-Invariant Set Normalization
GSE	Gene Expression Omnibus Series

GWAS	Genome Wide Association Studies
H	Heart
HCC	Hepatocellular Carcinoma
HER2	Human Epidermal Growth Factor Receptor 2
HMDD	Human MicroRNA Associated Disease Database
HuGE	Human Genome Epidemology
HuGENet	Human Genome Epidemology Network
ISMB	Intelligent Systems for Molecular Biology
IUPAC	International Union of Pure and Applied Chemistry
K	Kidney
KEGG	Kyoto Encyclopedia of Genes and Genomes
Li	Liver
Lu	Lung
Ly	Lymph Node
MAS5	Microarray Statistical Algorithm Software Developers Kit
mESAdb	MicroRNA Expression and Sequence Analysis Database
MeSH	Medical Subject Headings
MGI	Mouse Genome Informatics
MIAME	Minimum Information About a Microarray Experiment
N	Normal
NCBI	National center for Bioinformatics
Ng	Negative
O	Ovary
PCA	Principal Component Analysis

PCs	Principal Components
PHP	Hypertext Processor
Pl	Placenta
PR	Progesterone Receptor
Pr	Prostate
Ps	Positive
RISC	RNA-Induced Silencing Complex
RMA	Robust Multichip Average
RNA	Ribonucleic Acid
RV	Realised Volatility
SGD	Saccharomyces Genome Database
SM	Sceletal Muscle
SNAP	SNP Annotation and Proxy Search
SNP	Single Nucleotide Polymorphism
SOFT	Simple Omnibus Format in Text
SQL	Structured Query Language
Te	Testicle
Th	Thymus
TIGR	The Institute of Genomic Research
UCSC	University of California Santa Cruz
UI	User Interface
UTR	Untranslated Region

CHAPTER 1:INTRODUCTION

1.1 MICRORNAS

In 1993, Victor Ambross and colleagues announced that the transcript of the gene called *lin4*, regulating the timing of *C. elegans* larval development, did not code for a protein. Instead, this gene produced two short RNAs, 22 and 61 nucleotides in length, respectively (Lee, Feinbaum et al. 1993). This aforementioned gene has been the founding member of a non-coding RNA gene family, called microRNAs (Bartel 2004). The second member of this large RNA family is *Let-7*, that also has been discovered in *C. elegans* and was found to regulate the transition from late larval to adult stage in a similar way as *lin-4* regulated the timing between the first and the second larval stages of worm development (Reinhart, Slack et al. 2000; Slack, Basson et al. 2000).

Mature microRNAs are small (19–22 nt) RNAs that play crucial roles in many cellular processes via targeting mRNAs for translational repression or cleavage thus regulating gene expression (Bartel 2004). MicroRNAs, through their compatible 5'-seed sequences, exert regulatory functions primarily on the 3'-untranslated regions (UTRs) of targeted mRNAs (Lewis, Burge et al. 2005; Grimson, Farh et al. 2007; Iwama, Masaki et al. 2007). They are functional in crucial roles, such as development (Lee, Feinbaum et al. 1993; Boutet, Vazquez et al. 2003; Krichevsky, King et al. 2003; Alvarez-Garcia and Miska 2005; Shi and Jin 2009), apoptosis (Brennecke, Hipfner et al. 2003; Bartel 2004; Lynam-Lennon, Maher et al. 2009), differentiation (Kawasaki and Taira 2003; Kawasaki and Taira 2003; Shi and Jin 2009) and metabolism (Xu, Vernooy et al. 2003; Jordan, Kruger et al. 2011; Tamasi, Monostory et al. 2011) in animals. Evidence from earlier studies suggesting the participation of microRNAs in a long list of human diseases, especially different cancers, implies the importance of microRNAs (Alvarez-Garcia and Miska 2005; Gregory and Shiekhattar 2005; Chhabra, Dubey et al. 2010; Wang, Yu et al. 2011; Zhang, Yan et al. 2011).

According to *miRBase* (Ambros, Bartel et al. 2003; Griffiths-Jones 2004;

Griffiths-Jones, Grocock et al. 2006; Griffiths-Jones, Saini et al. 2008; Kozomara and Griffiths-Jones 2011) release 16, the number of discovered microRNA genes in different organisms varies in range between 1 to 1048. The highest entry number belongs to *Homo sapiens* in the range while the total number of microRNA entries was 15172 in release of *miRBase* declared above.

However, the discovery of new microRNAs has slowed down. Although the fact that mature sequences of a broad number of microRNAs are conserved among different taxa (Wheeler, Heimberg et al. 2009), the wide range of microRNA entries for different organisms in *miRBase* shows that new research technologies or broader application of *in silico* methods (Li, Xu et al. 2010) are needed to balance the entry numbers across organisms, which in turn will increase the pace of the novel microRNA discovery. Indeed, next generation sequencing has now provided a lead for novel microRNA discovery (Morin, O'Connor et al. 2008; Li, Chan et al. 2010).

1.2 MICRORNA TRANSCRIPTION, MATURATION AND FUNCTION

Although some microRNA genes are located in intronic regions of protein coding genes, many of them are intergenic (Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001; Lee and Ambros 2001; Aravin, Lagos-Quintana et al. 2003; Lagos-Quintana, Rauhut et al. 2003; Lai, Tomancak et al. 2003; Lim, Glasner et al. 2003; Lim, Lau et al. 2003; Saini, Griffiths-Jones et al. 2007), meaning that they are in between genes having their own transcription units (Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001; Lee and Ambros 2001).

microRNA genes can be located in an isolated fashion, as seen in human and worm genomes (Lim, Glasner et al. 2003; Lim, Lau et al. 2003), or can be clustered together producing multi-cistronic transcripts, as commonly seen in *Drosophila* genome (Aravin, Lagos-Quintana et al. 2003).

The initial transcript forms of miRNAs are called primary microRNAs or pri-miRNAs (Lee, Feinbaum et al. 1993). They are either a form of non-coding RNA transcribed by RNA polymerase II or spliced intronic parts of pre-mRNAs (Krol,

Loedige et al. 2010). Although some previous studies have attempted to derive a full definition of primary transcript for a microRNA or a cluster of them from a non-coding unit, it was not until 2007 that the boundaries of many pri-microRNAs in the human genome have been precisely predicted (Saini, Griffiths-Jones et al. 2007).

The first step in the maturation of an approximately ~22nt RNA including the 5' seed is splicing of the pri-miRNA by Drosha, an enzyme of RNase III type (Lee, Ahn et al. 2003). This enzyme recognizes stem loop structure and cuts the long stem and liberates a shorter, ~60-70 nt, hairpin, stem-loop molecule with 2-3 nt 3' overhang (Basyuk, Suavet et al. 2003). Next, this molecule is transported to the nucleus by exportin 5 (Yi, Qin et al. 2003; Lund, Guttinger et al. 2004). Another RNase III enzyme, Dicer cuts the loop and produces a double stranded RNA with 2-3 nt 3' overhang at both sides (Lee, Ahn et al. 2003). Afterwards, this dsRNA is integrated into miRNA, mediated by RNA interference genes silencing complex, miRISC (Bartel 2004; Meister and Tuschl 2004; Murchison and Hannon 2004; Krol, Loedige et al. 2010). Here, the mature single stranded microRNA is partially complementary to mRNA 3'UTR region of target gene and translation stop codons. Even in some cases, de-adenylation occurs after targeted mRNA is degraded (Giraldez, Mishima et al. 2006; Wu, Fan et al. 2006).

Intersetingly, few earlier studies that tried to report target:microRNA relationship in depth have found that there are some among-species conserved sequences in coding regions/ORFs (John, Enright et al. 2004; Lewis, Burge et al. 2005). Then in 2008, series of studies have been published reporting some microRNAs might target coding regions. In early 2008, it has been experimentally validated that p16 is a target of *miR-24* as predicted by Miranda (Enright, John et al. 2003) and that *miR-24* binds to regions both in 3' UTR and the coding region of p16. Later on, a study has announced that *miR-126* represses Hoxa9 by binding an across species conserved site at its Homeobox domain (Shen, Hu et al. 2008). Then, it was reported that a conserved site (between nucleotides 2382 and 2412) existed on DNA Methyl-transferase 3b (DNMT3b) mRNA as the target of *miR-148* (Duursma, Kedde et al. 2008). A comprehensive scanning study seeking for highly conserved sequences in coding regions of all genes from 17 genomes showed that conserved

sites were generally microRNA targets (Forman, Legesse-Miller et al. 2008). The study has also solidly shown that *let-7* family targets Dicer in three regions on its coding sequence, forming a negative feedback loop on microRNA function (Forman, Legesse-Miller et al. 2008). Apart from those which concentrated on conserved sequences in coding regions, another study has shown that some microRNA targets at exon-exon junctions seen in mouse are not conserved in humans (Tay, Zhang et al. 2008). However a relatively recent study claims that although the target sites for microRNAs in the coding regions are functional, the effects are weaker compared to the ones in the 3' UTRs (Forman and Coller 2010).

1.2.1 microRNA expression profiles

The first evidence for the expression profile specificity of microRNAs for different developmental stages and tissue types came from Northern blot studies and/or cloning efforts followed by sequencing (Pasquinelli, Reinhart et al. 2000; Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001; Lee and Ambros 2001; Lagos-Quintana, Rauhut et al. 2002; Aravin, Lagos-Quintana et al. 2003; Bashirullah, Pasquinelli et al. 2003; Basyuk, Suavet et al. 2003; Houbaviy, Murray et al. 2003; Lagos-Quintana, Rauhut et al. 2003; Lai, Tomancak et al. 2003; Lim, Glasner et al. 2003; Lim, Lau et al. 2003; Chen, Li et al. 2004). As expected, the first remarkable scientific discoveries in regard to these specificities were about the founding members of the microRNA gene family, *lin4* and *let-7*. Accordingly, they were found to be temporarily expressed in specific larval stages of *C. elegans* (Pasquinelli, Reinhart et al. 2000; Lau, Lim et al. 2001; Lagos-Quintana, Rauhut et al. 2002; Bashirullah, Pasquinelli et al. 2003; Lim, Lau et al. 2003).

After the discovery that microRNAs were not restricted to worms (Pasquinelli, Reinhart et al. 2000), many new members discovered by cloning and sequencing, also made their tissue specificity clear. Among those members were *miR-1*, expressed mainly in the mammalian heart (Lee and Ambros 2001; Lagos-Quintana, Rauhut et al. 2002), *miR-122* which was specific to liver (Lagos-Quintana, Rauhut et al. 2002), and *miR-223* expressed in mouse granulocytes and macrophages derived from bone marrow (Chen, Li et al. 2004). Another interesting discovery in

this field has been that there were embryonic stem cell specific microRNAs, namely, the *miR-290/mir-295* cluster that was only expressed in mouse embryonic stem cells (Houbaviy, Murray et al. 2003).

Those findings encouraged application of high throughput technologies to explore, more broadly, such findings as mentioned above. Finally, a study has shown that microRNAs have distinct expression patterns in different developmental stages and regions of the mammalian brain by using an array expression technology (Krichevsky, King et al. 2003). Based on further large-scale studies, microRNAs were annotated for their specificity for particular tissues, developmental stages and/or pathologies such as cancer (Houbaviy, Murray et al. 2003; Liu, Calin et al. 2004; Sempere, Freemantle et al. 2004; Sun, Koo et al. 2004). These individual studies then could be complied using meta-analysis methods: for example, Bargaje et al. (Bargaje, Hariharan et al. 2010) compiled and normalized multiple data sets from different sources to determine the tissue-specific and tissue-invariant consensus expression profiles. Others have surveyed microRNA expression profiles in large numbers of normal and cancerous tissues to decipher microRNA networks and conserved expression clusters in disease (Navon, Wang et al. 2009). There also is evidence suggesting that expression patterns of microRNAs are conserved at the species level (Hertel, Lindemeyer et al. 2006). However, development of database/tools that encompass tissue specific datasets with ability to analyze for a specific set of microRNAs in a multivariate fashion is needed.

1.2.2 MicroRNA databases

In recent years, several databases and analysis tools have also been published featuring high-throughput analysis results of microRNA sequence or expression. Among these, *miRBase* functions as a central repository for microRNA genomics for a variety of organisms and thus serves the community with up-to-date microRNA sequence, chromosome location and transcript information (Ambros, Bartel et al. 2003; Griffiths-Jones 2004; Griffiths-Jones, Grocock et al. 2006; Griffiths-Jones, Saini et al. 2008; Kozomara and Griffiths-Jones 2011). *mSigDB*, using motif lists from Xie et al. (Xie, Lu et al. 2005), provides microRNA target gene lists that could

be tested for enrichment with Gene Ontology (*GO*) functional terms, *KEGG* signaling pathways or other gene lists (Subramanian, Kuehn et al. 2007). Similarly, a manually curated database, called *Mir2DiseaseBase*, can be used for extracting associations between diseases and microRNAs (Jiang, Wang et al. 2009). Most recently, *miRBridge* has been developed to predict microRNA function and link microRNAs with cellular pathways using network algorithms (Tsang, Ebert et al. 2010). Among the expression analysis focused databases, *miRGator* is a comprehensive repository and analysis tool for microRNA expression, target and ontology data providing a graphical transcriptional evaluation of selected microRNA types for mice or humans (Nam, Kim et al. 2008). *MicroRNA.org* is another source of microRNA expression and functional data for understanding microRNA expression regulation through target prediction and examination of tissue transcript abundance (Betel, Wilson et al. 2008).

1.2.2.1 miRBase

It is the pioneer database that provides all microRNA sequence data, annotation and target information (Griffiths-Jones, Grocock et al. 2006). After the sharp increase in the number of annotated miRNAs from a variety of organisms (Pasquinelli, Reinhart et al. 2000; Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001; Lee and Ambros 2001; Lagos-Quintana, Rauhut et al. 2002; Aravin, Lagos-Quintana et al. 2003; Bashirullah, Pasquinelli et al. 2003; Basyuk, Suavet et al. 2003; Houbaviy, Murray et al. 2003; Lagos-Quintana, Rauhut et al. 2003; Lai, Tomancak et al. 2003; Lim, Glasner et al. 2003; Lim, Lau et al. 2003; Chen, Li et al. 2004), a registry was needed to bring the existing data together and thus, the microRNA registry was established (Griffiths-Jones 2004) according to the nomenclature proposed by Ambros (Ambros, Bartel et al. 2003). Now it survives as *miRBase* (Kozomara and Griffiths-Jones 2011).

In *miRBase*, the name of every microRNA entry in the database has a 3 or 4 letter prefix to specify their source species, e.g., ‘hsa’ for *Homo sapiens*, ‘mmu’ for *Mus musculus,* etc.. After this prefix, ‘mir’ is used to annotate precursor hairpins whereas ‘miR’ is assigned to mature sequences.

1.2.3 microRNA - target relationsip

~20 bp Dicer cut output for pre-mRNA or mirton, pre-microRNA derived from splicing events, interacts with mRNA in RISC complex (Krol, Loedige et al. 2010). Studies for determining the microRNA- target coupling features have shown that conserved perfect ~6-8 base pair matchings (seed matches) at the 5' end of microRNAs are relaible recognition sequences for those interactions (Lewis, Shih et al. 2003; Brennecke, Stark et al. 2005; Krek, Grun et al. 2005; Lewis, Burge et al. 2005; Lim, Lau et al. 2005; Jackson, Burchard et al. 2006). Seed sequences are generally well conserved also between paralogues microRNAs. However, the binding sites in microRNAs are classified into three types: (i) canonical, where 5' is dominant; (ii) seed only; and (iii) 3' compensatory (Maziere and Enright 2007).

In the first one, perfect base pairing is observed at the seed region and perfect match extends almost throughout the end of the 3' site. In this type there is a classical bulge in the middle. In the second type only the seed region has perfect match while in the last one, seed region has mismatches and there are long strecthes pf perfect matches towards the 3'end.

Another common rule that have been integrated to algorithms is the free energy of microRNA:mRNA duplexes (Min and Yoon 2010). To calculate the free energies of RNA foldings and base pairings some consensus programs have been used such as Vienna Package (Wuchty, Fontana et al. 1999), RNAfold (Hofacker 2003) and Mfold (Mathews, Sabina et al. 1999). The free energy thresholds are species specific (Watanabe, Tomita et al. 2007).

The core of microRNA target prediction methods relies on base paring method defined by Lewis et al. (2003). There are other prediction algorithms (Min and Yoon 2010) that use different features including evolutionary conservation (Lewis, Shih et al. 2003; Grun, Wang et al. 2005), secondary structure of target mRNA (Kertesz, Iovino et al. 2007; Long, Lee et al. 2007) and nucleotide composition of target mRNA (Grimson, Farh et al. 2007). The current published algorithms for predicting mRNA target sites have been well established (Watanabe, Tomita et al. 2007; Alexiou, Maragkakis et al. 2009; Min and Yoon 2010), and some essential ones are mentioned here.

TargetScan (Lewis, Shih et al. 2003; Lewis, Burge et al. 2005; Grimson, Farh et al. 2007; Friedman, Farh et al. 2009) considers 2nd to 8th nucleotides from 5' end of microRNA as seed sequence and seeks for perfect match for it in the target 3' UTR. Then it expands the seed match for other part of microRNAs. Another feature that the algorithm takes into account is free energy of binding of microRNA:miRNA dublexes via RNAFold. TargetScanS is an improved version of TargetScan. The differences between them are substantial ones. The most prominent difference between them is that TargetScanS uses two or more species for sequence conservation check instead of using thermodynamic stability check.

PicTar (Grun, Wang et al. 2005) focuses on multiple conserved seed sequence binding sites across species co-regulated by multiple microRNAs instead of looking for one seed sequence binding at an expected site on 3' UTR. miRanda (Enright, John et al. 2003) has originally been developed for fly microRNA targets, was then applied to predict targets of microRNAs in human. Three features have been used in the development of the algorithms. One is position-weighted base complementation, the other one is free energies of RNA:RNA duplexes and final feature is conservation of the target sites among 10 species. Further, a strict rule that requires perfect complementation of the seed region has been added to the algorithm (John, Enright et al. 2004). The targets predicted by this algorithm have been served as a service called MicroCosm targets among EMBL-EBI tools. DIANA (Maragkakis, Alexiou et al. 2009; Maragkakis, Reczko et al. 2009), uses a 38 bases window in length and sliding it through 3' UTR of the target. Then it selects the binding with the lowest free energy. In contrast to others, it allows weak pairing for seed regions.

Seed sequences and patterns on other parts of mature microRNAs also could be important in functioning of microRNAs, hence, this may be seen as a reflection in microRNA tissue specificity. All in all, a recent study has shown that unsupervised clustering of microRNA seequence was successful in separation of invasive and non-invasive breast cell carcinoma samples from different patients (Farazi, Horlings et al. 2011). In another study, a microRNA profiling has revealed a unique embryonic stem cell signature dominated by a single seed sequence (Laurent, Chen et al. 2008).

1.2.3.1 MicroCosm

MicroCosm (http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/) (Enright, John et al. 2003; Griffiths-Jones 2004; Griffiths-Jones, Saini et al. 2008) is a web tool for investigating microRNA targets in many species. The pipe for reaching the targets uses two main resources, miRBase for microRNA sequences and Ensembl (Flicek, Amode et al. 2011) for genomic sequences. The web resource currently uses *miRanda* algorithm (Enright, John et al. 2003) to select potential target sites. The current version (version 5) of *MicroCosm* uses dynamic programming for aligning the seed sequence of microRNAs with the sites identified by miRanda. Each alignment has a score between 0 and 100. The method applied uses strict rules. The most determining one is allowance of only one mismatch between seed sequence and the target sequence. Also the target site should be conserved in at least two species.

1.3 GO DATABASE

It is important to associate each microRNA and its target genes with a functional term. This can be accomplished by using ontology databases. Gene ontology is a set of controlled vocabulary that describes the role of genes in a cell. To produce an agreed upon and stable vocabulary regarding attributes of genes, efforts began as a collaborative work based on three databases, *Saccharomyces* Genome Database (Cherry, Adler et al. 1998), Mouse Genome Database (Blake, Bult et al. 2011) and *FlyBase* (Tweedie, Ashburner et al. 2009) in 1998 (Ashburner, Ball et al. 2000; Lewis 2005). As databases arose to meet the requirement of integrating information for different model organisms, scientists recognized that there was a common problem of classification. For separate projects, functional classifications had started individually, for example for *E. coli,* Monica Riley created one in 1993 (Riley 1993) and for the *FlyBase*, Ashburner created another one (Lewis 2005). The Institute of Genomic Research (*TIGR*) (http://www.jcvi.org/) also created its own functional classification system (Lewis 2005). At the end, Asburner proposed a solution involving a simple hierarchical controlled vocabulary to define the common

functional classification in bio-ontologies workshop, Intelligent Systems for Molecular Biology (*ISMB*) international conference, Montreal. Although the proposal had been dismissed by other participants, representatives of *MGI*, *FlyBase* and *SGD* have agreed to use the same vocabulary (Lewis 2005). Then the Gene Ontology Consortium was thus founded.

In GO, the vocabulary is grouped under three headings; *cellular component*, *molecular function* and *biological process*. Within these headings, vocabularies have a hierarchical relationship and are structured as directed acyclic graphs as seen in Figure 1.3.1 (Ashburner, Ball et al. 2000; Binns, Dimmer et al. 2009). This means that gene ontology entries can be queried at different levels from most general ones to the most specific one, and any member of the vocabulary set can be a more specific expansion of more than one general term. For example, GO:0016160 refers to the *GO* term amylase activity, 206 genes have the same *GO* identity while gene Amy1 from *Mus musculus* also has another *GO* term associated with it, GO:0003824, i.e., catalytic activity. Catalytic activity is a more general term then amylase activity. The relationship between these two terms can be seen in Figure 1.3.1.

Cellular Component is a part of a larger object in the cell. It may be an anatomical structure such as *nucleus* or *golgi* apparatus. It also may define a gene product group i.e., a protein dimer (Ashburner, Ball et al. 2000).

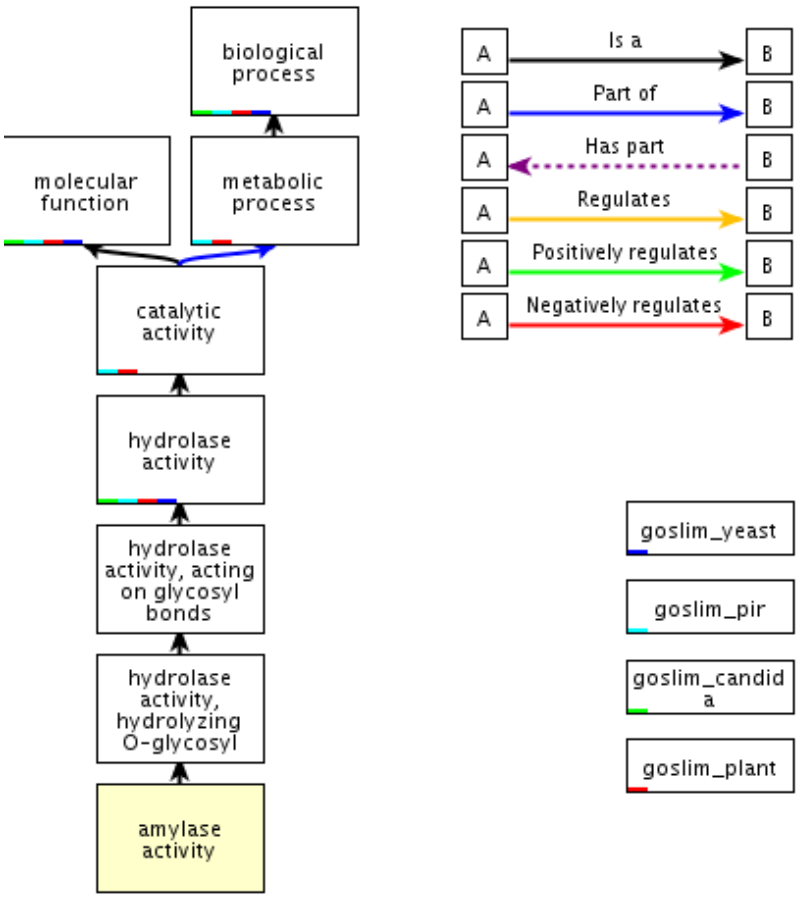


Figure 1.3.1: A sample graph view presents the hierarchical relationship between GO terms. The figure has been generated by QuickGO (Binns, Dimmer et al. 2009) linked in The Gene Ontology Consortium web page.

A biological process refers to sequential events that are accomplished by one or more ordered molecular functions. *Signal transduction* is an example in broad scope whereas *alpha-glucoside transport* is an example of a more specific term (Ashburner, Ball et al. 2000).

Molecular function ontology header describes a simple chemical activity such as catalyzing or binding at molecular level. Generally, it comprises of actions taken by one gene product, however it also includes the ones taken by assembled protein complexes. *Binding activity* is an example of general term in this category however, *toll receptor binding* is narrower one (Binns, Dimmer et al. 2009).

The Gene Ontology Consortium tool AMIGO gives associations between

some microRNA genes and *GO* terms, however it is very limited. For example although human genome has *miR-15a* as annotated gene, *AMIGO* does not show *Homo sapiens* among the species filter when the microRNA is searched in it.

1.4 KEGG: KYOTO ENCYCLOPEDIA OF GENES AND GENOMES

The KEGG database was launched as a project of Japanese Human Genome Program in 1995 (Kanehisa 1997). From its foundation up to its August 2010 update (Kanehisa, Goto et al. 2010) functions in the cell and organism behaviors has been kept in the forms that computers could process them. Molecular networks forms are the most popular ones among them. They are called as pathway maps. Another important form is hierarchical lists called as *BRITE* functional lists. These structures have been widely used for interpreting the large scale outputs of high-throughput experimental technologies such as sequencing and microarray technologies.

The knowledge in *KEGG* Project has started to be focused on human diseases and drugs. These new focuses have been integrated to the structures that are readily could be processed by computers in such a way that human diseases have been described as perturbed states of molecular systems that operate cells whereas drugs have been defined as perturbants to them.

The latest report describing updates completed up to August 2010 says that *KEGG* Project has been constructed from 16 main databases. However in the work described in this thesis has used only *KEGG PATHWAY* database, manually drawn pathways collection. Even the only pathways and diseases that could be matched with genes have been used, among the complete pathways and diseases chemical compounds and drugs also could be matched.

1.5 HUGE NAVIGATOR

A network called Human Genome Epidemiology Network (*HuGENet*) has constructed and maintained a database collecting publications of population-based epidemiological studies of human genes since 2001. Accordingly, *HuGE Navigator*

(Yu, Gwinn et al. 2008) has emerged as a knowledge base comprising a database integrating human genes, the human genome epidemiology associated with them and a number of tools to make the database easier to use for interdisciplinary researchers. Publications in the database are assigned to categories of study types such as observational studies or meta-analysis and of data type such as gene-disease association, gene-environment interaction or pharmacogenomics. Curators perform assignments weekly and add new entries as a new collection of publications entered into PubMed. Also, each publication is assigned a MeSH term (Medical Subject Headings), a hierarchical ontology by the National Library of medicine for indexing articles in *MEDLINE*, and to gene information from National Center for Bioinformatics (*NCBI*) *Entrez* gene database (Maglott, Ostell et al. 2011).

HuGE Navigator does not only include the database described above, it also includes some applications in its framework. They allow users to navigate and search the database in an integrated way. One of them is a search engine for finding published literature about human genome epidemiology such as genetic association studies, namely, *The HuGE Literature Finder* (Yu, Yesupriya et al. 2007). Another search engine is *The HuGE Investigator Browser*, developed for finding investigator networks for a given research interest. Huge Navigator also aids in following new trends in human genome epidemiology research. *HuGE Watch* (Yu, Wulf et al. 2008) has been in the list of those applications for this purpose.

The application set includes a tool called *Gene Prospector* (Yu, Wulf et al. 2008) for scientists who seek for candidate genes for an interested subject. Also all published Genome Wide Association Studies in GWAS catalogue (Hindorff, Sethupathy et al. 2009), curated by the National Human Genome Research Institute, can be queried in a robust way via *GWAS Integrator* (Yu, Gwinn et al. 2008). This bioinformatics tool also provides analytic functionalities for these studies. The data compilation is based on the GWAS Catalog (Hindorff, Sethupathy et al. 2009), *HapMap* (2003; 2004; 2005; Thorisson, Smith et al. 2005; Frazer, Ballinger et al. 2007; Altshuler, Gibbs et al. 2010), SNAP (Johnson, Handsaker et al. 2008). Interested GWAS study hits after some lookup can be converted to the *UCSC* browser (Kent, Sugnet et al. 2002) as query based custom tracks. Integration of SNPs

in close proximity and candidate genes from the *HUGE Navigator* to explore potential associations between GWAS hits and diseases/traits of interest is also possible. At the end the *HuGE Navigator* search results can be downloaded as a text file.

Most authors use historical or common names for annotating the genetic variants in the abstracts of their publications. This makes search criteria confusing for finding the studies relating a reference SNP number to some diseases or phenotypes. To match the *rs* numbers to the studies, *HuGE Navigator* provides a tool called *Variant Name Mapper* (Yu, Ned et al. 2009). This tool is a search engine that utilizes a database, mapping *rs* numbers to historical or common names of genetic variants. For validating genetic variations for health outcome predictions by calculating epidemiologic measures, *HuGE Risk Translator* also has been developed (Yu, Gwinn et al. 2008).

HuGE Navigator is an open source project and one can download presented data and tools from their website. The tool set includes also two online encyclopedias, *Phenopedia and Genopedia* (Yu, Clyne et al. 2010). They summarize the studies for gene-disease and gene phenotype associations respectively.

1.6 ENSEMBL PROJECT

Ensembl (Flicek, Amode et al. 2011) is the name of the project founded and maintained by collaborative efforts of two important institutions, EMBL-EBI and Welcome Trust Sanger Institute. Since 1999, before the accomplishment of the draft of Human genome, the aim of the project has been providing databases for vertebrates and software system automatically annotating the genomes in many ways through integration of other resources agreed. Altough the project has been initiated in 1999 the website has started to serve in July 2000. The assemblies and DNA sequences used in *Ensembl* gene builds have been provided by versatile global projects, each of which is documented in home pages of relevant species in www.ensembl.org.

Number of genomes handled in *Ensembl* project has been increasing. The amount of information provided is extensive (i.e., 56 species supported in *Ensembl*

build 59, completed in August 2010) data on human, mouse, rat and zebrafish are widely used.

Apart from the core sequence information of those 56 species and their annotations, *Ensembl* project also provides variation data, comparative genomics data, regulation data and Perl API (Stabenau, McVicker et al. 2004) for programmatic access. Among those, the ones, for which the most up-to-date developments have been reported, will be mentioned in this thesis.

The data stored in *Ensembl* is updated several times in year. All the softwares and data are freely available for download and installation.

1.6.1 Comperative Genomics

Ensembl project enlarges by the addition of new databases, especially the genome databases for the species genomic sequences of which are getting completed. Thus enormous computer power is needed to do genomic alignments for every update of *Ensembl* gene builds. *Ensembl* project has also solved this problem by creating an automatic pipe (Severin, Beal et al. 2010) for genomic alignments to determine homologues and orthologues genes and gene trees (Vilella, Severin et al. 2009).

1.7 RATIONALE AND AIMS

In the past, tissue specificity of microRNAs has been shown. Recent studies started to compile profiles and to perform meta-analyses. For example, Bargaje *et al*, have processed all datasets to find tissue specific and tissue invariant microRNA profiles (Bargaje, Hariharan et al. 2010). Furthermore some databases already house microarray expression data allowing for the presentation of such datasets for a queried microRNA, such as *miRGator* (Nam, Kim et al. 2008) and *MicroRNA.org* (Betel, Wilson et al. 2008), as mentioned earlier. Other databases exist that compile existing microRNA species from different organisms and report on their sequence and target specificity (Griffiths-Jones 2006; Sethupathy, Corda et al. 2006; Megraw, Sethupathy et al. 2007; Maselli, Di Bernardo et al. 2008; Wang 2008; Taccioli, Fabbri et al. 2009; Hsu, Lin et al. 2011; Yang, Li et al. 2011). However, there has not

been any microRNA database that incorporates expression data together with sequence data and allows multivariate visualization and expression enrichment/depletion analysis.

In the present study, I aimed to incorporate sequence and expression features of microRNAs together in a user friendly, modular and easily updatable manner within and among species. The importance of this aim comes from the fact that most of the microRNAs have been discovered by detecting conserved 3'UTR regions not only within species but also between species. This suggests that the presence of a seed sequence motif in a group of microRNAs may imply a common function since all of the microRNAs sharing the motif will target the similar mRNAs that bears the target sequence. This leads to the question of whether microRNAs having a particular seed sequence have similar expression patterns in terms of tissue or disease specificity.

Accordingly, the present thesis has focused on creating a tool for expression pattern analysis of a given set of microRNAs specified by their sequence motifs or tissue specificity. Furthermore, users might have their own datasets to compare with existing datasets increasing the need for user data upload facilities. For example, a study exemplifying this idea shows that AAGTGC motif is a stem cell specific seed motif. Furthermore, this set of microRNAs has been claimed as cancer discriminating microRNAs (Laurent, Chen et al. 2008). We can test this by:

A) Gathering and uploading different tissue and cancer microRNA datasets;
B) Selecting the group of microRNAs with the AAGTGC motif;
C) Visualizing the tissue and cancer specific expression profiles using multivariate techniques;
D) Defining the expression specificity by using an association index and;
E) Comparing expression profiles across different expression studies within and between species.

mESAdb then aims to provide an online tool by which the aims listed above (A-E) could be performed using an online tool and in an interactive and user-specified manner.

1.8 CONTRIBUTIONS

mESAdb contributes to the scientific community by permitting analysis of the relationship between expression patterns of microRNAs and their sequences via multivariate analysis techniques mentioned in the Materials and Methods section. One of the strengths of mESAdb originates from its use of R language for statistical calculations and visualization packages; this makes mESAdb modular and expandible. Others include the ability to select subsets of microRNAs for sequence and expression analysis via file upload, manual entry or through motif search options. This feature of mESAdb allows for integration of motif sequence with expression datasets. Other contributions by mESAdb can be summarized as:

1) Mining of default tissue-specific microRNA expression data sets across human and mouse and zebrafish;
2) Ability to upload any microRNA expression dataset in a .csv format and allow for automatic annotation based on *miRBase* entries;
3) Pair-wise multivariate analysis of expression data sets within and between taxa using MADE 4.0;
4) Application of phi-coefficient for enrichment analysis of microRNA expression for a given expression class and a set of microRNAs;
5) Comparison of a dataset with common motif sets in the seed regions of microRNAs;
6) Association of microRNA subsets with annotation databases, *HuGE Navigator*, while other microRNA databases focused *KEGG* and *GO*.

Expression pattern analysis and functional annotation for a single microRNA are also possible by the 'microRNA Search' module of mESAdb.

CHAPTER 2: METHODS

2.1 STATISTICAL METHODS USED IN MESADB

A unique feature of mESAdb is the on-the-fly utilization of various statistical methods on collected microRNA array data and other gene oriented data integrated from other sources by using the statistical environment R (R 2010). For multivariate analysis solutions the MADE4 (Culhane, Thioulouse et al. 2005) packcage from Bioconductor (Gentleman RC 2004) has been used. For example, 'co-inertia analysis' provided in MADE4 has been used for comparing two array datasets or comparing any dataset with motif distribution whereas 'correspondence analysis' has been used for representing the match between tissues and microRNAs in two-dimensional space.

Besides the R packages for specific statistical analyses, generic R functions e.g., for term enrichment, '*hyperp*', hyper geometric distribution function (Johnson, Kotz et al. 1992), have also been used.

A modified version of ϕ-coefficient (Guilford 1941), a basic example of item set analysis, has been used to assess the significance of expression values of the selected microRNAs.

2.1.1 PCA-based multivariate data analysis

Principal Components Analysis (PCA) (Pearson 1901; Hotelling 1933; Jollife 2002) is the representation of multivariate data using new set of axes, the number of which is much smaller than the interrelated variables such that new axis set could capture as much variation as possible, in the multivariate data. These new set of axes are called Principal Components (PCs) (Jollife 2002). Those new axes are uncorrelated and orthogonal to each other. Another property of these axes is that PCs are ordered according to the variance that they carry i.e., the first one caries the most of the variance and second one represents the second highest variance and so on (Jollife 2002).

For the first time Hilsenbeck et al, (1999) introduced the PCA method to

microarray data analysis. In this study they have found the genes, expression levels of which have been changed during the tamoxifen resistance acquisition in MCF7 cells (Hilsenbeck, Friedrichs et al. 1999). They have determined three components: 1-) Genes with average expression, 2-) Gene expression levels that differ between the estrogen stimulated cells and tamoxifen applied cells, 3-) and again gene expressions that can discriminate tamoxifen resistant and sensitive cells (Hilsenbeck, Friedrichs et al. 1999).

There are some other studies in which PCA is applied to analyze microarray data. The main focus of two of the studies was to catch the linear trends in microarray data in which variables had been accepted as conditions and the genes had been treated as observations (Raychaudhuri, Stuart et al. 2000; Crescenzi and Giuliani 2001). However, one of the following studies have applied PCA as correspondence analysis to microarray data to discover the association between samples and genes (Fellenberg, Hauser et al. 2001). Finally, the use of PCA arrived at cross platform microarray data analysis and comparison via co-inertia analysis (Culhane, Perriere et al. 2003). Culhane et al. (2005) have developed an R package that could cover those PCA applications mentioned above for all kinds of expression data (Culhane, Thioulouse et al. 2005).

In mESAdb, I have used the PCA technique indirectly by utilizing correspondence and co-inertia analysis as provided by *MADE4* package. Following sub sections those applications have been described.

2.1.2 Correspondence Analysis

PCA and Correspondence Analysis (CA) (Fellenberg, Hauser et al. 2001) are similar that both reduce the dimensions of a space in a data matrix. CA deals with two variables at the same time whereas the PCA deals with one. Also, via CA, it is possible to plot both genes, vectors of the condition space, and conditions, vectors of the gene space, onto the same space with reduced dimensionality. These projections accomplished by CA method aim to reveal the associations between the two variables (e.g., microRNAs and tissue types).

In the work of Fellenberg et al. 2001, (Fellenberg, Hauser et al. 2001) the algorithmic procedure for CA has been defined as follows and is summarized herein:

Let M be the K by L data matrix with N elements where K represents genes and L represents conditions in output of a microarray experiment. To formulate the correspondence analysis, for practical reasons, some concepts have been denoted as following:

$for 1 \le i \le K \; and \; 1 \le j \le L$

n_{i+} : $sum \; of \; the \; kth \; row$,

n_{+j} : $sum \; of \; the \; lth \; column$,

n_{++} : $sum \; of \; the \; all \; N \; elements \; in \; the \; matrix \; M$,

$c_j = n_{+j} / n_{++}$:$The \; mass \; of \; the \; lth \; column$,

$r_i = n_{i+} / n_{++}$:$The \; mass \; of \; the \; kth \; row$,

$P \; is \; the \; correspondence \; matrix \; with \; K \; rows \; and \; L \; columns \; where \; elements \; of \; it$:

$P_{ij} = n_{ij} / n_{++}$,

$The \; matrix \; S \; is \; a \; K \; by \; L \; matrix \; where \; s_{ij} = (P_{ij} - r_i c_j) / \sqrt{r_i c_{lj}}$

$Where \; S \; can \; be \; seen \; as \; the \; product \; of \; three \; matrices : S = U \lambda V^T$.

λ is the diagonal matrix carrying the singular values of the matrix S. Elements of this diagonal matrix should be ordered from the highest to the lowest since they are positively correlated with the co-variances captured. The new axes are the matrices U and V. Hence, the new coordinates of genes and the conditions are:

$g_{ik} = \lambda_k u_{ik} / \sqrt{r_i}$ and $f_{jk} = \lambda_k v_{jk} / \sqrt{c_j}$ respectively for k=1,…,L.

2.1.3 Co-inertia Analysis

Co-inertia analysis (Culhane, Perriere et al. 2003) is a mathematical method for capturing and determining the co-relationships in multivariate datasets. As in the CA and PCA, for CIA the principle is the same: finding two or three axes that maximize the variances of the points plotted on them. Again as mentioned in correspondence analysis the axes and Eigen values are ordered according to amount of variances that they carry or represent, such that the first axis is the one that carries the maximum variances of projected points and the second is so among the remaining axes orthogonal to the previous one and so on. Here some concepts again

have been denoted to explain the mathematical basis of CIA as explained in Culhane et al., 2003 (Culhane, Perriere et al. 2003):

From the notations of previous section those are handled;

$$R = [r_1, ..., r_K]$$
$$C = [c_1, ..., c_L]$$
$$X = [\frac{P}{r_i c_j} - 1]$$
$$D_r : Diagonal\ matrix\ with\ the\ values\ of\ R$$
$$D_c : Diagonal\ matrix\ with\ the\ values\ of\ C$$
$$D_{cx} : Diagonal\ matrix\ with\ the\ values\ of\ C\ derived\ from\ dataset\ X\ of\ type\ M$$
$$D_{cy} : Diagonal\ matrix\ with\ the\ values\ of\ C\ derived\ from\ dataset\ Y\ of\ type\ M$$
$$B = D_{cx}^{1/2} X D_r Y^T D_{cy}^{1/2}$$

Here, B is a K by K matrix but this time it is not correspondence matrix, it is correlation matrix. Again we can decompose it to its singular values where $B = U\lambda V^T$. Scaling by the c values each of which belongs to one dataset while determining the coordinates in new space would create two points for each gene one from each dataset. Hence the vectors seen on the microRNA plots after co-inertia analyses represent this case. In those plots, while the starting point of an arrow is belonging to the projection from the first dataset, the end point of it represents the projected coordinates from the second dataset.

2.1.4 ϕ-Coefficient

ϕ-coefficient is a statistical measure introduced by Karl Pearson (Cramér 1946). Its value represents the association between two binary variables, observations of which are held in a contingency table. Let the binary variables be x and y, then to contingency table will be:

Table 2.1.1: A sample contingency table of two binary variables, x and y.

	y=1	y=0	total
x=1	n_{11}	n_{10}	n_{1+}
x=0	n_{01}	n_{00}	n_{0+}
total	n_{+1}	n_{+0}	n

Then the ϕ coefficient will be $\Phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1+}n_{0+}n_{+1}n_{+0}}}$.

2.1.4.1 ϕ-coefficient based barplot:

ϕ-coefficient has been applied to barplots in mESAdb to show the association between selected microRNAs and each selected tissue (e.g., brain vs 10 other tissues). Let the expression data for the barplot be a matrix M of *m* rows and *n* columns, where the individual microRNAs are the rows, the classes (tissues) are the columns and each M*ij* is the expression level of microRNA *i* in condition *j*. Also, without loss of generality, let the microRNAs in the selected group (group *P*) be rows from 1 to *k* and the rest (group *N*) be from *k*+1 to *n*. For any class *j* of the *n* classes, the ϕ coefficient of the selected microRNAs can be calculated by ranking the centroid of these microRNAs. The centroid of the P-group microRNAs from 1 to *k* in class *j* is defined as:

$$C_{Pj} = \frac{1}{k}\sum_{i=1}^{k} M_{ij}$$

The centroid of group N is likewise defined as:

$$C_{Nj} = \frac{1}{n-k}\sum_{i=k+1}^{n} M_{ij}$$

The rank *R* of a centroid is the number of rows (microRNAs) in that column that have an expression higher than that centroid. For example, if in a case, C_{Pj} is less than 25 microRNA expression levels in class *j*, then R_{Pj} will be 25.

Since the ϕ coefficient is also affected by the absence of expression in other classes, as well as the presence of expression in the given class, a *virtual* column ¬*j* has been defineed such that it represents all the classes (columns) other than *j*. The

values of each row in ¬*j* are calculated as follows:

$$M_{i,\neg j} = \frac{1}{n-1}\left[\left(\sum_{k=1}^{n} M_{ik}\right) - M_{ij}\right]$$

Given these calculations, the ϕ coefficient is simply:

$$\phi_{Pj} = \frac{R_{Nj}R_{P\neg j} - R_{Pj}R_{N\neg j}}{\sqrt{R_{Pj}R_{Nj}R_{P\neg j}R_{N\neg j}}}$$

The ϕ coefficient varies between 1.0 and -1.0, positive values showing positive differential expression (enrichment) for the microRNAs in group *P* in class *j* and negative values showing negative differential expression (depletion). Values of ϕ at or around zero denote the independence of group P expression from class *j*.

The ϕ coefficient is related to the χ^2 statistic given the population. Since here two classes (*j* and ¬*j*), each with *m* microRNAs (rows) have been concerned, the population is 2*m*. Hence, the χ^2 statistic becomes:

$$\chi^2_{Pj} = 2m\left(\phi_{Pj}\right)^2$$

This can in turn be used to test significance and calculate a *p* value using Pearson's χ^2-test.

2.1.5 K-Means Clustering

It is a widely used clustering algorithm that clusters M points of N dimensions into desired K clusters, where K<M. So the inputs of the algorithm are an M by N matrix and a K centers with N dimensions (Hartigan and Wong 1979). The cluster assignment criteria is to minimize the sum of squares in each clusters (Hartigan and Wong 1979).

2.2 DATABASE DESIGN

mESAdb enables access and retrieval of microRNAs with specified motifs to associate and analyze them functionally as well as based on expression profiles (Figure 2.2.1). An initial version of this work was presented in abstract form in BioSysBio 2007: Systems Biology, Bioinformatics, Synthetic Biology (Kaya, Karakulah et al. 2007).

Figure 2.2.1: Screenshot of the mESAdb main page. The modules, 'motif-expression', 'expression-expression', 'motif-function' and 'microRNA search', are shown.

2.2.1 Data collection and storage

Data used in mESAdb are obtained periodically from multiple sources and processed for integration into the underlying MySQL database using a series of routines which download, parse and integrate these data from relevant sources (*Ensembl, miRBase, microCosm, HuGE,* KEGG and *GO*) either directly or through the Biomart integration service (Figure 2.2.1and Figure 2.2.2) (Ashburner, Ball et al. 2000; Kanehisa and Goto 2000; Durinck, Moreau et al. 2005; Griffiths-Jones, Saini et al. 2008; Yu, Gwinn et al. 2008).

The mentioned integration of different databases has been accomplished by a python script (Rossum May 1995). To get the mature microRNA names and their sequences for four species, *miRBase* (Ambros, Bartel et al. 2003; Griffiths-Jones 2004; Griffiths-Jones, Grocock et al. 2006; Griffiths-Jones, Saini et al. 2008;

Kozomara and Griffiths-Jones 2011) has been used by the script. First, all mature names and their sequences are downloaded from this database. Then a unique union of mature microRNA name list has been constructed. This part of the script ends up with the MySQL table which is constructed from mature microRNA names and their sequences for four species. The column names of the table are ‘mirna’ and short names of considered species, namely, ‘hsa’, ‘mmu’, ‘cel’ and ‘dre’. This table establishes the core of the mESAdb. It enables selection of mature microRNAs with their sequence properties when required. Another utility of this table is up-to-date annotation of the probe sequences in any high-throughput expression dataset to be used in mESAdb.

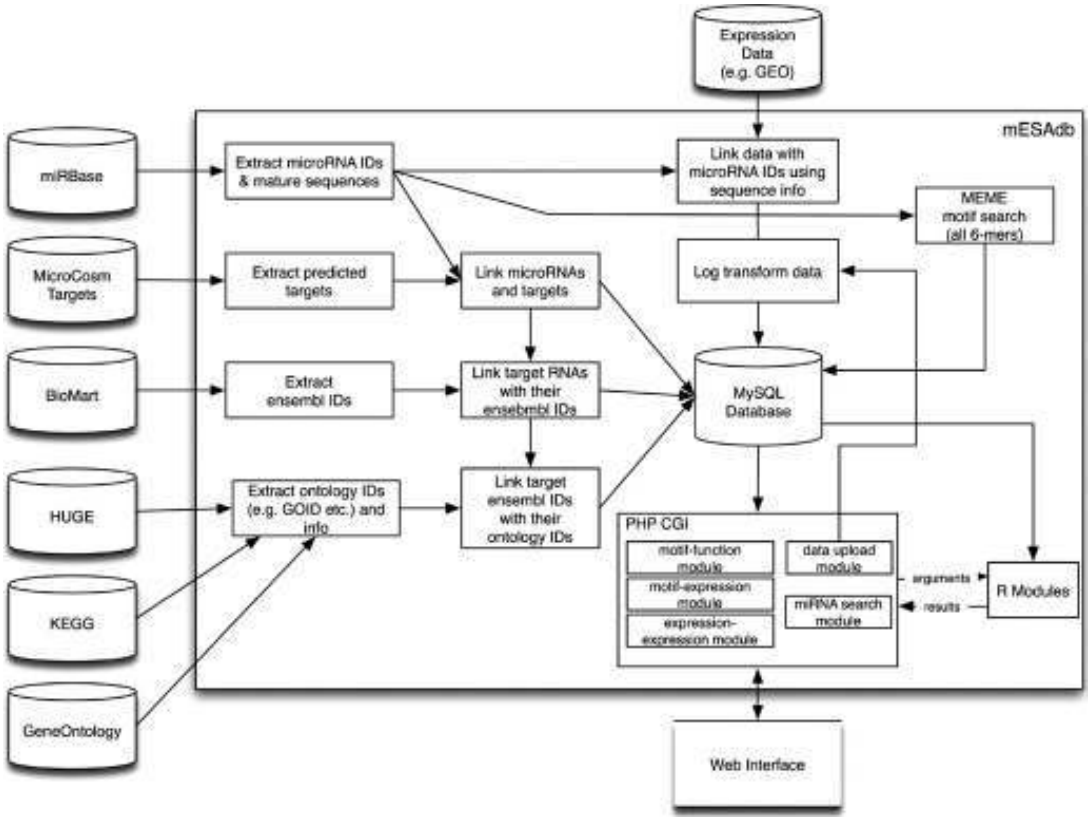


Figure 2.2.2: Workflow diagram of mESAdb. MESAdb combines data from a variety of external data sources. For example, microRNA mature sequences and IDs are retrieved from miRBase and matched with microRNA data sets (e.g. from GEO). microRNA sequences are processed by the MEME motif finder for conserved motifs. The microRNA targets are fetched from EBI's MicroCosm Targets for each species; BioMart is used to get the ENSEMBL Gene IDs of the targets' transcript IDs. These ENSEMBL Gene IDs are then linked to HUGE Navigator Disease IDs, KEGG Pathway IDs and GO IDs. A user-friendly interface has been developed in PHP for accessing data in the system and allowing versatile analysis via various R scripts (http://php.net; http://www.r-project.org./;http://www.mysql.com/).

In the current version of mESAdb, mature microRNA names and sequences were downloaded from *miRBase* Release 15 (Griffiths-Jones, Grocock et al. 2006).

MicroRNA microarray experiment data sets for human, mouse and zebrafish, primarily focusing on expression from different tissues and developmental stages were stored separately as default data sets (Barad, Meiri et al. 2004; Thomson, Parker et al. 2004; Baskerville and Bartel 2005; Beuvink, Kolb et al. 2007; Ach, Wang et al. 2008; Navon, Wang et al. 2009; Meiri, Levy et al. 2010) (

Table 2.2.1). Tables containing the normalized expression values were

associated with sequence data linked with the corresponding *miRBase* names for these microRNAs (Figure 2.2.2). Where available, the probe sequences printed on microarrays that match exactly with the species-specific reverse complementary sequences in *miRBase* were included resulting in increased stringency; thus the number of microRNAs from each microarray study incorporated into mESAdb might be smaller than that reported in the original study. Expression data were logarithmically transformed where necessary, and quantile normalized (Bolstad, Irizarry et al. 2003).

To link sequence and expression properties with functional information, the predicted human targets were retrieved from *MicroCosm* Targets (Figure 2.2.2) (Griffiths-Jones, Saini et al. 2008). *MicroCosm* microRNA-target gene matching files have been used to construct the species specific microRNA-target tables using the same python script.

These targets were further processed on the R environment (Version 2.11.1) (R 2010); transcript IDs were matched with *Ensembl* Gene IDs (Ensembl Relese 59) using the package *biomaRt* (Durinck, Moreau et al. 2005). Only a single *Ensembl* ID was retrieved for each target gene with multiple transcript entries. Species-specific microRNAs were paired with target gene IDs associated with ontology terms and these matched pairs were stored in mESAdb's underlying DBMS (Figure 2.2.2; MySQL). *KEGG* and Gene Ontology terms associated with microRNA targets were extracted and matched with the corresponding microRNA ID (Ashburner, Ball et al. 2000; Kanehisa and Goto 2000). The disease terms associated with microRNA targets were obtained from the *phenopedia* view of *HuGE Navigator*; these terms were parsed and matched with microRNA targets and stored in the MySQL table underlying mESAdb. *HuGE* and *KEGG* databases use *Entrez* gene IDs. To convert those *Entrez* gene IDs to target *Ensembl* IDs, the script uses the *BioMart* database. The database also has been used for target gene-gene ontology term matching in the species specific MySQL tables. Target and associated terms are updated as the script is called either by hand or periodically (Figure 2.2.2).

Table 2.2.1: Default data sets provided in mESAdb

Name	Species	GSE NO.	Platform	Pubmed ID	Tissues
Meiri et al., 2010	*Homo sapiens*	GSE20414	GPL10067	20483914	Ly, K, En, Lu, Bl, B, H, Li
Navon et al., 2005	*Homo sapiens*	GSE14985	GPL8227	19946373	Br, Pr, Ly, O, Co, Li, Te, Lu
Ach et al., 2008	*Homo sapiens*	GSE11806	GPL6955	16783629	Pl, B, Br, H, Th, Li, O, SM,Te
Barad et al., 2004	*Homo sapiens*	-	MOE-ER array	15574827	HeLa, B, Li, Th, Te, Pl
Baskerville & Bartel 2005	*Homo sapiens*	-	MWG biotech	15701730	BM, B, H, K, Li, Lu, Pa, Pr, SM, Sp, Th, FC, Ly, Co, HelaS3, Ce, Bl, Te, A, U, Br, F, Sl, Pl, O
Wienholds et al., 2005	*Danio rerio*	GSE2628	GPL2023	15919954	B, Ey, SM, H, Gi, Fi, Sk, G, Li, Te, O
Thomson et al., 2004	*Mus musculus*	GSE1635	GPL1391	15782152	Li, K, Lu, O, H, B, Th, ES, EBD3, EBD28, E7, E11, E15, E17
Beuvink et al., 2006	*Mus musculus*	-	Custom	17355992	HeLa, B, Li, Lu, Sl, SM, H, K, Sp

Tissue Abreviations. Fallopian tube (F), Uterus (U), Lymph node (Ly), Placenta (Pl), Breast(Br), Pancreas (Pa), Liver (Li), Brain (B), Thymus (Th), Heart (H), Lungs (Lu), Spleen (Sp), Testicle (Te), Ovary (O), Kidney (K), Skeletal muscle (SM), Small intestine (SI), Colon (Co), Prostate (Pr), Bladder (Bl), Cervix (Ce), Adrenal gland (A), Stomach (St), Bone Marrow (BM), Frontal Cortex (FC), Eye (Ey), Gill (Gi), Fin (Fi), Skin (Sk), Gut (G), HeLa Cells (HeLa), HeLa S3 (HL3S), Endometrium (En).

2.2.2 User-specified expression data set management

mESAdb incorporates a tool for the upload of user-specified expression data sets provided as comma separated files (Figure 2.2.3). The user is free to add, view and remove expression data sets having expression data for arbitrary numbers of microRNAs against arbitrary number of expression classes (e.g. tissues, developmental stages, disease states). The format for the input file is straightforward: a comma delimited file with the first row giving the names of the classes, the subsequent lines of the file each beginning with the name of the microRNA (e.g. the *miRBase* ID) and, optionally, the probe sequence, followed by the measured expression for each of the classes given in the header line. The file uploaded is preprocessed line-by-line; for each, if the reverse complement of the probe sequence given for that line contains the mature sequence for the corresponding microRNA as given in the latest *miRBase*, the line is verified. For lines that cannot be thus verified, the system searches for a match in the *miRBase* sequences for the relevant species. If found, the microRNA is renamed to its *miRBase* standard name, if not, the line is discarded. Subsequently, the lines for duplicate microRNAs are averaged. The upload module generates a downlodable text file listing the actions performed while parsing and processing the *csv* file. mESAdb uses nomenclature by *miRBase* for cross taxa comparisons performed in 'expression–expression' module where microRNAs with the same name from two different species are matched. Most microRNAs carrying the same name exhibit high sequence similarity across species

while ~5% are relatively divergent.

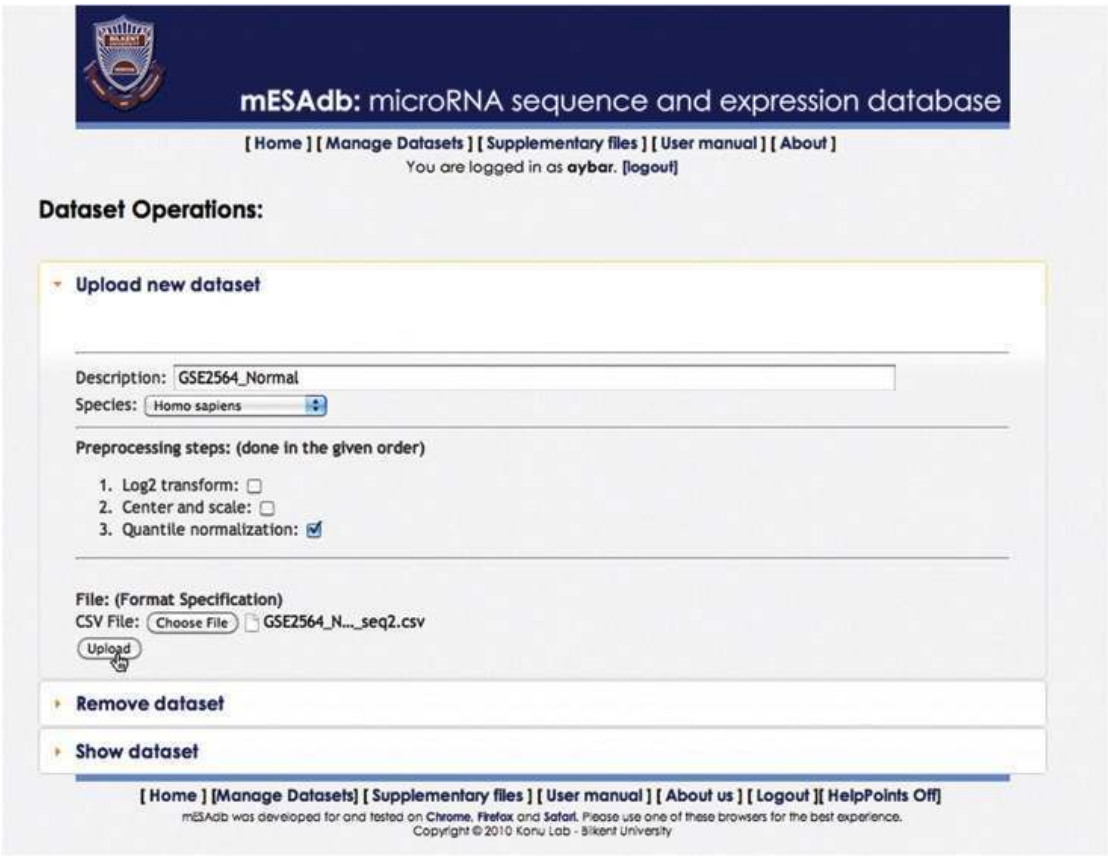


Figure 2.2.3: Screenshot of the data upload module. User can select from species and microarray normalization options and then browse to upload a data set.

The data upload utility also warns users in such cases. The module also provides utilities to log-transform, center and scale or quantile normalize the expression data upon verification by *miRBase* (Figure 2.2.3). A user-uploaded dataset is tied to the specific user account that creates it and may have been retrieved from another source or may be the product of the users' own research. The designed system protects privacy of proprietary data by keeping uploaded sets visible only to the account that owns it and no data is retained once a user removes a data set.

An exemplary data set was provided in the current version of the mESAdb (i.e. GSE2564NORMAL_seq.csv; mESAdb supporting material; http://konulab.fen.bilkent.edu.tr/mirna/supplementary_files.php) (Lu, Getz et al. 2005). Accordingly, GSE2564 expression series matrix was downloaded from GEO

(Barrett, Troup et al. 2009). This data set includes normal tissues from stomach (n=5), colon (n=5), pancreas (n=1), liver (n=3), kidney (n=3), bladder (n=2), prostate (n=8), uterus (n=9), lung (n=4), breast (n=3) and brain (n=2), together with cancer samples for different tissues. For the example used herein, only the expression data on the normal tissues were obtained; linked with microRNA annotations and probe sequences in the GPL1986 description file; and a comma separated file was formed for upload. An account of processing of the microRNAs in the .csv file was generated by mESAdb. The data set, called GSE2564_normal, could be uploaded using the 'Manage Datasets' facility of mESAdb (Figure 2.2.3) and compared with the existing data sets listed in

Table 2.2.1.

2.3 INTEGRATION OF R PACKAGES

mESAdb uses a hybrid of PHP and R as a computational environment. The basic operations and the web interface elements are coded in PHP whereas more significant statistical analyses such as co-inertia, correspondence, the ϕ coefficient and p-value calculation and hyper geometric distribution tests for *HuGE*, *KEGG* and *GO* terms are performed in R (Figure 2.2.2). The web interface has been made as responsive and user-friendly as possible with the addition of dynamic elements created with *JavaScript* and the *JQuery UI* (http://jqueryui.com/) library.

The communication between the PHP and R environments is performed using the common underlying MySQL database and Unix based operating system pipes. Briefly, a PHP script creates a child R process to which command line arguments are passed onto. The R process uses this information to retrieve the relevant information from the MySQL database and subsequently prepares the output (e.g. graphics; the bar plot, correspondence plots) which it passes onto the calling PHP script to display on the page. If the output is mostly textual (e.g. tabular data), it is passed on the output stream of the R program via PHP serialization package. If it is a larger result like an image, the R program saves it under a predetermined, random and client specific file name in a temporary location, which the PHP script retrieves from once the child R process is finished. This two-way communication between the PHP code

and its R child processes has been implemented as a simple but effective API, which allows new R scripts to be easily integrated into the mESAdb tool as needed.

2.4 MESADB MODULES

2.4.1 Motif Expression

mESAdb has a motif selection tool with a pulldown menu in which users might select from different options to group retrieved microRNAs with a given motif, i.e. dinucleotide motifs or motifs up to 6 nucleotide long using the IUPAC code (Panico, Powell et al. 1993). It is also possible to upload user-specified microRNA lists. 'Motif-expression' module integrates the motif selection tool with default microarray data sets found in mESAdb as well as those uploaded by the user (Figure 2.2.1, Figure 2.2.3 and

Table 2.2.1).

Accordingly, mESAdb provides a platform for visualization of microRNA expression in humans, mouse and zebrafish. Once a microRNA list is selected, expression of this set of microRNAs can be investigated using three different analysis options: 'expression analysis', 'correspondence analysis' and 'co-intertia analysis' (Culhane, Thioulouse et al. 2005).

The 'expression analysis' option enables the user to compare, using bar plots, the amount of mean expression of the selected microRNAs with those of the remaining microRNAs across the studied expression classes, i.e. tissues or developmental stages. Expression data (

Table 2.2.1) for the selected microRNAs and those for the unselected microRNAs are extracted from the quantile normalized log transformed expression tables that have been generated by the mESAdb data upload facility. The class (e.g. tissue) specific mean values for the selected and unselected microRNAs then are plotted separately for each column of the data set (e.g. each tissue) using a bar plot. Bars are color coded by the value of the ϕ-coefficient to assess the association of the selected microRNAs with the tissue in consideration (also called the Yule-ϕ) (Guilford 1941). A dynamic hover feature has been implemented for user to see

exact information about each column by hovering the mouse pointer over it in the barplot. Expression data sets are accessible in the html format, and the $\chi 2$ and P-values for the ϕ-coefficient also are generated. Help boxes are made available for data plots and analysis tools.

mESAdb performs multivariate analysis of expression using the R package MADE4 customized for visualization and analysis in mESAdb (Culhane, Thioulouse et al. 2005). 'Correspondence analysis' of the selected set of microRNAs produce three graphical outputs, allowing for visualization of the expression patterns across classes (e.g. tissues), or microRNAs, or both the classes and microRNAs. 'Co-inertia analysis' (Culhane, Thioulouse et al. 2005) of the selected set of microRNAs helps visualize the similarities between microRNA expression and occurrence of common 6-mer MEME (Bailey and Elkan 1994) motifs found among the microRNA sequences housed in mESAdb. Users can link from a motif to back to the 'expression analysis' module explained above to visualize the expression data as bar plots per expression class (e.g. tissue), of the group of microRNAs used in the co-inertia plot containing the specified motif . MEME motif outputs we generated for the human, mouse and zebrafish microRNAs can be accessed from the supporting material (http://konulab.fen.bilkent.edu.tr/mirna/supplementaryfiles.php) found at the mESAdb.

2.4.2 Expression-expression

This module provides a tool for meta-analysis of microRNA expression data sets. Selected sets of microRNAs can be investigated with regard to the datasets listed in

Table 2.2.1 in a pairwise fashion; other user-defined data sets can be uploaded and analyzed as well (Figure 2.2.1). 'Expression–expression module' outputs co-inertia graphics for (i) classes (e.g. tissues) and (ii) microRNAs, and also a heatmap of both data sets using customized MADE4 (Culhane, Thioulouse et al. 2005) and heatplus (http://bioconductor.org/packages/2.6/bioc/vignettes/Heatplus/inst/doc/Heatplus.pdf) packages in R (www.bioconductor.org). The output has been customized for better

visualization; and the degree of association, indicated by the RV coefficient (Robert and Escoufier 1976; Culhane, Thioulouse et al. 2005) between two different microarray data sets also is provided. A high RV score suggests better correlation among data sets. For the microRNA oriented co-inertia graph, several utilities are provided in order to facilitate the visualization of potentially high numbers of data points. It is possible to visualize the microRNA data points with or without labels on the co-inertia graph. The co-inertia tool also provides an automatic clustering of the microRNAs based on the similarity of their expressions in both data sets using k-means clustering (Hartigan and Wong 1979); the default clustering displayed is the clustering with the maximum silhouette coefficient (Lovmar, Ahlford et al. 2005). Since k-means clustering is not deterministic, for each k-value the module performs 20 runs of the algorithm and the best clustering for each k is selected using highest silhouette. The clustering with the overall best silhouette is displayed by default. The user can manually set a cluster number between 2 and 10 clusters (i.e. $2 \leq k \leq 10$) if desired. These clusters can further be investigated to visualize the expression profiles for the given datasets using expression bar plots of in-cluster and out-of-cluster microRNAs, by clicking on the cluster centroids.

2.4.3 Motif-function

This function may be useful for functional analysis of, for example, a set of differentially expressed microRNAs (Figure 2.2.1 and Figure 2.2.2). In the present study, information from *HuGE Navigator*, in addition to *GO* and *KEGG* databases can be associated with the selected microRNAs (Ashburner, Ball et al. 2000; Kanehisa and Goto 2000; Yu, Gwinn et al. 2008). For any selected subset of microRNAs, mESAdb then can be used to retrieve the mappings of the selected functional terms, with the targets of these microRNAs and subsequently to calculate a probability value based on the hypergeometric distribution (Kachitvichyanukul and Schmeiser 1985). Functional and expression correlates of a single microRNA can be assessed using this module to enable a quick search involving multiple modules of mESAdb (Figure 2.2.1). Terms from *GO*, *HuGE*, *KEGG* and target genes associated with the given microRNA can be extracted; and the observed and expected counts as

well as hypergeometric P-values can be downloaded. Expression profile of the selected microRNA also can be visualized using the aforementioned bar plots and downloaded as .txt files.

2.4.4 microRNA search module

This module is a partial combination of the other three modules for a single microRNA of interest. The main objective of this module is to associate a single microRNA with target information as well as with expression profiles. One can also look up relevant disease relations through *HuGE phenopedia* and *genopedia*, *GO* terms, *KEGG* pathways for that microRNA (Figure 2.4.1).

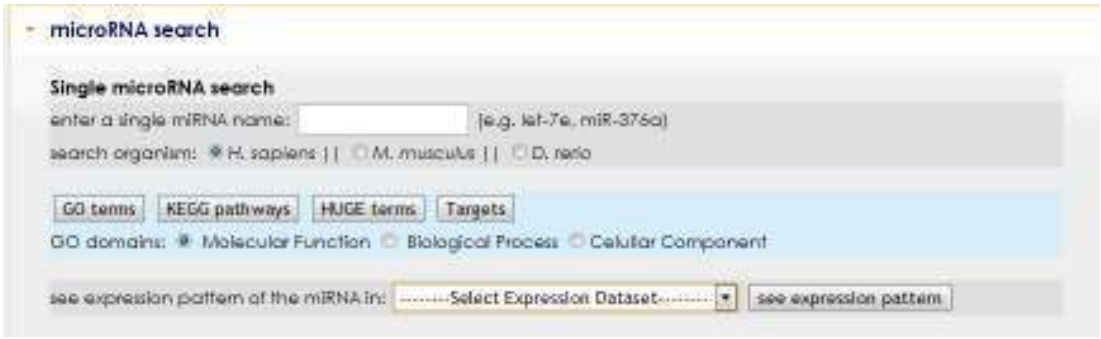


Figure 2.4.1: Snapshot of microRNA search module. In this module, one can apply every analysis function, that are applicable in mESAdb for a single microRNA.

2.4.5 Data processing for default expression datasets

2.4.5.1 Ach et al., (2008) (Ach, Wang et al. 2008):

This dataset has been downloaded from *NCBI*'s *GEO* (Edgar, Domrachev et al. 2002; Barrett, Troup et al. 2011) web site through files in SOFT format (Simple Omnibus Format in Text) linked with the GSE11806 accession number and read locally by *Bioconductor*'s *GEOquery* (Sean and Meltzer 2007) package. The data matrix of hybridization intensities and probeset information of the array have been fetched from it. The values in the data matrix have been log transformed first and then the arithmetic means of the replicate columns have been taken. The column

names have been replaced by agreed upon tissue abbreviation codes. Then the data was turned into a data frame satisfying our data upload format and written to a .csv file. Since it has no sequence information the column specified as "sequence" field in upload data format has been left empty. After all, the data was uploaded to the database as a human dataset after checking the quantile normalization (Bolstad, Irizarry et al. 2003) option.

2.4.5.2 Barad et al., (2004) (Barad, Meiri et al. 2004):

This dataset comes with an annotated and non log transformed background corrected signal intensity matrix; the annotation table includes sense probe sequence information, which was first converted to antisense, and then matched with microRNA names and assigned to each row of the expression matrix before uploading. When using upload module, data was log transformed and quantile normalized (Bolstad, Irizarry et al. 2003).

2.4.5.3 Baskerville & Bartel (2005) (Baskerville and Bartel 2005):

This dataset comes as a signal intensity matrix in which the values below the background intensity have been replaced with the background value. The sequence annotations of miRNAs have been also provided. The same procedure as in Barad data was followed.

2.4.5.4 Beuvink et al., (2007) (Beuvink, Kolb et al. 2007):

This dataset normally includes raw signal intensity matrix neither normalized nor background corrected. Sequence annotation was provided and used to update microRNA names while being uploaded. The data were log transformed and quantile normalized.

2.4.5.5 Meiri et al., (2010) (Meiri, Levy et al. 2010):

This dataset was downloaded from the *NCBI*'s *GEO* (GSE20414). Only the array samples from the normal tissues were selected. The dataset given in the SOFT formatted file type had already been normalized through best fit to a reference array and log transformed at base 2. The GPL contains sequence information, thus during the preparation for upload this has been considered. Data was quantile normalized

(Bolstad, Irizarry et al. 2003) before upload.

2.4.5.6 Navon et al., (2009) (Navon, Wang et al. 2009):

This dataset available in *NCBI*'s *GEO* (Edgar, Domrachev et al. 2002; Barrett, Troup et al. 2011) (GSE14985) is composed of array samples derived from normal and tumor tissues from 14 different patients. The data had been pre-processed by replacing values smaller than the background with the background value. By using R and *GEOquery* package (Sean and Meltzer 2007), this dataset has been processed further by log transforming the dataset before selecting normal tissue samples only and taking the mean of samples from the same tissues. Data was quantile normalized (Bolstad, Irizarry et al. 2003) before upload.

2.4.5.7 Thomson et al., (2004) (Thomson, Parker et al. 2004):

The dataset includes array samples derived from either adult mice or mice embryos. Embryonic cell types were separated from the ones of adult origin. The dataset comes as log transformed, background corrected and median centered. Only additions to these preprocessing events were that replicated samples were averaged, and sequence information was added to the data frame. Before upload, the data was also quantile normalized (Bolstad, Irizarry et al. 2003).

2.4.5.8 Wienholds et al., (2005) (Wienholds, Kloosterman et al. 2005):

This zebrafish dataset was found normalized and mean centered in *NCBI*'s *GEO* (GSE2628) but not log transformed. SOFT formatted files were downloaded, the expression matrix was log transformed after adding a fixed value to the whole data matrix making the minimum value in the matrix, 1. Next, the sequence information from platform data has been integrated to the matrix and data was quantile normalized before upload.

CHAPTER 3:RESULTS

mESAdb is a highly interactive and flexible database with an ability to analyze and visualize selected expression profiles for a given subset of microRNAs in a multivariate manner using correspondence and co-inertia analyses. One can also study a single microRNA of interest using bar plots associated with a gene expression enrichment index, based on the ϕ-coefficient. This index provides a significance value for the relative enrichment of a microRNA(s) in a particular class with respect to others. Furthermore, the user can obtain information about the functional enrichment of a microRNA or a group of microRNAs using different databases, including *GO*, *KEGG* and *HuGE* Navigator.

3.1 ADDING NEW DATASETS TO MESADB

The default expression data sets currently focus on tissue- and stage-specificity; however, users can add any microarray data containing other types of expression classes, e.g. cancer versus normal, treatment versus control (Figure 3.1.2) (please see Methods section 3.2.2 for another example). This allows for great flexibility in analyzing one's own research data.

For example, to demonstrate the data upload utility, the dataset with GEO accession number GSE10694 (Li, Xie et al. 2008) has been used. The dataset is already normalized with local background subtraction and quantile normalization. This dataset, however, has no sequence information in its platform data, GPL6542. The arrays have been prepared from liver cancer tissues of 78 different patients with hepatocellular carcinoma, corresponding neighboring non-cancerous tissues of the same cases and 10 normal liver tissues.

The typical *csv* file format, where each entry is comma separated and each line has the same number of entries, was prepared for this dataset to upload it to mESAdb. First line of the file is header line. The first two columns, "mirna" and "sequence", are fixed and represent the mature microRNA name and probe sequence respectively. The remaining columns represent the arrays containing expression values. They are re-named such that hcc sample taken from patient number x has

been named as px_hcc, and normal liver sample taken from the same patient has been named as pxn where n (tissue from non-cancer individuals) is between 1 and 10. Other lines contain the corresponding values where sequence field is empty since there is no information coming from the platform data. The sample file prepared can be downloaded from: http://konulab.fen.bilkent.edu.tr/mirna/hcc.csv. After the upload, a snapshot of the dataset view is given below (Figure 3.1.1).

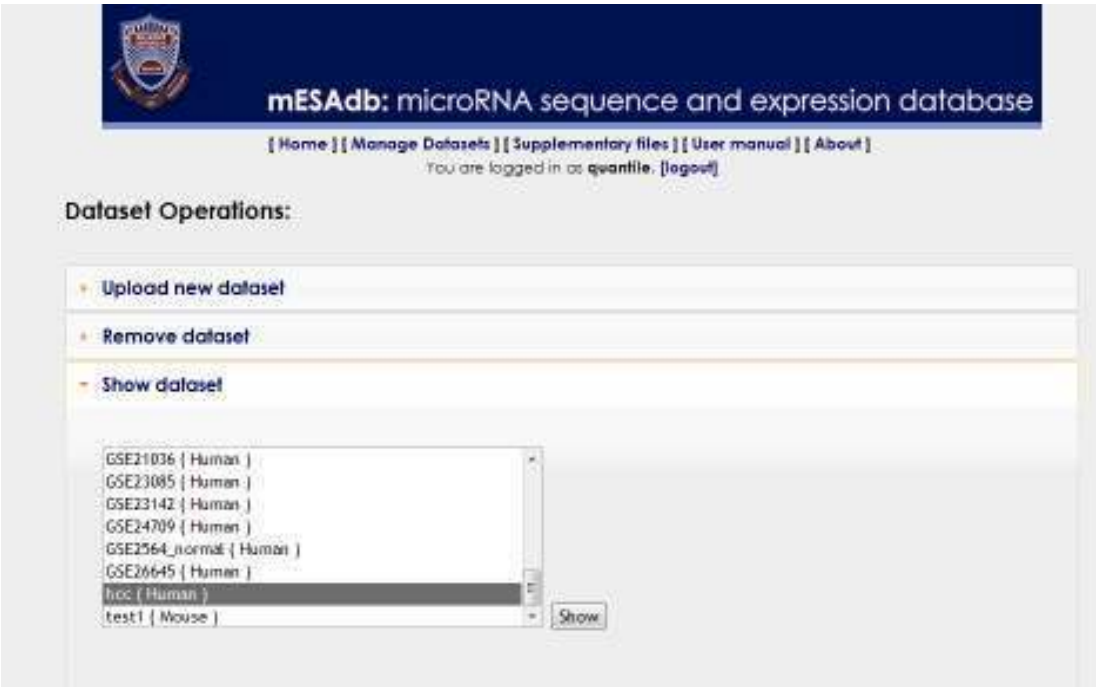


Figure 3.1.1: GSE10964 has been added to the database with the name 'hcc'.

To demonstrate the utility of this dataset in mESAdb, a microRNA seed sequence motif associated with embryonic stem cell enrichment, i.e., AAGTGC, was searched within this '*hcc*' Human dataset using the motif-expression motif (Laurent, Chen et al. 2008). Correspondence analysis visualization was performed showing that the microRNAs with the given motif (i.e., *miR-106a, miR-106b, miR-20a, miR-20b, miR-93, miR-520f, and miR-520b*; Table 3.1.1) successfully separated the tumor samples from the adjacent normal and normal samples (Figure 3.1.2). It is apparent from the correspondence analysis which microRNAs separate the hcc samples from the normal and vice versa. In fact, the expression of the *miR-520f* and *miR-520b* diverges from the rest and they are relatively more expressed in normal tissues where

the rest of the microRNAs are overexpressed in hcc samples (Figure 3.1.3). The findings suggest that stem cell signature might be an important factor in diagnosis of the hcc.

Table 3.1.1: The list of microRNAs that contain the AAGTGC motif particularly specific to stem cell populations (Laurent, Chen et al. 2008).

microRNA	Sequence
miR-106a	AA**AAGTGC**TTACAGTGCAGGTAG
miR-106b	TA**AAGTGC**TGACAGTGCAGAT
miR-20a	TA**AAGTGC**TTATAGTGCAGGTAG
miR-20b	CA**AAGTGC**TCATAGTGCAGGTAG
miR-520b	A**AAGTGC**TTCCTTTTAGAGGG
miR-520f	**AAGTGC**TTCCTTTTAGAGGGTT
miR-93	CA**AAGTGC**TGTTCGTGCAGGTAG

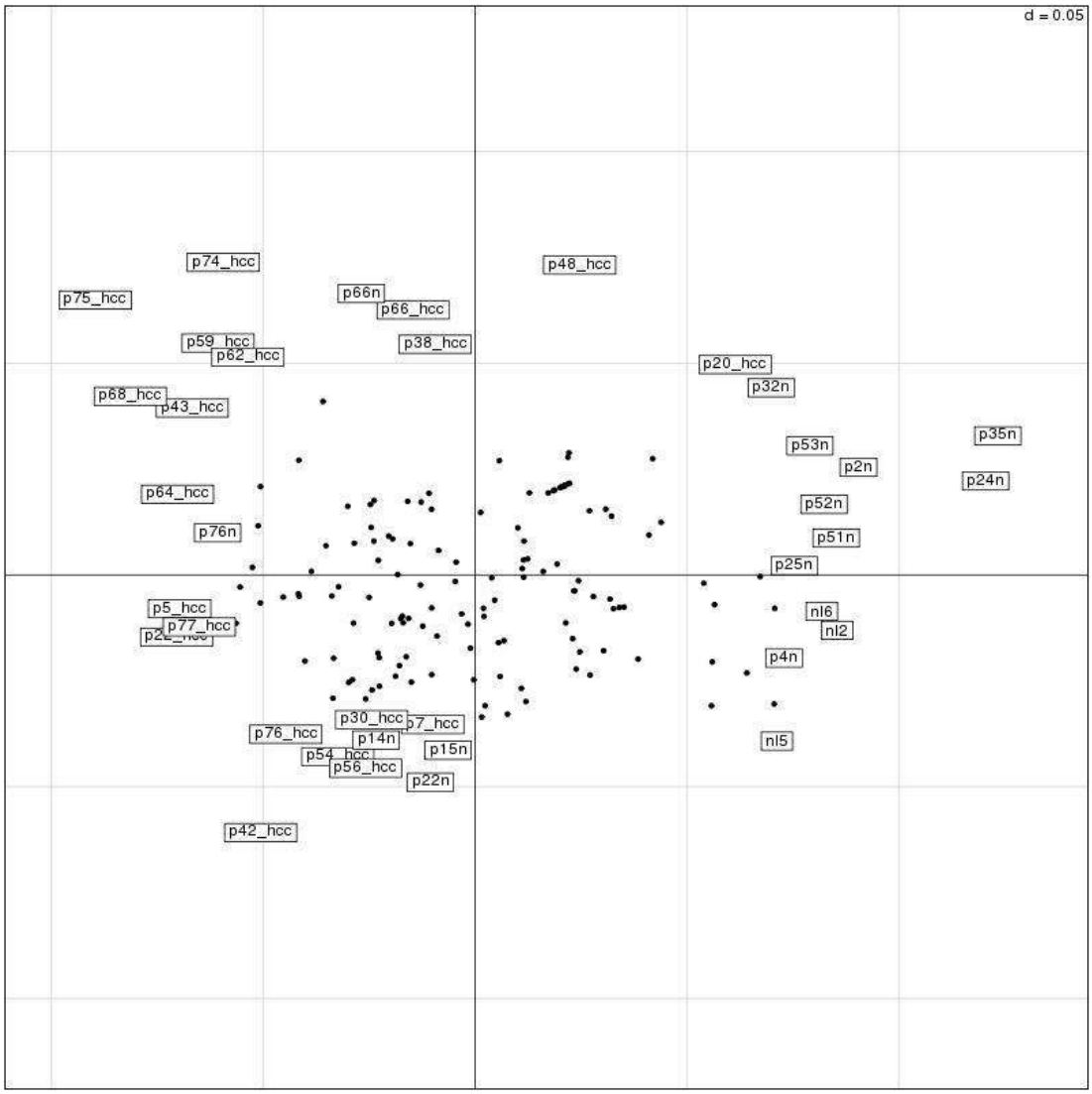


Figure 3.1.2: Plot of samples after the correspondence analysis of the dataset GSE10964 with microRNAs having 'AAGTGC' seed motif.

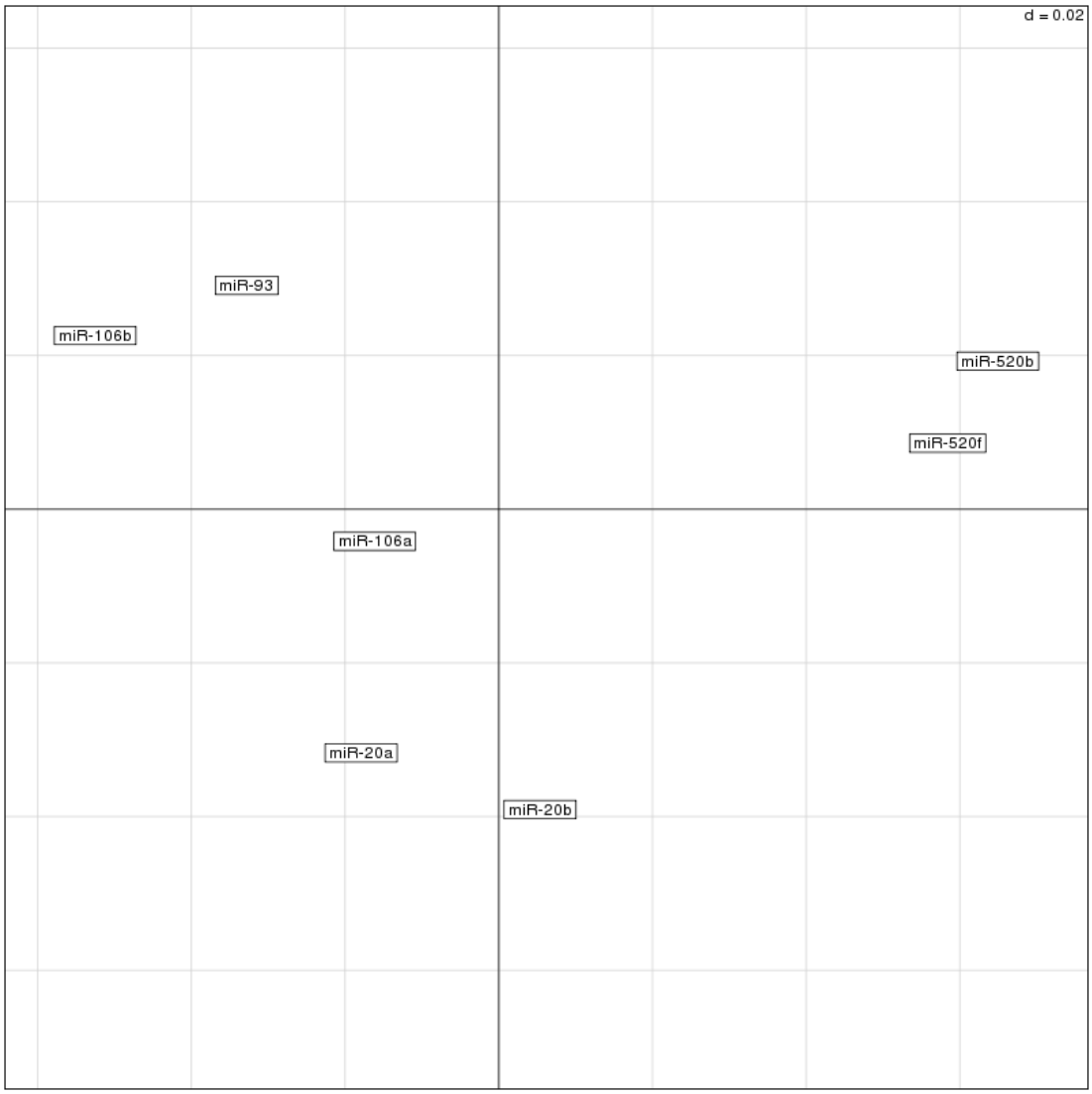


Figure 3.1.3: Plot of the microRNAs having 'AAGTGC' seed motif after the correspondence analysis of the dataset GSE10964.

3.2 COMPARISON OF DATASETS ACROSS TAXA FOR A GIVEN SET OF MICRORNAS

As an example, it has been demonstrated that the user can compare two data sets with respect to a list of microRNA clusters that are common to both mice and humans. Using the 'expression–expression' module of mESAdb, a human (Meiri, Levy et al. 2010) and a mouse (Thomson, Parker et al. 2004) data set have been chosen (

Table 2.2.1) and then a microRNA list has been uploaded (Table 3.2.1).The list includes *let-7a-i, miR-130a-b, miR-15a-b, miR-181a-b, miR-200a-b, miR-23a-b, miR-26a-b, miR-29a-c, miR-30a-d* and *miR-99a-b* clusters). Then the coinertia analysis was performed using only the tissues common to both data sets, namely, brain (B), liver (Li), lung (Lu), kidney (K) and heart (H) (Figure 3.2.1). Using coinertia analysis, mESAdb allows for comparison and visualization of two expression data sets by plotting them side by side in terms of the expression of selected microRNAs for the given tissues.

Table 3.2.1: Mature sequences of microRNAs that are used for co-ineria analysis between Meiri et al., 2010 and Thomson et al., 2004.

miRBase ID	Marture sequence in *Homo sapiens*	Marture sequence in *Mus musculus*
let-7a	TGAGGTAGTAGGTTGTATAGTT	TGAGGTAGTAGGTTGTATAGTT
let-7b	TGAGGTAGTAGGTTGTGTGGTT	TGAGGTAGTAGGTTGTGTGGTT
let-7c	TGAGGTAGTAGGTTGTATGGTT	TGAGGTAGTAGGTTGTATGGTT
let-7d	AGAGGTAGTAGGTTGCATAGTT	AGAGGTAGTAGGTTGCATAGTT
let-7e	TGAGGTAGGAGGTTGTATAGTT	TGAGGTAGGAGGTTGTATAGTT
let-7f	TGAGGTAGTAGATTGTATAGTT	TGAGGTAGTAGATTGTATAGTT
let-7g	TGAGGTAGTAGTTTGTACAGTT	TGAGGTAGTAGTTTGTACAGTT
let-7i	TGAGGTAGTAGTTTGTGCTGTT	TGAGGTAGTAGTTTGTGCTGTT
miR-130a	CAGTGCAATGTTAAAAGGGCAT	CAGTGCAATGTTAAAAGGGCAT
miR-130b	CAGTGCAATGATGAAAGGGCAT	CAGTGCAATGATGAAAGGGCAT
miR-15a	TAGCAGCACATAATGGTTTGTG	TAGCAGCACATAATGGTTTGTG
miR-15b	TAGCAGCACATCATGGTTTACA	TAGCAGCACATCATGGTTTACA
miR-181a	AACATTCAACGCTGTCGGTGAGT	AACATTCAACGCTGTCGGTGAGT
miR-181b	AACATTCATTGCTGTCGGTGGGT	AACATTCATTGCTGTCGGTGGGT
miR-200a	TAACACTGTCTGGTAACGATGT	TAACACTGTCTGGTAACGATGT
miR-200b	TAATACTGCCTGGTAATGATGA	TAATACTGCCTGGTAATGATGA
miR-23a	ATCACATTGCCAGGGATTTCC	ATCACATTGCCAGGGATTTCC
miR-23b	ATCACATTGCCAGGGATTACC	ATCACATTGCCAGGGATTACC
miR-26a	TTCAAGTAATCCAGGATAGGCT	TTCAAGTAATCCAGGATAGGCT
miR-26b	TTCAAGTAATTCAGGATAGGT	TTCAAGTAATTCAGGATAGGT
miR-29a	TAGCACCATCTGAAATCGGTTA	TAGCACCATCTGAAATCGGTTA
miR-29b	TAGCACCATTTGAAATCAGTGTT	TAGCACCATTTGAAATCAGTGTT
miR-29c	TAGCACCATTTGAAATCGGTTA	TAGCACCATTTGAAATCGGTTA
miR-30a	TGTAAACATCCTCGACTGGAAG	TGTAAACATCCTCGACTGGAAG
miR-30b	TGTAAACATCCTACACTCAGCT	TGTAAACATCCTACACTCAGCT
miR-30c	TGTAAACATCCTACACTCTCAGC	TGTAAACATCCTACACTCTCAGC
miR-30d	TGTAAACATCCCCGACTGGAAG	TGTAAACATCCCCGACTGGAAG
miR-99a	AACCCGTAGATCCGATCTTGTG	AACCCGTAGATCCGATCTTGTG

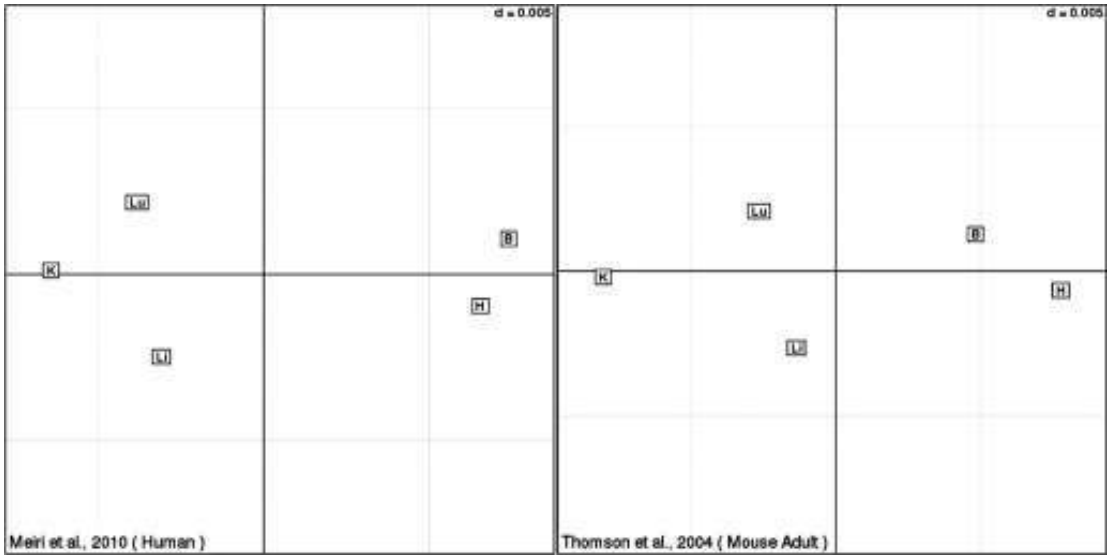


Figure 3.2.1: Coinertia plot of Meiri and Thomson expression data sets for a set of microRNA clusters with sequence similarity. Similarity of microRNA expression patterns between mice and humans are shown for brain (B), heart (H), kidney (K), liver (Li), and lung (Lu).

Accordingly, it has been found that microRNAs in the uploaded list were expressed similarly in human and mouse data sets because the location of the projected tissues closely corresponded between the two plots (Figure 3.2.1). mESAdb also enables visualization of the expression of selected microRNAs from both data sets by simultaneously overlaying them on a two-dimensional plot. In this microRNA-oriented view, similarly expressed microRNAs are found closer in space.

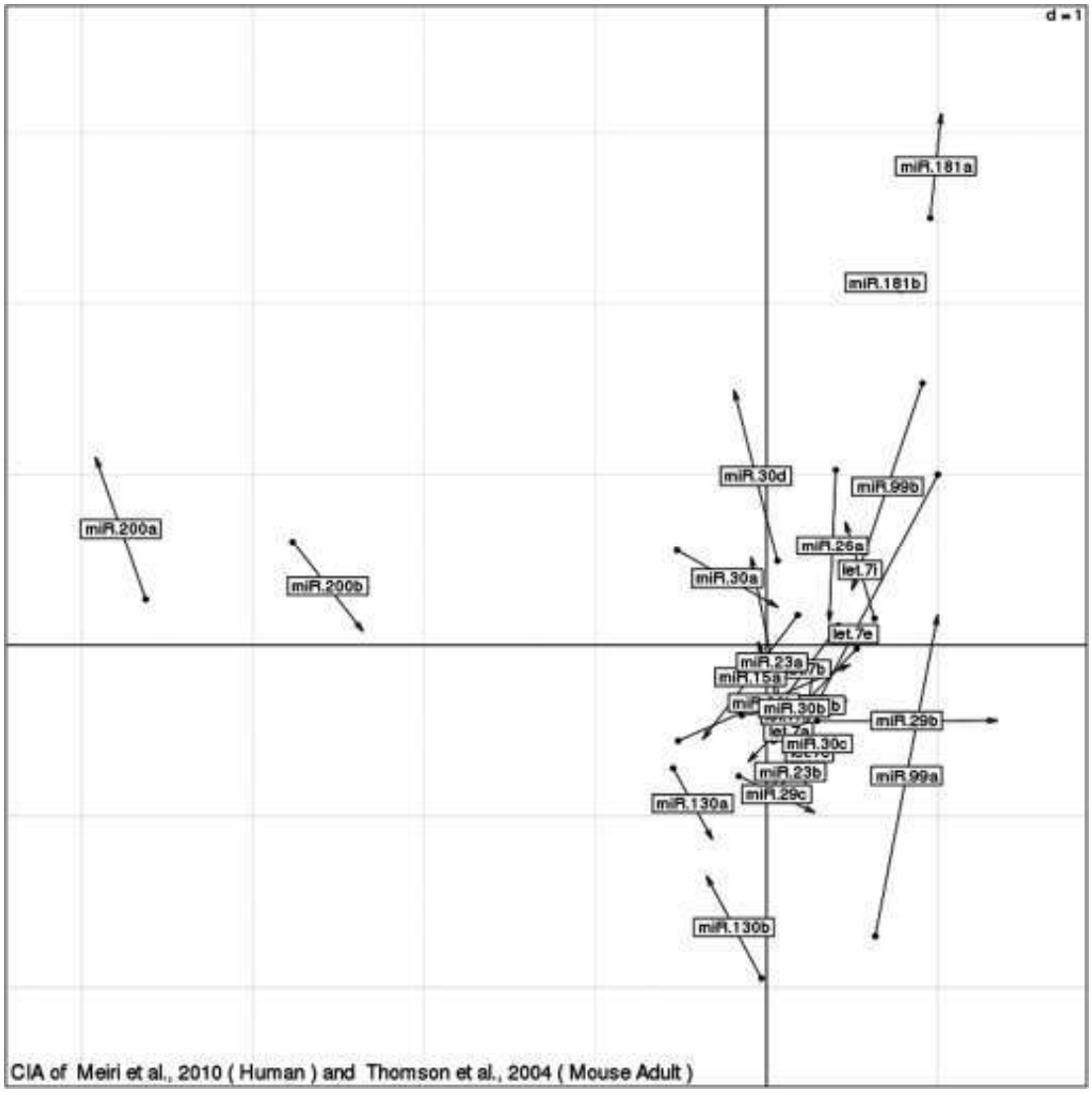


Figure 3.2.2: Distribution of microRNAs after dimension reduction by co-inertia analysis. microRNAs related in expression clustered together.

The figure (Figure 3.2.2) shows the analysis of the uploaded microRNA list via co-inertia between two dataset of different species. The length of an arrow coupled to a microRNA correlates with the amount of expression divergence for a particular microRNA between the two data sets, i.e. human versus mouse. The analysis result shown in Figure 3.2.2 indicated that several microRNAs formed clusters based on their expression, in particular, *miR-181a* and *miR-181b*, and *miR-200a* and *miR-200b* (Figure 3.2.2). Indeed, *miR-181a* and *miR-181b*, similar in sequence and diverging only with 3 nucleotides, exhibit a common sequence motif

(i.e. AACATTCA) in their first 8th nucleotides microRNAs having 'AAGTGC' seed motif. Similarly, *miR-200a* and *miR-200b* are similar in their sequences containing a common motif (i.e. TAA[C][T]ACTG) in their first 8 nucleotides (Table 3.2.1).

Table 3.2.2: Locations of microRNAs in human genome that are used for co-ineria analysis between Meiri et al., 2010 and Thomson et al., 2004.

Chromosome	Start	End	Strand	ENSEMBL ID	miRBase ID
9	96938239	96938318	+	MI0000060	hsa-let-7a-1
11	122017230	122017301	-	MI0000061	hsa-let-7a-2
22	46508629	46508702	+	MI0000062	hsa-let-7a-3
22	46509566	46509648	+	MI0000063	hsa-let-7b
21	17912148	17912231	+	MI0000064	hsa-let-7c
9	96941116	96941202	+	MI0000065	hsa-let-7d
19	52196039	52196117	+	MI0000066	hsa-let-7e
9	96938629	96938715	+	MI0000067	hsa-let-7f-1
X	53584153	53584235	-	MI0000068	hsa-let-7f-2
3	52302294	52302377	-	MI0000433	hsa-let-7g
12	62997466	62997549	+	MI0000434	hsa-let-7i
11	57408671	57408759	+	MI0000448	hsa-miR-130a
22	22007593	22007674	+	MI0000748	hsa-miR-130b
13	50623255	50623337	-	MI0000069	hsa-miR-15a
3	160122376	160122473	+	MI0000438	hsa-miR-15b
1	198828173	198828282	-	MI0000289	hsa-miR-181a-1
9	127454721	127454830	+	MI0000269	hsa-miR-181a-2
1	198828002	198828111	-	MI0000270	hsa-miR-181b-1
9	127455989	127456077	+	MI0000683	hsa-miR-181b-2
1	1103243	1103332	+	MI0000737	hsa-miR-200a
1	1102484	1102578	+	MI0000342	hsa-miR-200b
19	13947401	13947473	-	MI0000079	hsa-miR-23a
9	97847490	97847586	+	MI0000439	hsa-miR-23b
3	38010895	38010971	+	MI0000083	hsa-miR-26a-1
12	58218392	58218475	-	MI0000750	hsa-miR-26a-2
2	219267369	219267445	+	MI0000084	hsa-miR-26b
7	130561506	130561569	-	MI0000087	hsa-miR-29a
7	130562218	130562298	-	MI0000105	hsa-miR-29b-1
1	207975788	207975868	-	MI0000107	hsa-miR-29b-2
1	207975197	207975284	-	MI0000735	hsa-miR-29c
6	72113254	72113324	-	MI0000088	hsa-miR-30a
8	135812763	135812850	-	MI0000441	hsa-miR-30b
1	41222956	41223044	+	MI0000736	hsa-miR-30c-1
6	72086663	72086734	-	MI0000254	hsa-miR-30c-2
8	135817119	135817188	-	MI0000255	hsa-miR-30d
21	17911409	17911489	+	MI0000101	hsa-miR-99a

Table 3.2.3: Locations of microRNAs in mouse genome that are used for co-ineria analysis between Meiri et al., 2010 and Thomson et al., 2004.

Chromosome	Start	End	Strand	ENSEMBL ID	miRBase ID
13	48633548	48633641	-	MI0000556	mmu-let-7a-1
9	41344799	41344894	+	MI0000557	mmu-let-7a-2
15	85537749	85537833	+	MI0000558	mmu-let-7b
16	77599902	77599995	+	MI0000559	mmu-let-7c-1
15	85537033	85537127	+	MI0000560	mmu-let-7c-2
13	48631381	48631483	-	MI0000405	mmu-let-7d
17	17967316	17967408	+	MI0000561	mmu-let-7e
13	48633198	48633286	-	MI0000562	mmu-let-7f-1
X	148346889	148346971	+	MI0000563	mmu-let-7f-2
9	106081171	106081258	+	MI0000137	mmu-let-7g
10	122422696	122422780	-	MI0000138	mmu-let-7i
14	62250864	62250947	-	MI0000564	mmu-mir-15a
3	68813694	68813757	+	MI0000140	mmu-mir-15b
8	86732417	86732491	+	MI0000571	mmu-mir-23a
13	63401792	63401865	+	MI0000141	mmu-mir-23b
9	118940914	118941003	+	MI0000573	mmu-mir-26a-1
10	126432586	126432669	+	MI0000706	mmu-mir-26a-2
1	74440884	74440968	+	MI0000575	mmu-mir-26b
6	31012660	31012747	-	MI0000576	mmu-mir-29a
6	31013023	31013093	-	MI0000143	mmu-mir-29b-1
1	196863234	196863314	+	MI0000712	mmu-mir-29b-2
1	196863741	196863828	+	MI0000577	mmu-mir-29c
1	23279108	23279178	+	MI0000144	mmu-mir-30a
15	68168977	68169072	-	MI0000145	mmu-mir-30b
4	120442139	120442227	-	MI0000547	mmu-mir-30c-1
1	23298540	23298623	+	MI0000548	mmu-mir-30c-2
15	68172770	68172851	-	MI0000549	mmu-mir-30d
16	77599181	77599245	+	MI0000146	mmu-mir-99a
2	84581272	84581335	-	MI0000156	mmu-mir-130a
16	17124154	17124235	-	MI0000408	mmu-mir-130b
1	139863032	139863118	+	MI0000697	mmu-mir-181a-1
2	38708255	38708330	+	MI0000223	mmu-mir-181a-2
1	139863216	139863295	+	MI0000723	mmu-mir-181b-1
2	38709350	38709438	+	MI0000823	mmu-mir-181b-2
4	155429005	155429094	-	MI0000554	mmu-mir-200a
4	155429790	155429859	-	MI0000243	mmu-mir-200b

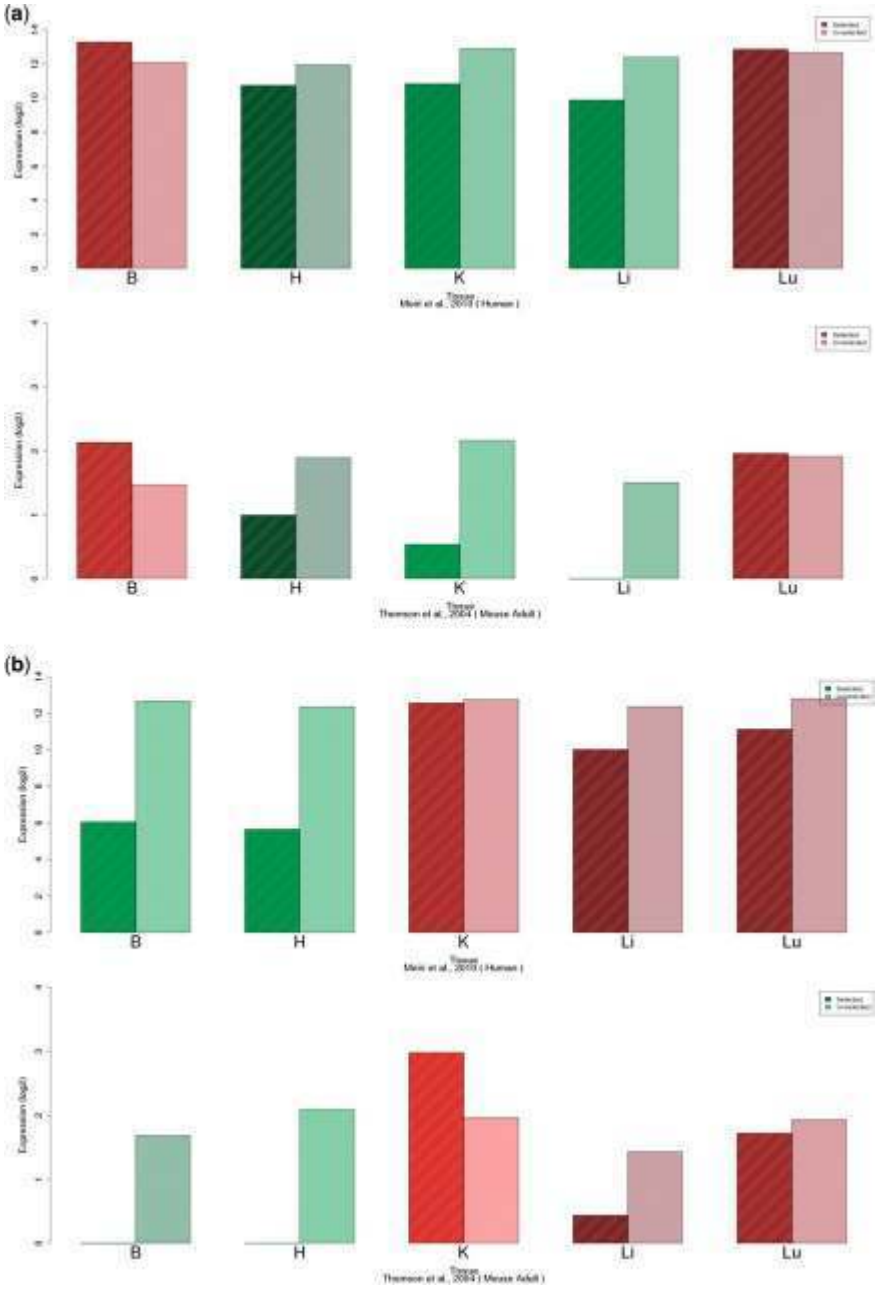


Figure 3.2.3: Similarity of expression of microRNA expression from Meiri and Thomson. Expression bar plot of (a) *miR-181a* and *miR-181b* cluster (b) *miR-200a* and *miR-200b* for five different tissues [i.e. brain (B), heart (H), kidney (K), liver (Li), and lung (Lu), respectively.]. For each tissue, the bar on the left indicates the mean expression of the members of the cluster and the right hand bar indicates the mean expression of the remainder of the data set.

Using the 'expression analysis' module, *miR-181a* and *miR-181b* were found to be expressed primarily in the brain and lung (Figure 3.2.3-a) whereas the *miR-*

200a-b cluster was clearly expressed mostly in the kidney and lung both in mice and humans (Figure 3.2.3-b). These findings suggested that expression patterns of *miR-181a-b* and *miR-200a-b* were highly conserved between human and mice.

3.3 SEARCHING FOR A DISEASE ASSOCIATION MICRORNAS USING *HUGE NAVIGATOR*

This feature is only available for human microRNAs since the *HuGE* database deals with human epidemiological studies. Searching for a disease association for microRNA(s) is possible in two ways. One can either use 'motif & function' module or 'microRNA search' module.

In the first option, one can upload a text file, can enter a motif to select a list of microRNAs that bears it or can select a dinucleotide as directed by the selection list menu 'select miRNA upon:'.

Figure 3.3.1: Motif and function module. You can associate GO terms, HUGE terms and KEGG pathways for a set of selected microRNA by a criteria that you set here.

After choosing the selection criteria the user is directed to a page where one can analyze the selected microRNAs for either of the following: 1) *GO*, 2) *HuGE* and 3) *KEGG* terms. An example is provided here while focusing on liver specific microRNAs (Chen 2009). In this editorial, a table showing the list of microRNAs proven to be expressed in liver according to various studies together with their expression levels has been adopted from Girard et al (Girard, Jacquemin et al. 2008; Chen 2009). From this table, the top thirteen microRNAs have been selected, and a file containing those has been created as a liver-specific microRNA file example.

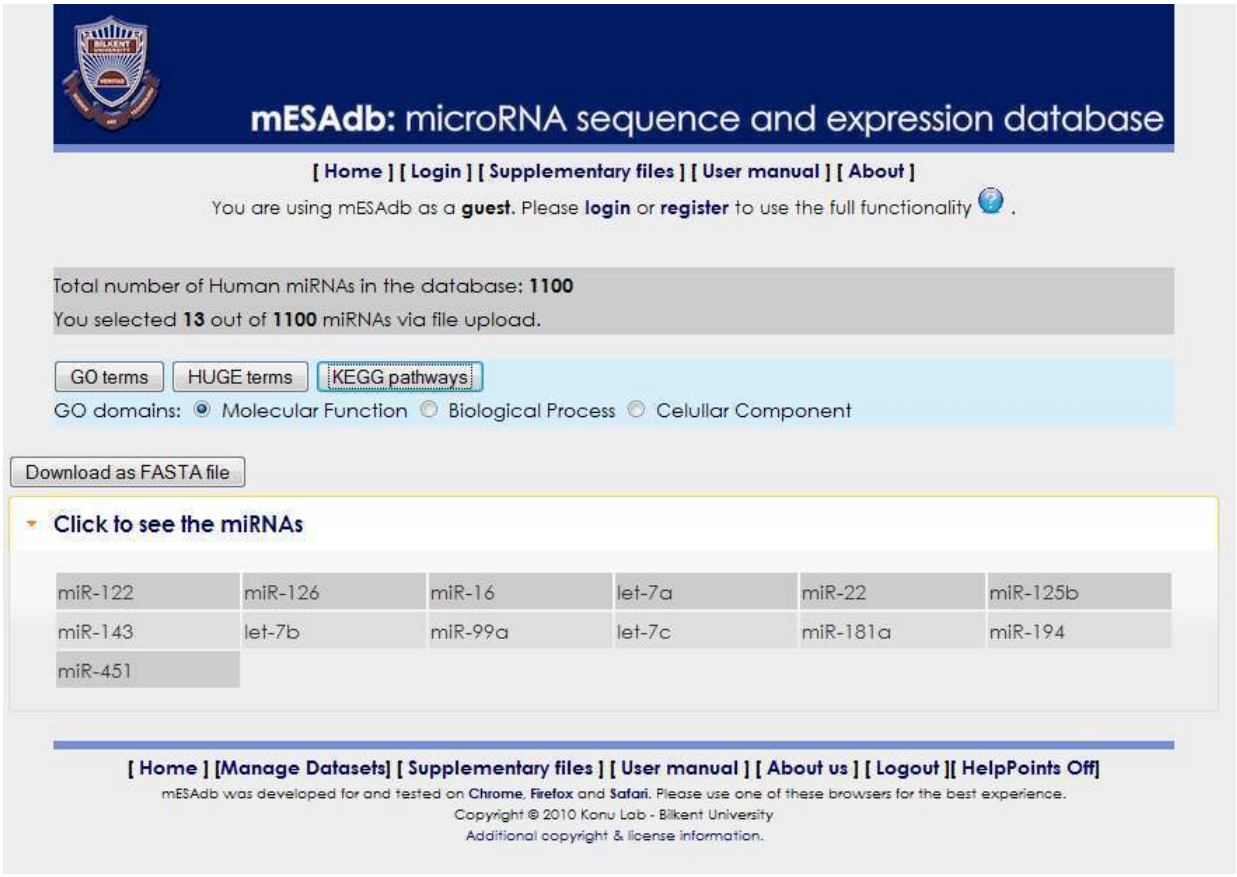


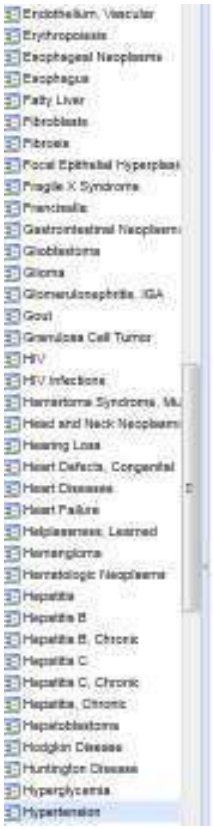
Figure 3.3.2: After the upload of liver related microRNAs, the page to which the client is directed. Here one can associate the microRNAs to HUGE, KEGG and GO terms. FASTA formats of uploaded microRNAs could also be seen.

As mentioned above, mESAdb allows for upload of a text file containing a list of microRNAs. In this case, the file containing a list of liver specific microRNAs obtained from Chen et al, 2009 (Chen 2009), and was uploaded from which one could associate these microRNAs with a HUGE term together with a probability of this association being due to change. The result has been shown in the figure below (Figure 3.3.3). The findings were as expected in that liver specific microRNAs were found to be associated with liver specific diseases such as Hepatitis B and C.

Figure 3.3.3: The HUGE terms associated with the uploaded microRNAs primarily expressed in liver.

One of the ways of associating a microRNA with a disease is investigating the relationship between its target and the disease. In mESAdb it has been applied. However, in 2008, Lu et al, have built a database called human microRNA-associated disease database (*HMDD* http://202.38.126.151/hmdd/mirna/md/) (Lu, Zhang et al. 2008). This database has been built by inferring microRNA-disease association by literature scanning (Lu, Zhang et al. 2008).

A comparison of two database results has been shown here. First, from the *HMDD*, hypertension disease was selected (Figure 3.3.3).



MiRNA	note	Reference
hsa-mir-1-1	miR-1: deregulation	18690400
hsa-mir-1-2	miR-1: deregulation	18690400
hsa-mir-133a-1	miR-133a: deregulation	18690400
hsa-mir-133a-2	miR-133a: deregulation	18690400
hsa-mir-155	miR-155: deregulation	18690400
hsa-mir-208b	miR-208: deregulation	18690400
hsa-mir-122	a polymorphism of the 3'UTR of the SLC7A1 gene affects the binding with miR-122	19067360
hsa-mir-204	Reestablishing miR-204 expression should be explored as a potential new therapy for this disease	21321078
hsa-mir-637	A common genetic variant (rs938671) in the 3'-UTR of Vacuolar H+-ATPase ATP6V0A1 creates a micro-RNA (hsa-miR-637) motif to alter Chromogranin A (CHGA) processing and hypertension risk.	21558123

Figure 3.3.4: A snapshot of HMDD. From the left panel "hypertension" was selected and on the right frame related microRNAs appeared.

microRNAs related to this disease were put into a file and uploaded to the mESAdb through 'motif & expression' module and all 6 microRNAs were found on mESAdb (Figure 3.3.5). Obviously one should expect a difference between two databases, since the association methods are different i.e., *HMDD* uses direct relation mined from published texts whereas mESAdb uses target gene-disease association via *HuGE Navigator*.

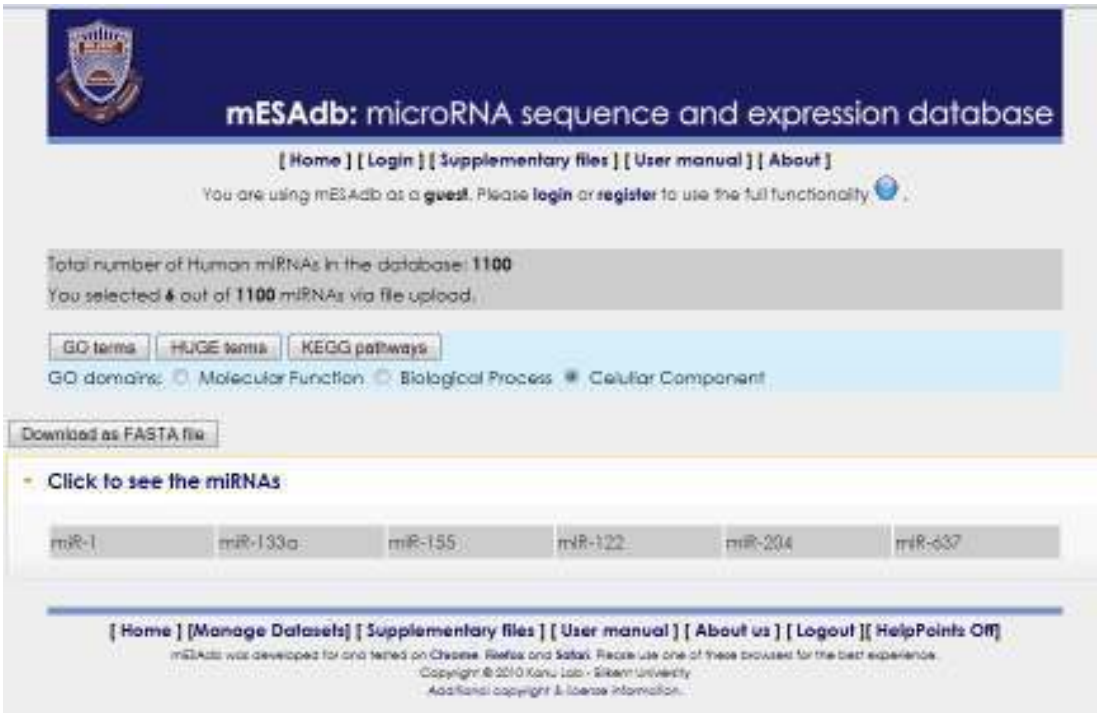


Figure 3.3.5: microRNAs associated to hypertension according to the HMDD were uploaded to mESAdb.

The association results from mESAdb have been shown in Figure 3.3.6. The top term is not "hypertension", however it is a relevant term, namely, "Cardiovascular Diseases" which includes "hypertension".

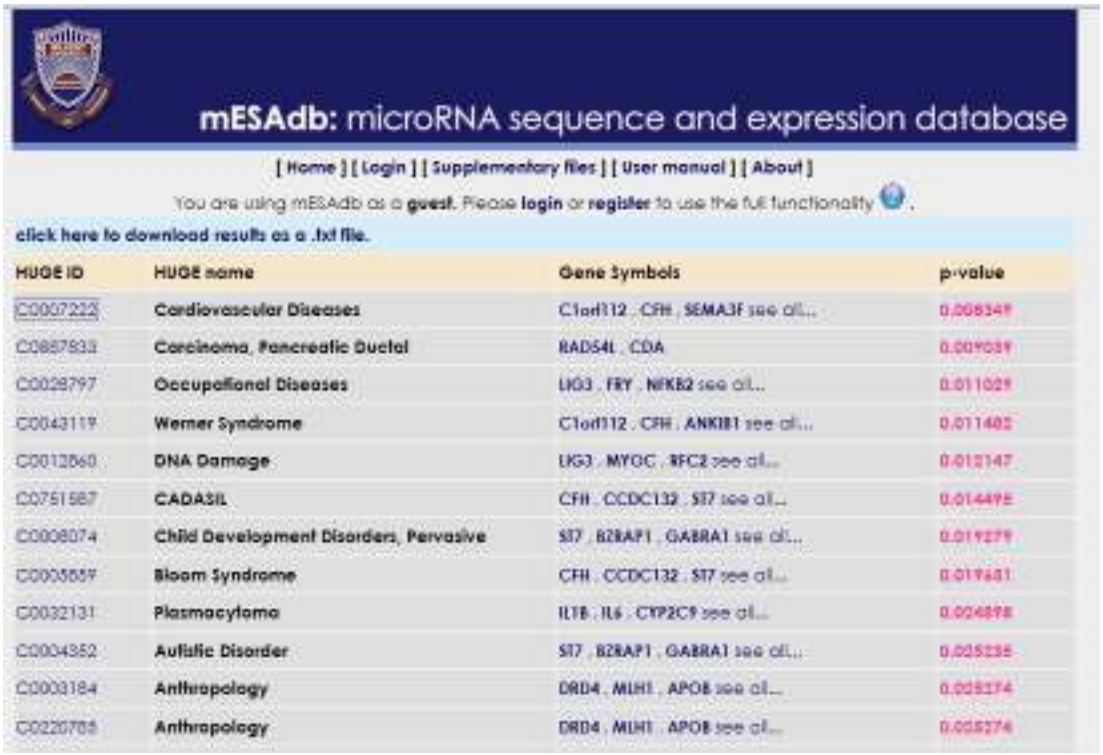
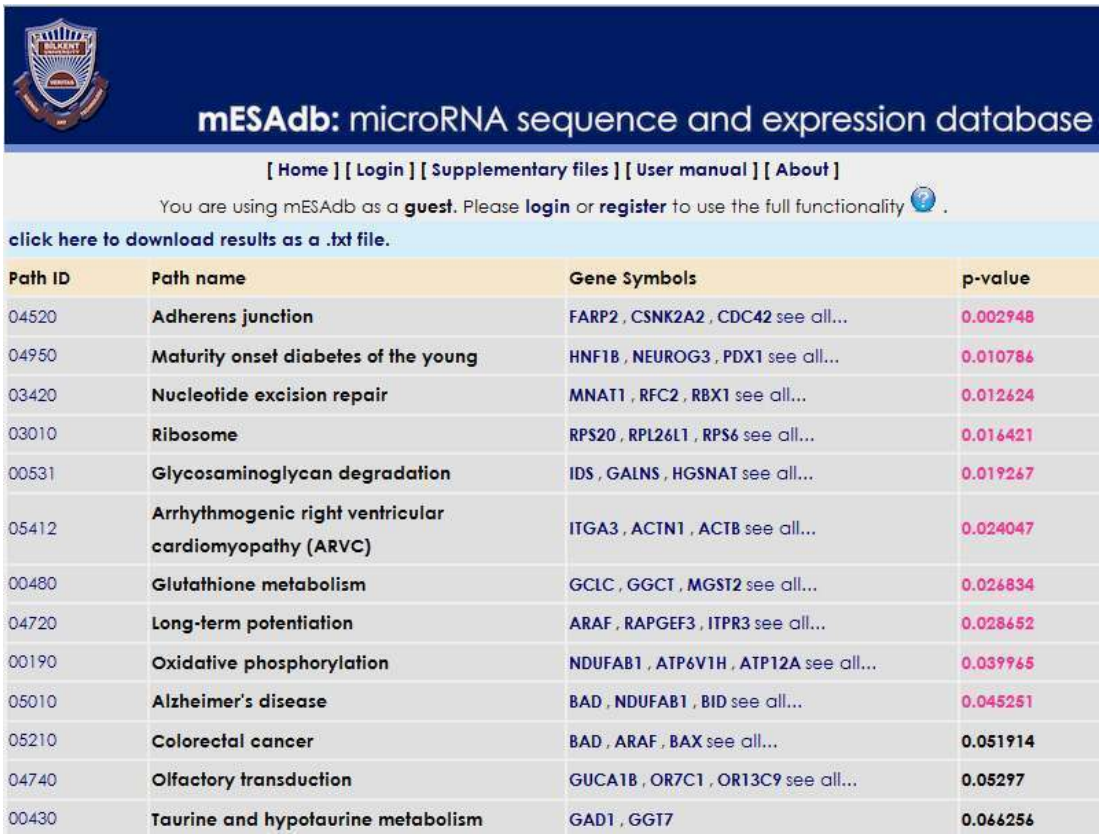


Figure 3.3.6: mESAdb association of the selected microRAs to HUGE terms.

3.4 SEARCHING FOR KEGG ASSOCIATED WITH MICRORNAS

The same set of miRNAs has also been used for finding significantly relevant KEGG terms. Results are shown in Figure 3.4.1. There are 10 significant KEGG terms associated with those micrroRNAs from liver (Chen 2009). These top terms include some keywords, such as diabetes and glycosaminoglycan degradation, related to liver function, as they are metabolism related. This is discussed further in the discussion section.



mESAdb: microRNA sequence and expression database

[Home] [Login] [Supplementary files] [User manual] [About]

You are using mESAdb as a **guest**. Please **login** or **register** to use the full functionality .

click here to download results as a .txt file.

Path ID	Path name	Gene Symbols	p-value
04520	Adherens junction	FARP2 , CSNK2A2 , CDC42 see all...	0.002948
04950	Maturity onset diabetes of the young	HNF1B , NEUROG3 , PDX1 see all...	0.010786
03420	Nucleotide excision repair	MNAT1 , RFC2 , RBX1 see all...	0.012624
03010	Ribosome	RPS20 , RPL26L1 , RPS6 see all...	0.016421
00531	Glycosaminoglycan degradation	IDS , GALNS , HGSNAT see all...	0.019267
05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	ITGA3 , ACTN1 , ACTB see all...	0.024047
00480	Glutathione metabolism	GCLC , GGCT , MGST2 see all...	0.026834
04720	Long-term potentiation	ARAF , RAPGEF3 , ITPR3 see all...	0.028652
00190	Oxidative phosphorylation	NDUFAB1 , ATP6V1H , ATP12A see all...	0.039965
05010	Alzheimer's disease	BAD , NDUFAB1 , BID see all...	0.045251
05210	Colorectal cancer	BAD , ARAF , BAX see all...	0.051914
04740	Olfactory transduction	GUCA1B , OR7C1 , OR13C9 see all...	0.05297
00430	Taurine and hypotaurine metabolism	GAD1 , GGT7	0.066256

Figure 3.4.1: The KEGG terms associated with the uploaded microRNAs primarily expressed in liver.

3.5 CHRNA5 TARGETING MICRORNAS AND THE ESTROGEN RECEPTOR

GSE15885 is a dataset comprising 29 early stage breast cancer samples and expression of 353 microRNAs from this were investigated in this study in the context of estrogen receptor status. The estrogen receptor (ER), progesterone receptor (PR) and Human Epidermal growth factor Receptor (HER2/neu) expression status of the samples were given. The SOFT formatted family files contained background subtracted, median normalized and log transformed data.

By using the *GEOquery* library, a *Biconductor* package for R, the meta-data that contain expression status of the aforementioned receptors, the normalized signal data and the names of mature microRNAs were fetched to construct the mESAdb .csv file format for upload. While labeling the samples only the estrogen status of the cells

was considered. The uploaded file has been named 'GSE15885_er' and quantile normalized.

The gene with ENSEMBL ID 'ENSG00000169684' is CHRNA5. The microRNAs targeting it were queried in mESAdb underlying *MySQL* database where the target information had been collected from the *MicroCosm* of *ENSEMBL*. The release mESAdb has got 27 targeting microRNAs:

miR-106a, miR-106b, miR-126, miR-137, miR-15a, miR-15b, miR-16, miR-17, miR-20a, miR-20b, miR-220b, miR-223, miR-23a, miR-23b, miR-28-3p, miR-301a, miR-301b, miR-409-3p, miR-424, miR-497, miR-520g, miR-554, miR-561, miR-588, miR-607, miR-655, miR-93.*

These microRNAs were saved in a text file to facilitate later selection. To see how the samples project to two new dimensions that are captured top two highest mean variability in the space of microRNAs, 'motif & expression' module of mESAdb was used. After selecting the dataset, the list prepared previously was used to select CHRNA5 targeting microRNAs. The result of correspondence analysis, showing sample projections on to two dimensions based on variances of the selected microRNAs in them, has been illustrated in Figure 3.5.1.

Figure 3.5.1: The clustering of samples of GSE15885 dataset labeled according to only ER status of the cells.

First of all, the most striking thing in this figure was the existence of numerous clusters of samples. This separation also covers the discrimination of ER positive and ER negative samples. Although a small number of samples violate the trend, a line having approximately a 135 degrees angle with the x axis is separating the estrogen positive and negative samples. The ER negative samples were observed as three clusters in the Figure 3.5.1.

The figure from microRNA tab of the *Correspondence Analysis* output of the mESAdb has been illustrated in Figure 3.5.2.

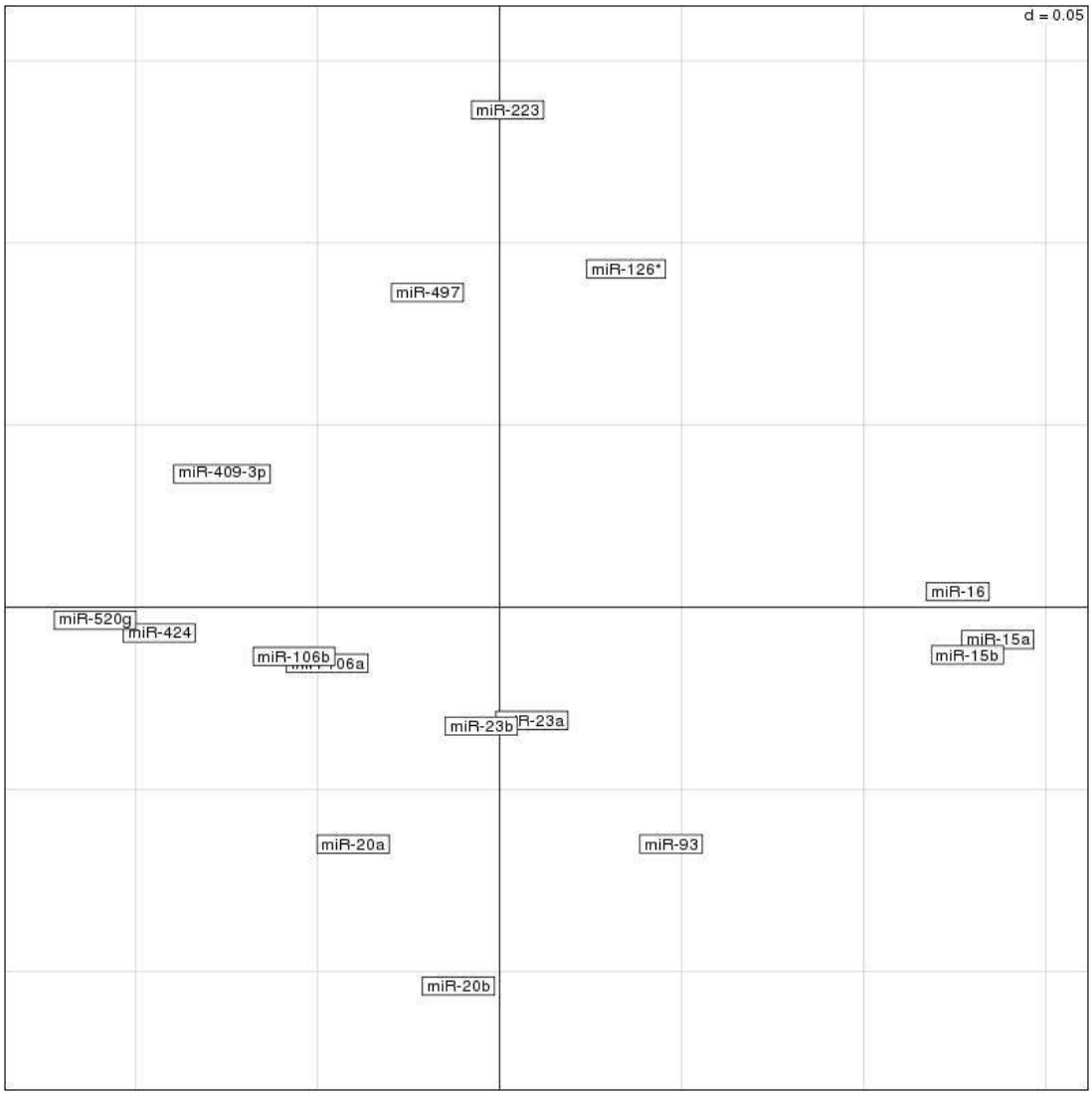


Figure 3.5.2: CHRNA5 targeting microRNA clustering after correspondence analysis of GSE15885 dataset.

The pattern observed in previously was also seen here. A line with approximately 135 degrees with the x axis seemingly would cluster the microRNAs such that every cluster could correspond to a sample cluster formed with the same separating line. In this case, most of the microRNAs should be responsible for clustering of ER negative samples. However, the Figure 3.5.3, which shows corresponding groups will make the picture clearer.

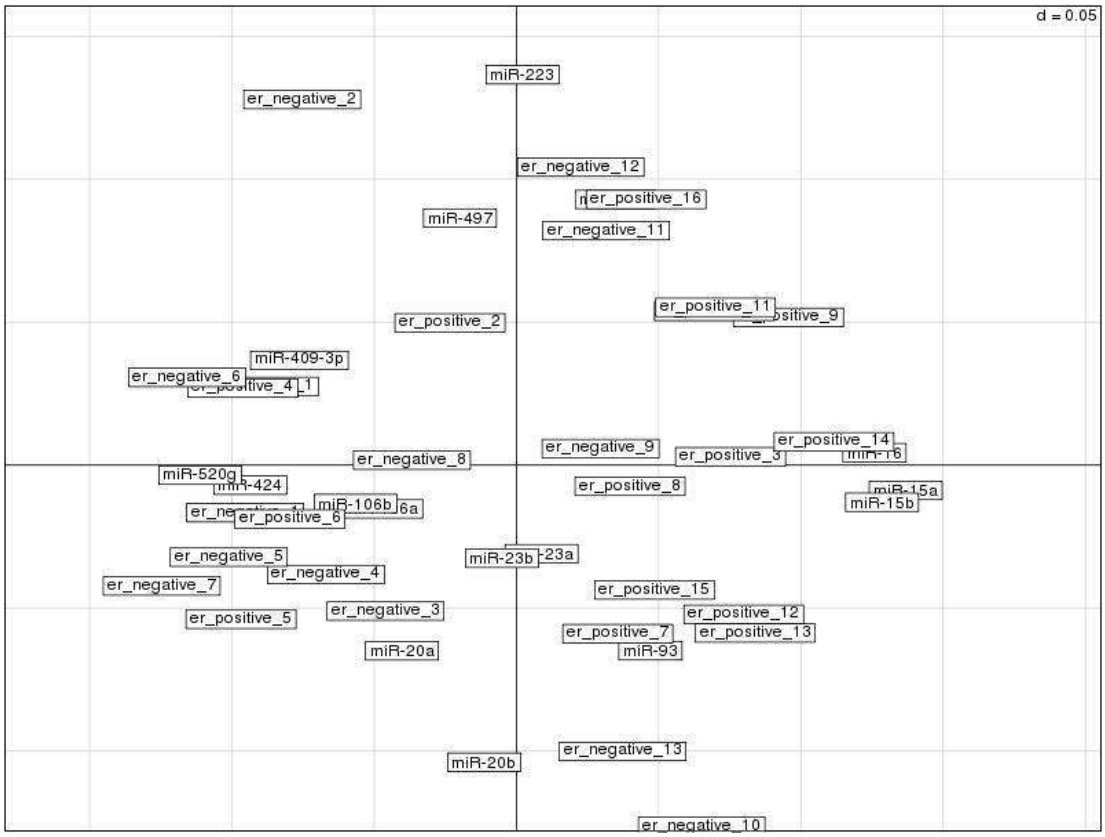


Figure 3.5.3: The correspondence tab of the output for GSE15885 dataset.

According to the figure, *miR-15a*, *miR-15b*, *miR-16*, *miR-93* was clearly related with ER positive samples. This also could be observed via expression analysis of those micro-RNAs in the dataset. Figure 3.5.4 shows the result.

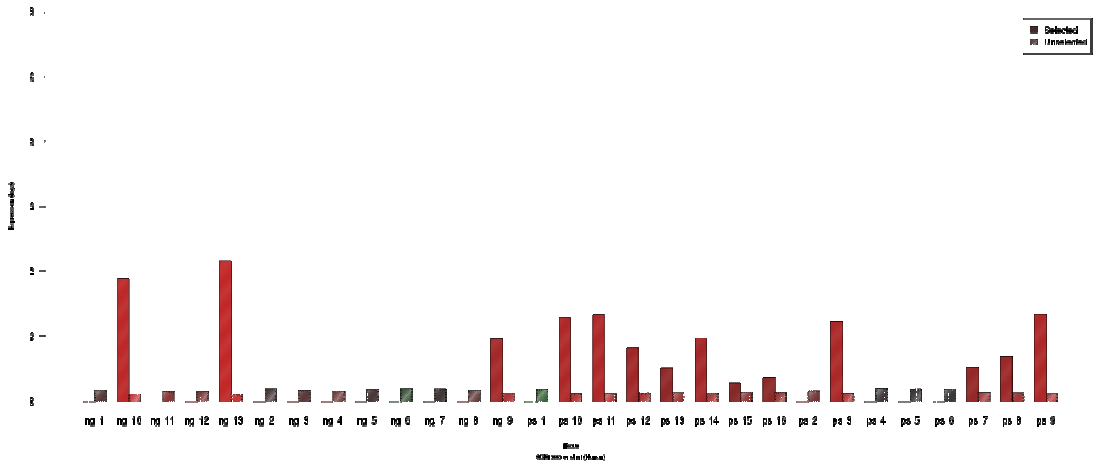


Figure 3.5.4: The expression profiles of micoRNAs that are associated with ER positive samples in GSE15885 dataset. Red bars indicate expression of the given set of microRNAs; height of the bar refers to the magnitude of expression. There are more samples annotated by ps (positive) than ng (negative) exhibiting red bars.

In this figure (Figure 3.5.4) we can clearly see that 11 out of 14 red bars are belonging to ER positive samples (the ones with the ps_ suffix.). Three samples are negative ones, ng_10, ng_9 and ng_13. To find out where theses microRNAs are bound on CHRNA5 mRNA, *microRNA.org* (Betel, Wilson et al. 2008) was used. The results are shown in Figures 3.5.5 and 3.5.6.

An mRNA can be controlled by numerous microRNAs either at one site or at many sites. For the CHRNA mRNA, the case is the latter as could be seen in figures 3.5.5 and 3.5.6. Obviously it is not expected that all microRNAs that can potentially suppress the CHRNA5 RNA are expressed at the same time. As a starting point, the binding sites of CHRNA5 targeting microRNAs have been investigated through microrna.org.

Figure 3.5.5: First part of the alignment that shows which microRNAs bind to which part of CHRNA5 mRNA.

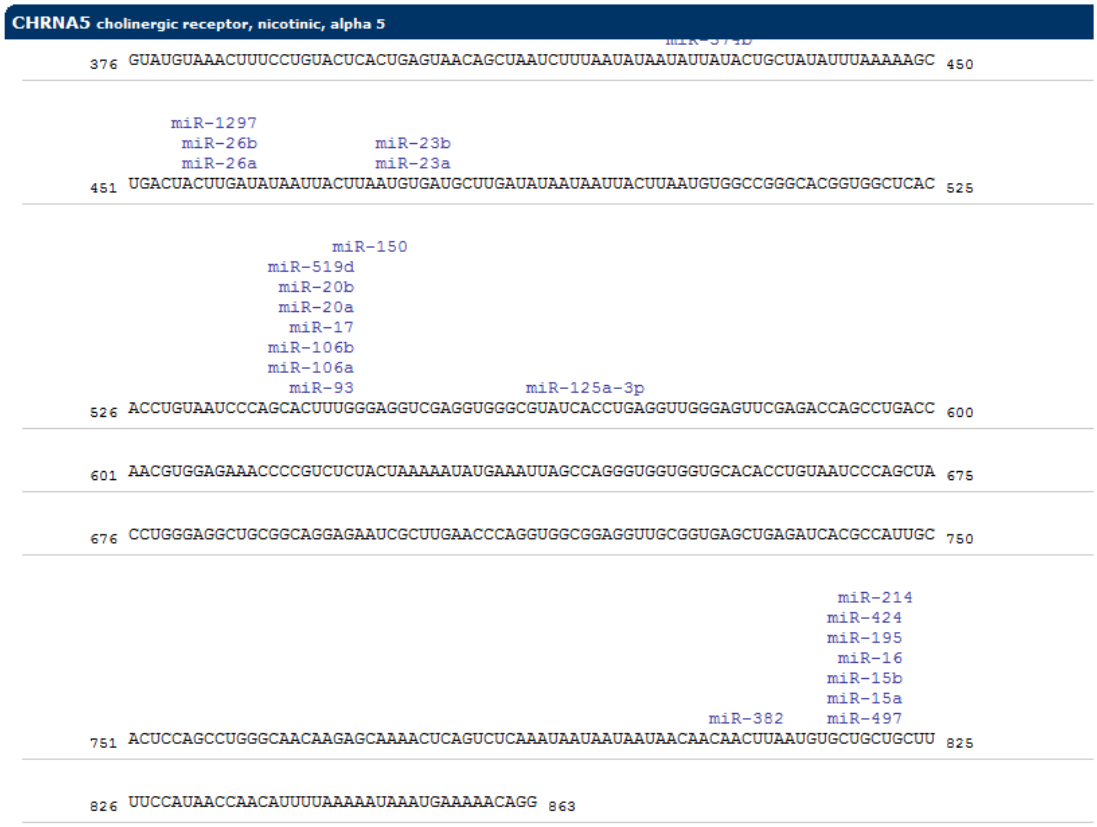


Figure 3.5.6: Remaining part of the alignment that shows which microRNAs bind to which part of CHRNA5 mRNA.

The two figures (Figure 3.5.5 and Figure 3.5.6) show the binding sites of CHRNA5 targeting microRNAs. There are two major sites that collect most of the microRNAs binding to the CHRNA5 mRNA. One of the sites is around 800th nucleotide of the CHRA5 mRNA whereas the other is around 524th base. The case has been tabularized in Tables 3.5.1 and 3.5.2.

Table 3.5.1: The microRNAs having potential binding site around 524th base of CHRNA5 mRNA

microRNA	Sequence	Bindining site
hsa-miR-106b	TAAAGTGCTGACAGTGCAGAT	526
hsa-miR-519d	CAAAGTGCCTCCCTTTAGAGTG	526
hsa-miR-93	CAAAGTGCTGTTCGTGCAGGTAG	524
hsa-miR-106a	AAAAGTGCTTACAGTGCAGGTAG	524
hsa-miR-17	CAAAGTGCTTACAGTGCAGGTAG	524
hsa-miR-20a	TAAAGTGCTTATAGTGCAGGTAG	524
hsa-miR-20b	CAAAGTGCTCATAGTGCAGGTAG	524

Table 3.5.2: The microRNAs having potential binding site around 800th base of CHRNA5 mRNA

microRNA	Sequence	Bindining site
hsa-miR-15b	TAGCAGCACATCATGGTTTACA	802
hsa-miR-214	ACAGCAGGCACAGACAGGCAGT	802
hsa-miR-497	CAGCAGCACACTGTGGTTTGT	800
hsa-miR-15a	TAGCAGCACATAATGGTTTGTG	800
hsa-miR-195	TAGCAGCACAGAAATATTGGC	800
hsa-miR-424	CAGCAGCAATTCATGTTTTGAA	800
hsa-miR-16	TAGCAGCACAGAAATATTGGC	799

The seed sequences of microRNAs in the first group are exactly the same, AAAGTGC (Table 3.5.1). For the other group, the seed sequence is a motif, AGCAG [C][G] [A][C] (Table 3.5.2). To discover any tissue specific groups of microRNAs among those, correspondence analysis has been applied to three datasets (Ach, Wang et al. 2008; Navon, Wang et al. 2009; Meiri, Levy et al. 2010) with them. Results have been shown in Figure 3.5.7, Figure 3.5.8 and Figure 3.5.9.

Figure 3.5.7: Projection of microRNAs that hit around 800th and 524th nucleotides of CHRNA5 mRNA according to the Ach et al, 2008.

According to the figure (Figure 3.5.7), except the five microRNAs all of them has been accumulated together on to the space generated by the reduction of tissue space of the Ach dataset to the two dimensions via correspondence analysis. The five microRNAs showing divergence in expression have been *miR-195*, *miR-497*, *miR-214*, *miR-424* and *miR-519d*.

Figure 3.5.8: Projection of microRNAs that hit around 800th and 524th nucleotides of CHRNA5 mRNA according to the Meiri et al, 2010.

A similar projection pattern has been produced by reducing the tissue space of the Meiri dataset. There are differences and similarities between these figures Figure 3.5.8 and Figure 3.5.7. Again a bulk accumulation of microRNAs was seen in this figure as it was seen in the previous one. This time the members of microRNAs that were observed from the central part of the graphics were *miR-20b*, *miR-214*, *miR-424* and *miR-519d*. The last three ones were common in both correspondence analyses for the two datasets.

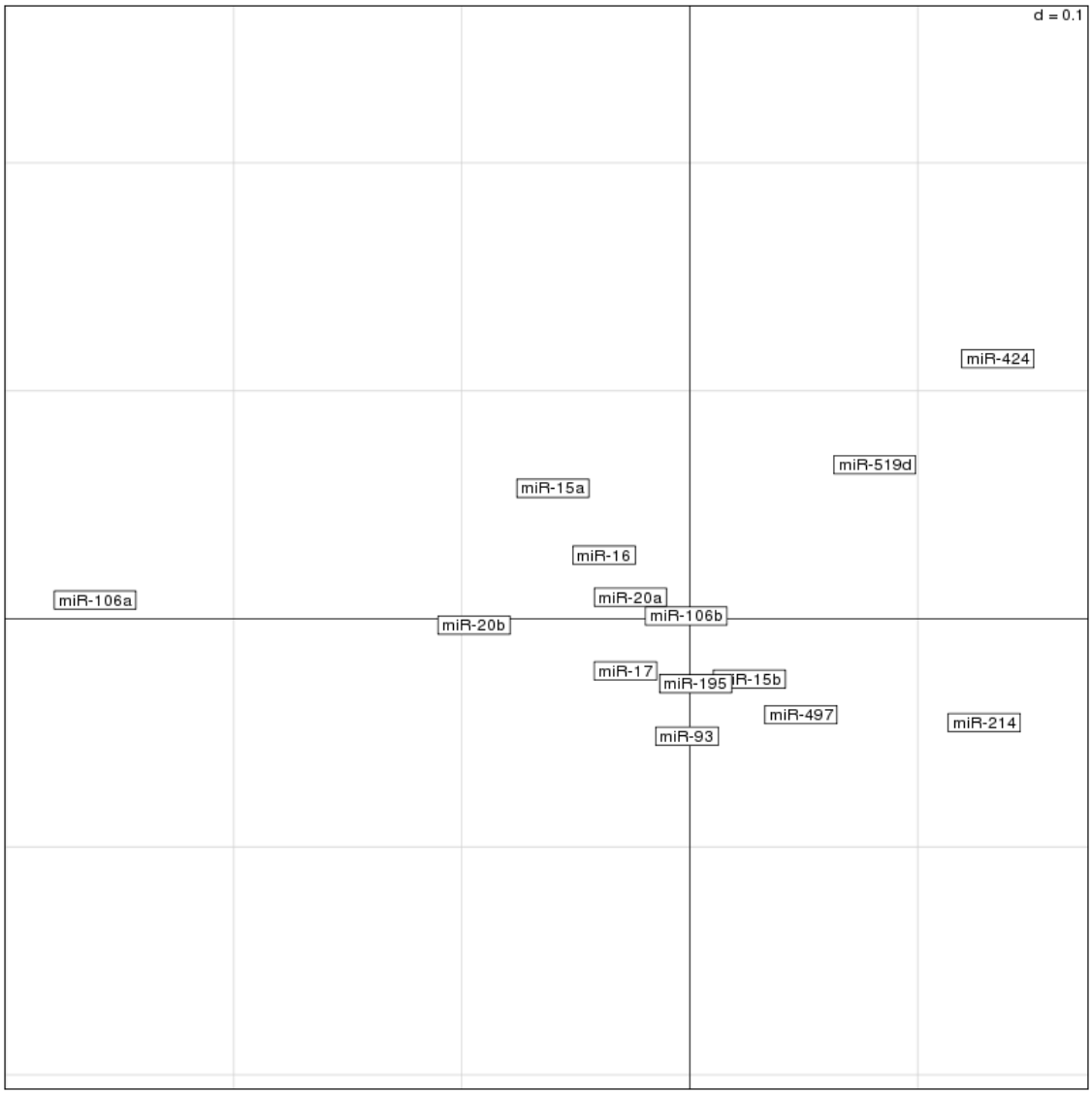


Figure 3.5.9: Projection of microRNAs that hit around 800th and 524th nucleotides of CHRNA5 mRNA according to the Navon et al, 2009.

Projections of the two groups of microRNAs don't show too much variability and wide range distribution. Although Navon et al. 2009 dataset shows difference in the view of microRNA set that formed major cluster in the previous two figures, the general trend is the same such that *miR-519d*, *miR-424* and *miR-214* deviated from the central point where all other microRNAs have tended to accumulate.

To find out if there is a consistent relationship between any microRNA and tissue, the correspondence analysis has been applied to above datasets. Tissue-microRNA correspondences for each dataset have been shown in the Figures 3.5.10-

12.

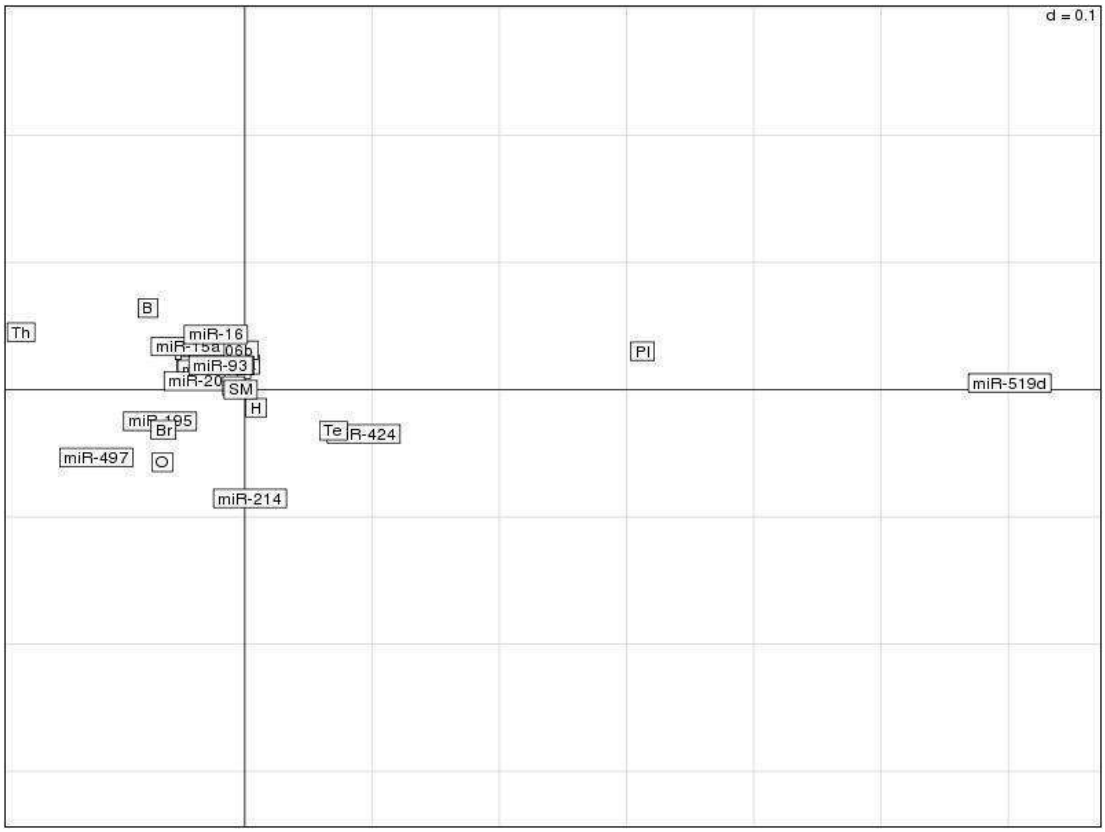


Figure 3.5.10: Projections of microRNAs that hit around 800th and 524th nucleotides of CHRNA5 mRNA and tissues according to corresponsence analysis of Ach et al, 2008.

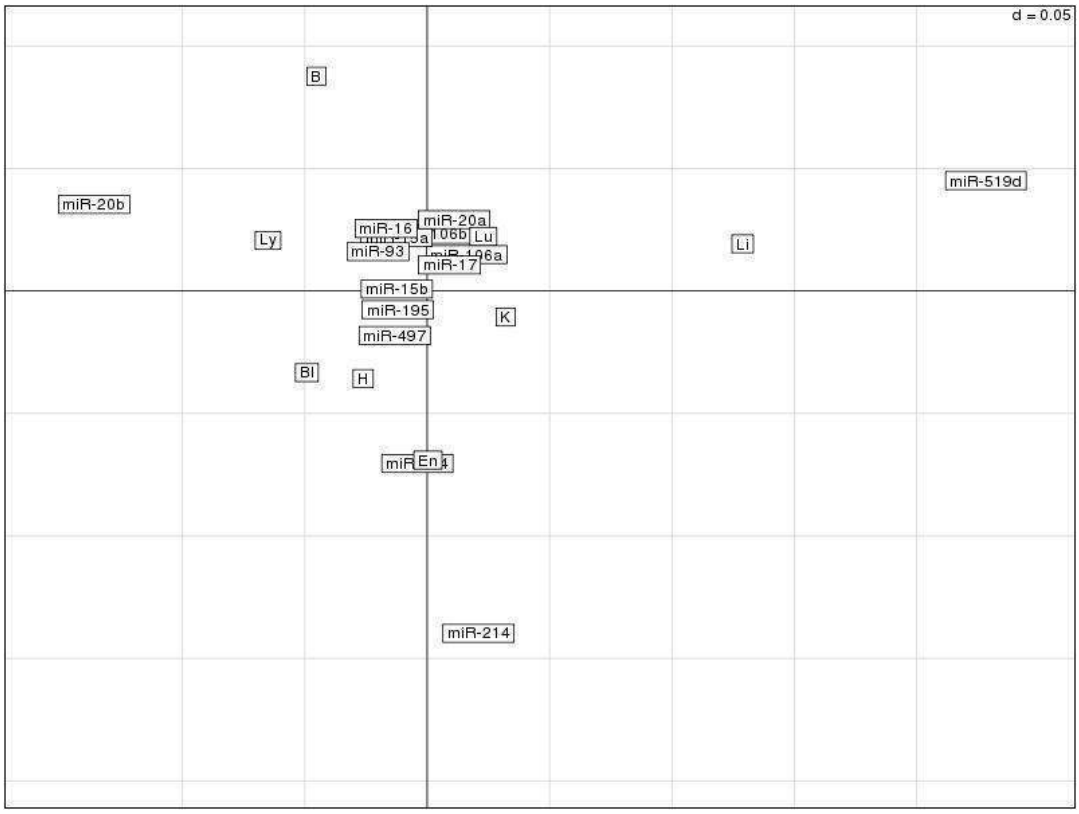


Figure 3.5.11: Projections of microRNAs that hit around 800th and 524th nucleotides of CHRNA5 mRNA and tissues according to corresponsence analysis of Meiri et al, 2010.

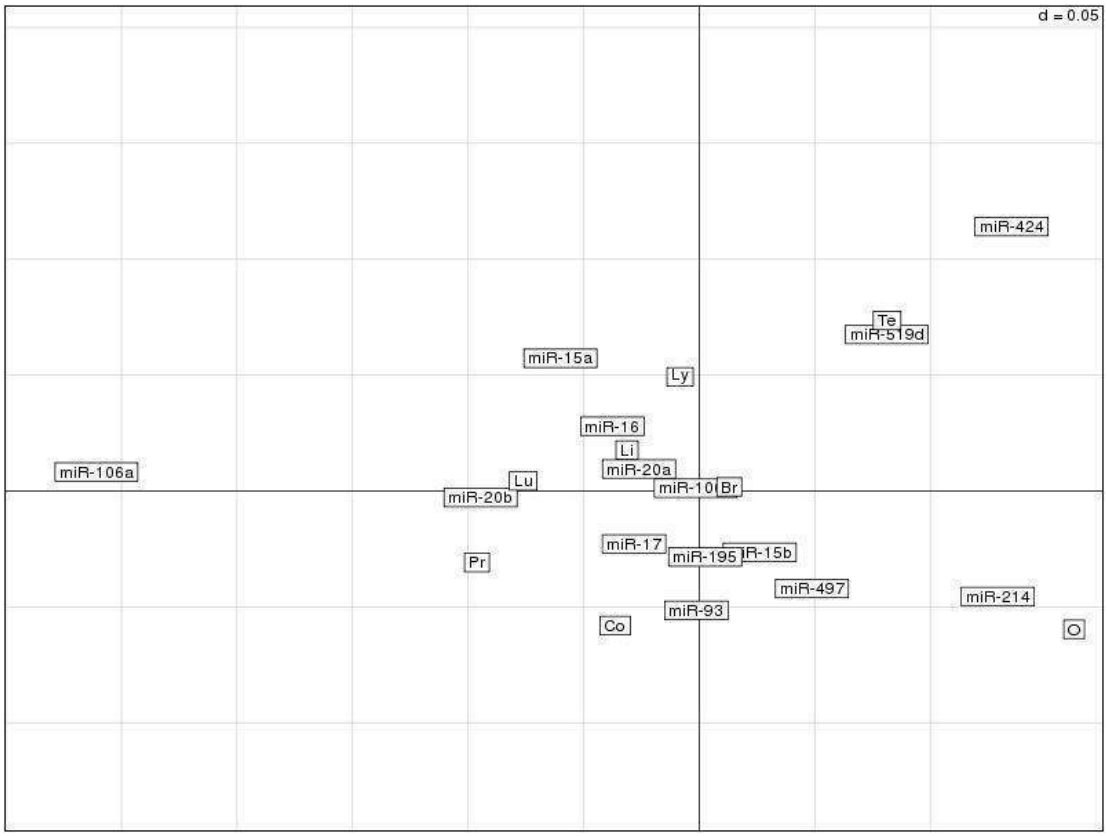


Figure 3.5.12: Projections of microRNAs that hit around 800th and 524th nucleotides of CHRNA5 mRNA and tissues according to corresponsence analysis of Navon et al, 2009..

The only tissue common to all three datasets is the liver. So, from the given figures, it is very difficult to infer any microRNA-tissue relationship. However, Navon et al, 2009 (Navon, Wang et al. 2009) and Ach et al, 2008 (Ach, Wang et al. 2008) datasets have four common tissues, namely, breast, ovary, testicle and liver. Considering this fact, from Figure 3.5.10 and Figure 3.5.13, it can be said that *miR-214* has a relation with ovary while *miR-424* has a relation with testicle. To clarify the results co-inertia analysis was applied to Ach et al, 2008 and Navon et al, 2009 with those microRNAs shown in the tables 3.5.1 and 3.5.2 by only selecting the four common tissues. The result has been illustrated in the following three figures (Figure 3.5.13, Figure 3.5.14 and Figure 3.5.15).

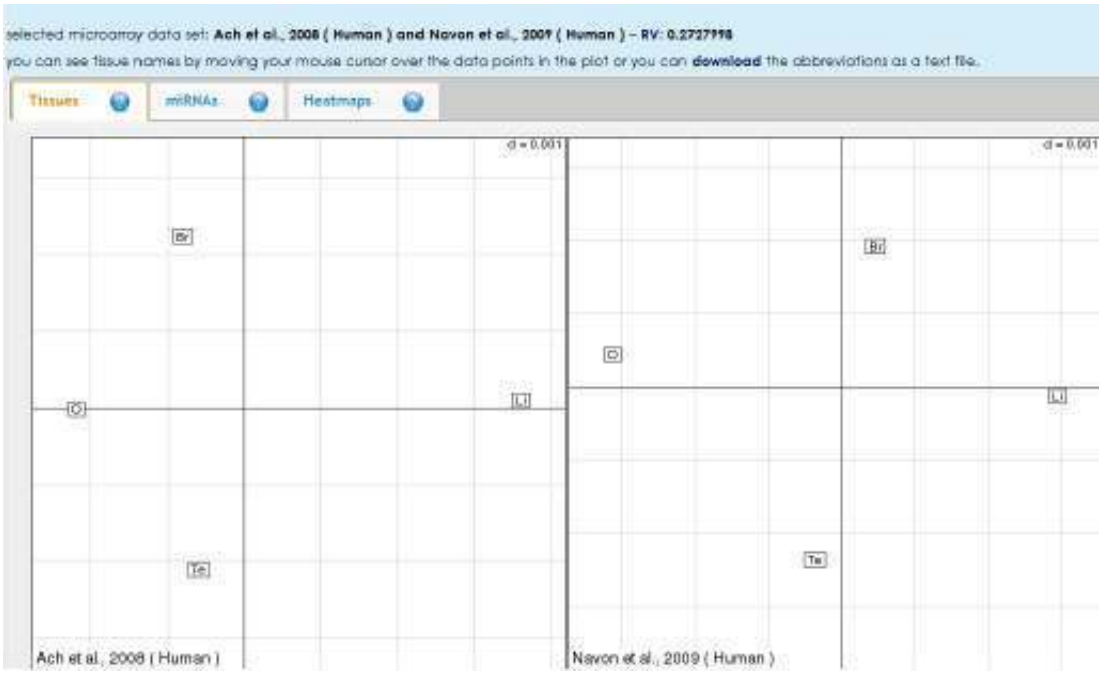


Figure 3.5.13: The distributions of common tissues on the two dimensions created by co-inertia analysis of the two datasets, Ach et al., 2008 and Navon et al., 2009, with the microRNA set listed in Table 1.3.1 and 1.3.2.

Although the RV coefficient was not very close to 1, the result of co-inertia analysis was two projections where the projected locations of tissues were similar in terms of their positions according to the origins of their maps. This means that a set of similar covariance axes exist in microRNA spaces of both datasets. The microRNA projections on to two dimensions derived from common tissue spaces of both datasets via co-inertia analysis have been illustrated in Figure 3.5.14.

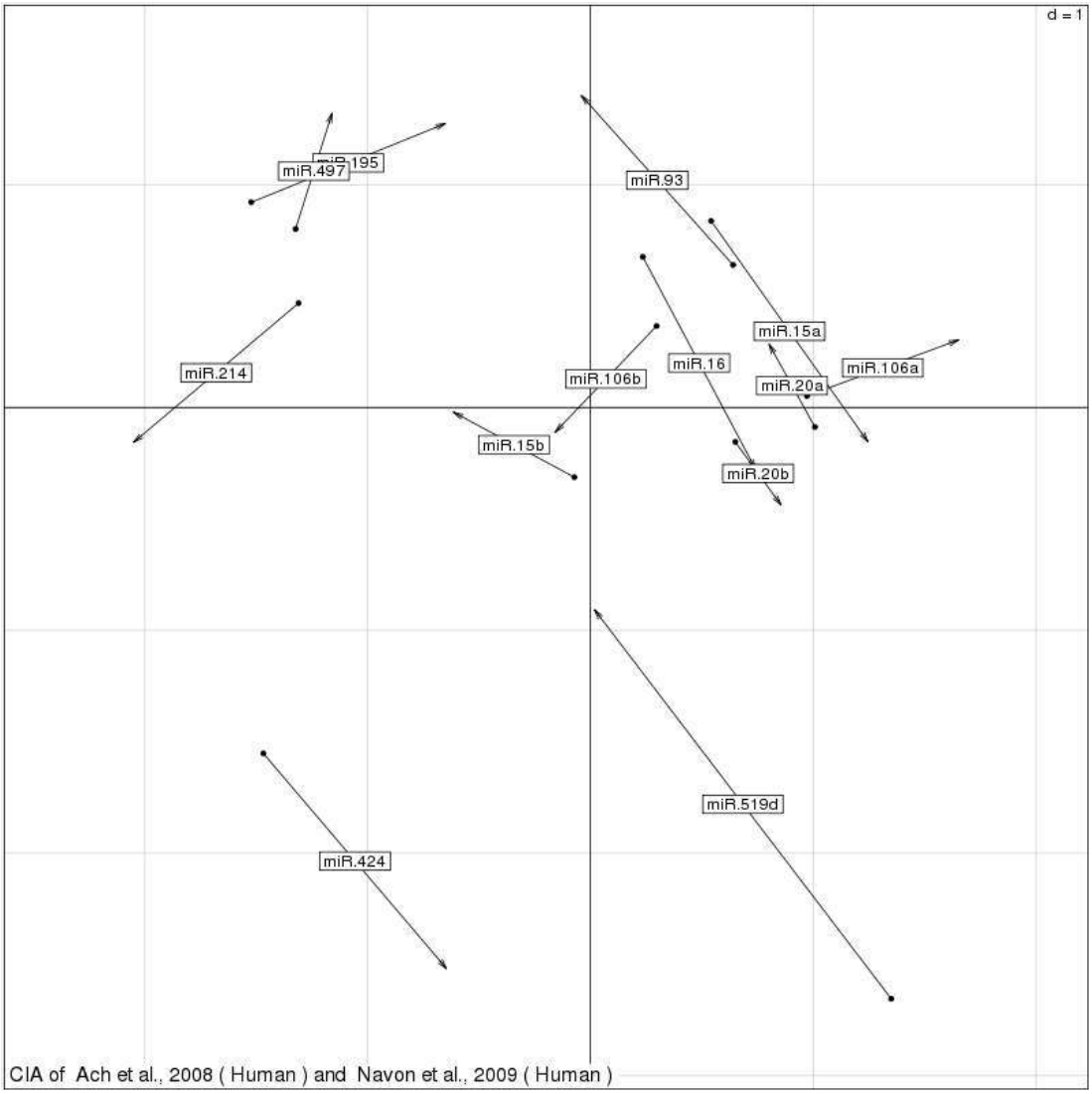


Figure 3.5.14: Projections of microRNA data points after co-inertia analysis of both datasets, Ach et al., 2008 and Navon et al., 2009, with their common tissues.

Expectedly, the three divergent microRNAs that have been determined in Figures 3.5.7-12, *miR-214*, *miR-424* and *miR-519d* again separated clearly from others. In addition to those, another two member cluster has arisen, *miR-195* and *miR-497* cluster (Figure 3.5.14).

In order to see the expression patterns of those microRNAs, cluster view representing those microRNAs in minimum cluster number has been investigated. In this trial, the minimum number that covers those microRNAs as separate clusters was 8. The case has been illustrated in Figure 3.5.15.

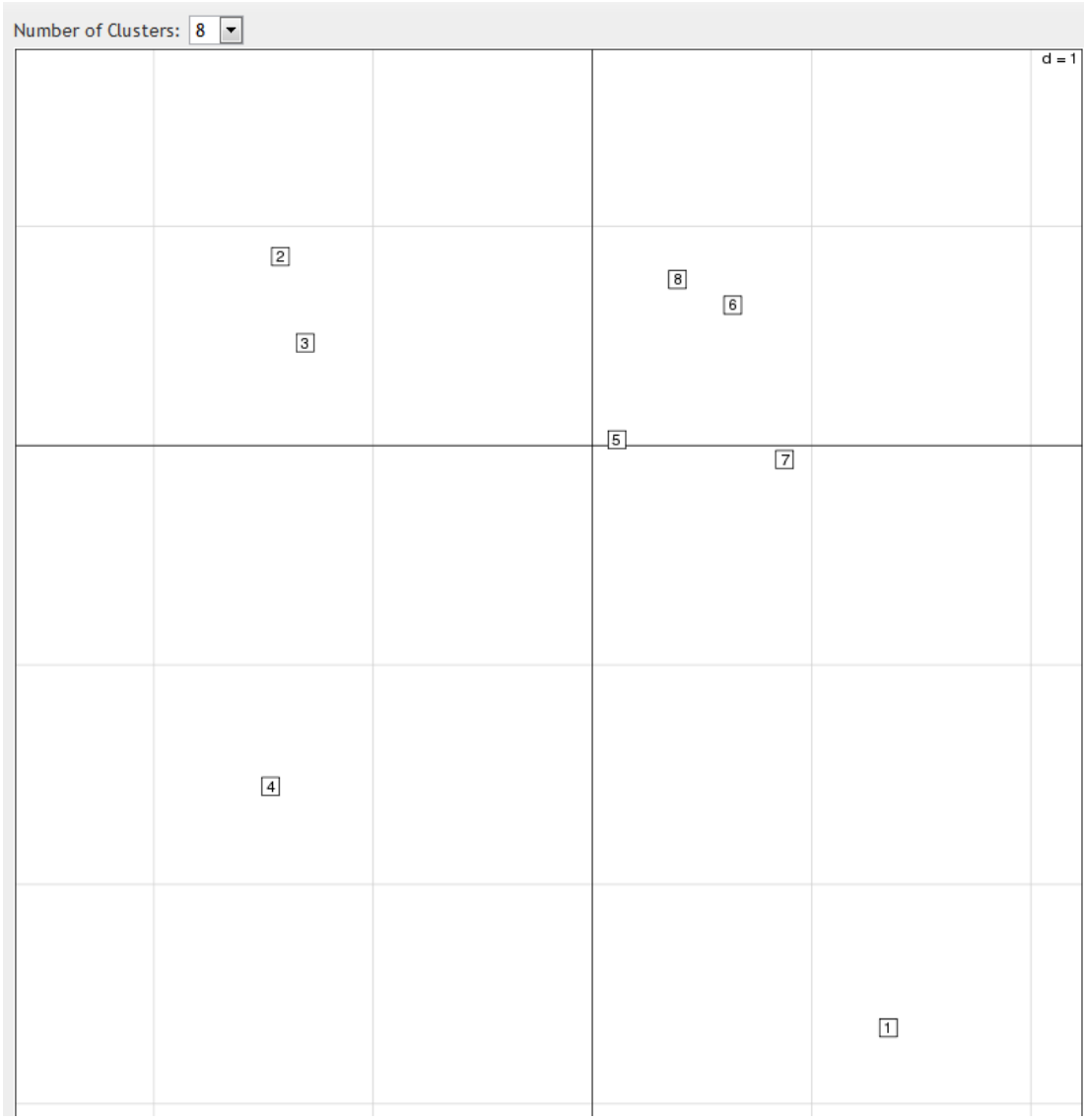


Figure 3.5.15: K-means cluster output view of the projections of the microRNA data points where K=8.

In this Figure (Figure 3.5.15), members of some cluster data points could easily be recognized when it is compared to Figure 3.5.14. For example, cluster number 1 represents the *miR-519d*. Accordingly, Cluster number 4 comprises only *miR-424* whereas cluster number 3 includes *miR-214*. Finally *miR-497* and *miR-195* have constituted the cluster 2 in this figure. As mentioned in section 3.2, those cluster data points are clickable and if they are clicked, the expression patterns of cluster members relative to out-of-cluster members are visualized in the datasets that are subjected to co-inertia analysis via 'expression & expression module'. So the facts

about earlier inferences indicating the relationships between *miR-214*-ovary and *miR-424*-testicle (Figure 3.5.10 and Figure 3.5.12) could be revealed by clicking on the respective cluster points to see the cluster specific expression patterns. Following figures (Figure 3.5.16, Figure 3.5.17, Figure 3.5.18 and Figure 3.5.19) show those relationships as expression barplots.

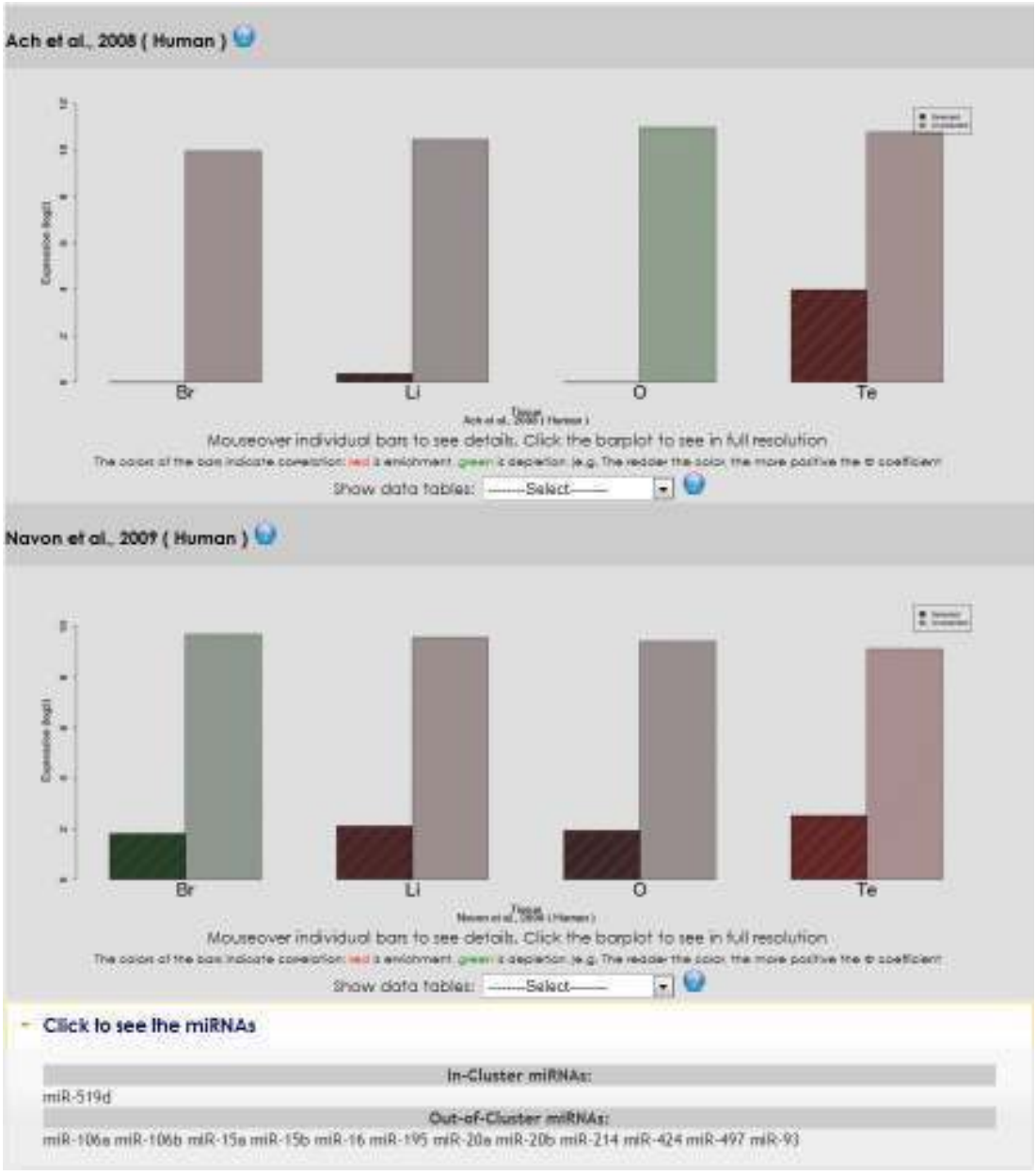


Figure 3.5.16: Expression pattern of cluster point number 1 (Figure 3.5.15).

From Figure 3.5.13 and Figure 3.5.14 it is observed that *miR-424* and *miR-*

519d are relatively more expressed in testes than any other microRNA represented in the Figure 3.5.14. Thus expression pattern seen in Figure 1.3.27 is expected as an output of expression analysis of cluster 1 seen in Figure 3.5.15.

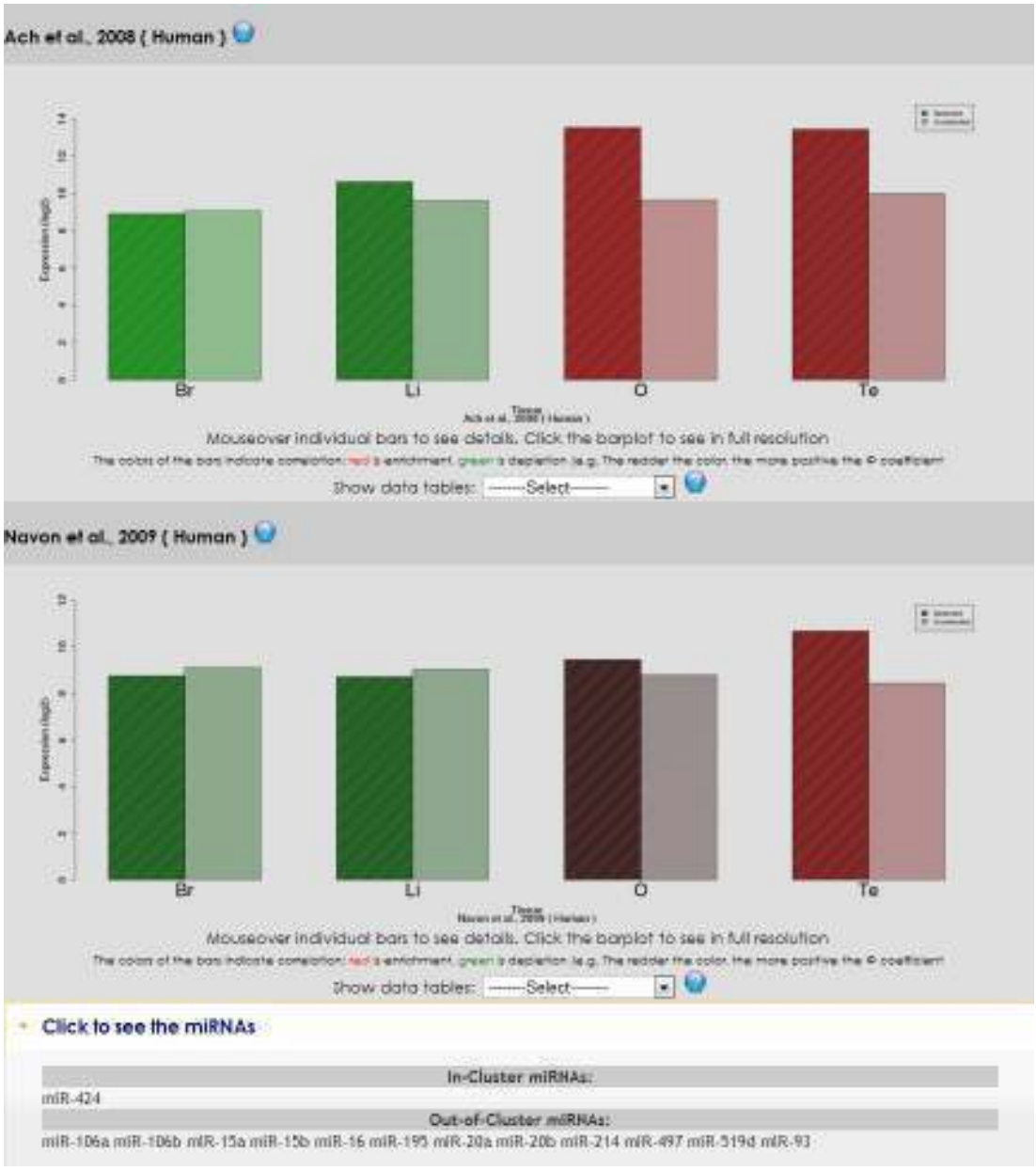


Figure 3.5.17: Expression pattern of cluster point number 4 (Figure 3.5.15).

This figure has been generated via mESAdb by clicking the cluster number 4 illustrated in the Figure 3.5.15. Previous observations from Figure 3.5.10, Figure 3.5.11 and Figure 3.5.12 stating that *miR-424* expression has been related to testes

was supported by this figure (Figure 3.5.17) as it could be done by comparing the locations of tissues and microRNAs from Figure 3.5.13 and Figure 3.5.14. In those, it is again seen that location of testes corresponded to the location of *miR-424*.

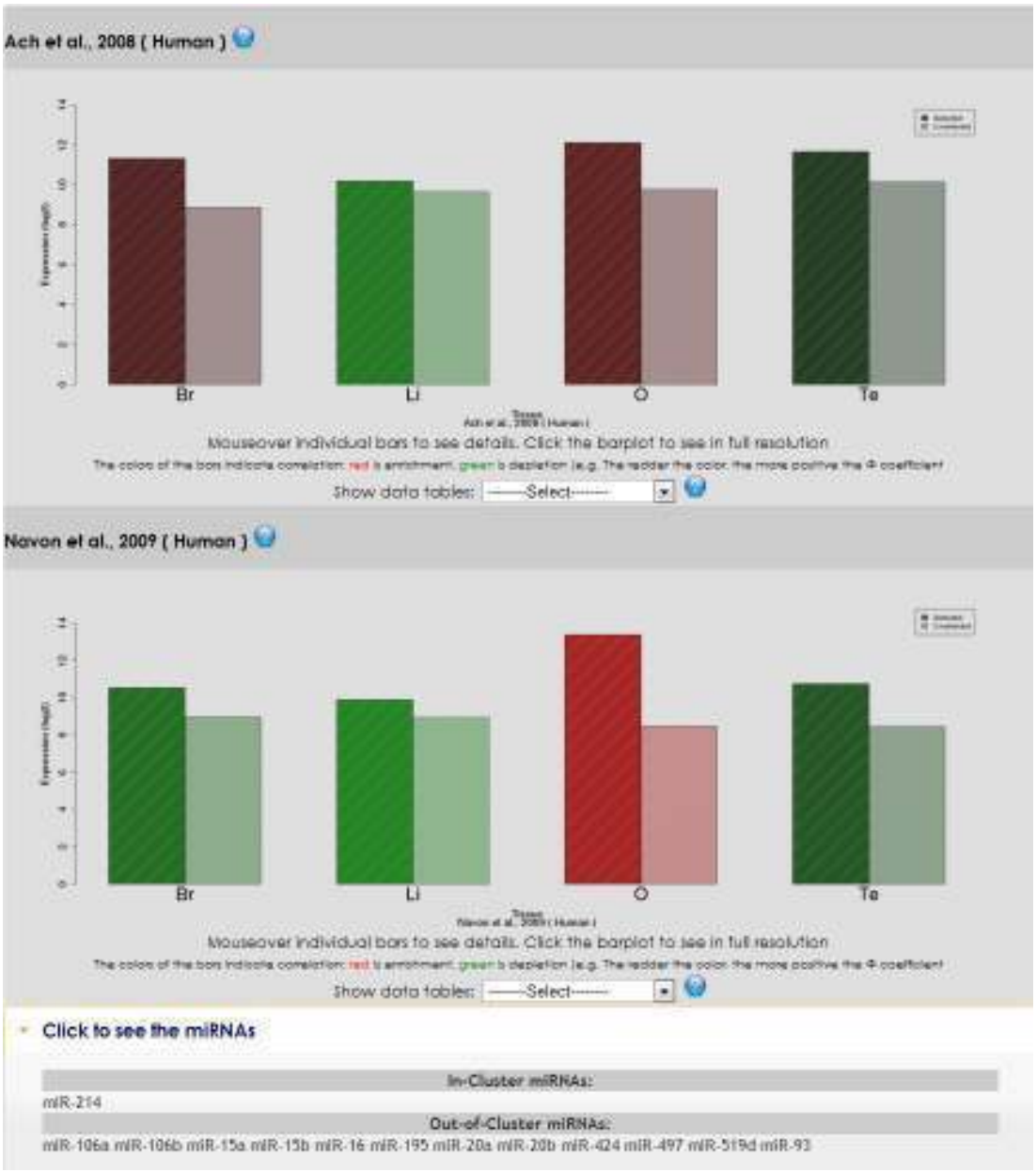


Figure 3.5.18: Expression pattern of cluster point number 3 (Figure 3.5.15).

In a similar way, Figure 3.5.18 shows the *miR-214*-ovary relation as Figure 3.5.17 did between *miR-424* and testes. Again, they were one of the consistent co-localized microRNA-tissue couples that have been observed in Figure 3.5.10, Figure

3.5.11 and Figure 3.5.12.

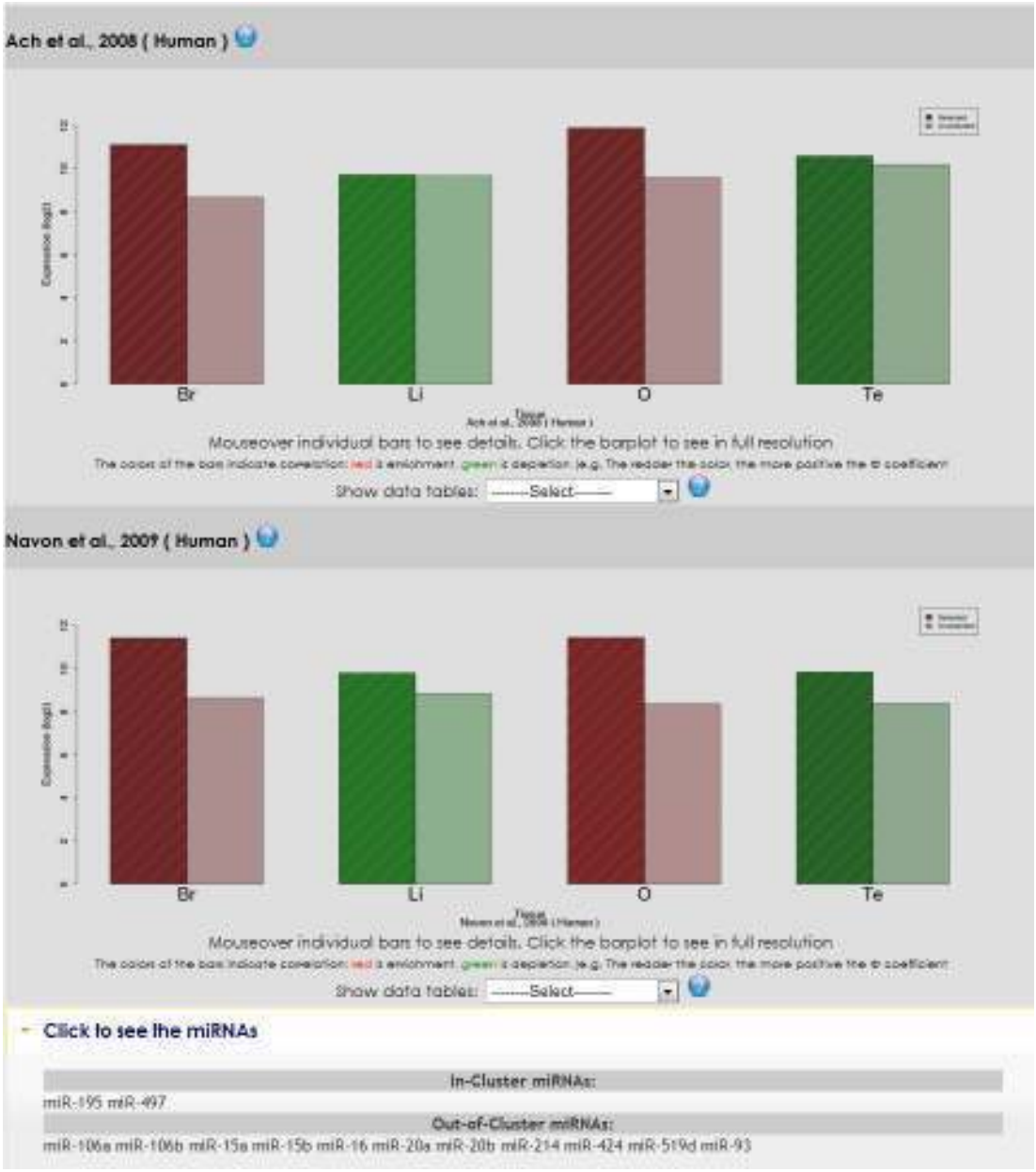


Figure 3.5.19: Expression pattern of cluster point number 2 (Figure 3.5.15).

Restricting the co-inertia analysis to the common tissues between the two datasets created another microRNA cluster that is separated from the group seen as a bulk accumulated close to the origin in the +/+ area of the axes illustrated in the Figure 3.5.14. Comparison of locations of cluster number 2 and locations of tissues from Figure 3.5.15 and Figure 3.5.13 respectively reveals that location of the new

cluster is closer to breast tissue than any other tissues in Figure 3.5.13 where those figures are the outputs of the same co-inertia analysis. Figure 3.5.19 supports those observations. Furthermore, the very short distance between the clusters 2-3 implies that members of both clusters could be enriched in the ovary. This figure supports also this observation. We see red bars only in two tissues in Figure 3.5.19, breast and ovary.

CHAPTER 4:DISCUSSION

mESAdb is a meta-analysis tool based on a database which has been built via integration of a multitude of resources relevant to microRNA biology. Main features of mESAdb can be listed as followings: I-) User specific high-throughput microRNA expression data upload utility; II-) Comparisons of two high-throughput microRNA expression datasets via coinertia analysis; III-) Sequence-expression relationship analysis for microRNA lists; IV-) Expression analysis for one or a group of microRNAs; V-) Functional annotation for one or a group of microRNAs via *GO*, *KEGG* and *HuGE* databases based on their targets stored in *MicroCosm*.

For a meta-analysis tool, providing user specific data upload is an important feature. First of all, users can re-annotate the microRNA probe sequences according to up to date miRBase (Ambros, Bartel et al. 2003; Griffiths-Jones 2004; Griffiths-Jones, Grocock et al. 2006; Griffiths-Jones, Saini et al. 2008; Kozomara and Griffiths-Jones 2011) nomenclature while uploading the data. It has been made possible to normalize the uploaded data as requested using quantile normalization. Similarly, linearization of illumination intensity data via log transformation also is made possible. Those novel features enable comparison of the uploaded data with the existing ones in the database in a uniform way.

Uploaded datasets can be analyzed in different manners: For example if a cancer dataset is uploaded such as the one exemplified in Section 3.1.1 and Figures 3.1.1-3, mESAdb could help test the similarity among different patient samples in terms of the tissue specific microRNA expression using multivariate analyses such as correspondence analysis and co-intertia analysis. In the case study exemplified here in the Section 3.1 By selecting 'AAGTGC' seed motif, found to be human embryonic stem cell specific by Laurent et al., 2008, mESAdb has separated most of the samples with hepatocellular carcinoma from others as clusters (Figure 3.1.2). This is actually plausible since earlier it has been reported that cancer cells have a hierarchy of organized groups of malignancies in which only a rare subset of cancer cells drives the tumor growth (Brennecke, Stark et al. 2005). Nevertheless CSCs (Cancer Stem Cells) have also been described in HCC (Krek, Grun et al. 2005). None

of the microRNAs listed in Table 3.1.1 has been reported as liver cancer stem cell specific miccroRNAs in the literature, although a comprehensive study has been done on this subject recently (Lewis, Burge et al. 2005). According to the distribution of projections of the microRNAs in the Table 3.1.1 on the same space on which samples from hcc patients and normal livers have been projected (Figure 3.1.3) except the *miR-520b* and *miR520f*, all microRNAs are co-localized with different hcc clusters.

Though, possible clusters of samples might also be associated with the clinical annotation of samples to deduce whether tissue expression patterns of the selected microRNAs are correlated with a given set of patients or controls. In that case we could investigate the direction of projections of expression values of microRNAs like it has been done for across species dataset (Figure 3.2.2).

To solidly demonstrate this aspect of mESAdb is to enable comparative analysis of microRNA expression patterns across taxa. To exemplify in the results section we compared a mouse and a human dataset (Thomson, Parker et al. 2004; Meiri, Levy et al. 2010) (Figure 3.2.1). The microRNA list was composed of let-7a-i, mir-130a-b, mir-15a-b, mir-181a-b, mir-200a-b, mir-23a-b, mir-26a-b, mir-29a-c, mir-30a-d and mir-99a-b clusters (Table 3.2.1, Table 3.2.2 and Table 3.2.3) and expression was studied across taxa for the brain (B), liver (Li), lung (Lu), kidney (K) and heart (H). The high RV coefficient, ~0.730, indicated that the mouse and human expression data for the abovementioned tissues and microRNAs were highly comparable and similar. Interestingly, the RV coefficient for the selected microRNAs is much higher than the one, generated by the selection of all available microRNAs with the same tissue set (~0.464; not reported earlier). This suggests that selected microRNAs could be functionally more conserved between the two taxa.

Among the conserved microRNAs, one pair is *miR*-181a and *miR-181b*. In the literature, both mir-181a and mir-181b were shown to be tumor suppressors in glioma where mir-181b were more effective than the former (Shi, Cheng et al. 2008). mESAdb analysis has shown that in both mice and humans, *miR-181a* and *miR-181b* were more abundant in brain and also in lung than in the other tissues. This tissue specificity is in support of their function in human gliomas and our findings suggest

that these microRNAs might also be involved in tumor suppression of mouse brain cancers. *miR-181a* also was recently implicated in the prognosis of non-small cell lung cancer (Gao, Yu et al. 2010).

mESAdb phi-coefficient analysis of *miR-200a* and *mir-200b* in these mouse and human expression datasets also revealed that these two microRNAs have very highly similar expression patterns as expected since they have similar sequences. Their higher expression in Kidney, Liver and Lung but lower expression in Brain and Heart suggested that they exhibited epithelial cell specificity. Indeed, previous research has shown this was the case (Bracken, Gregory et al. 2008). The findings in addition showed that their expression was similar also among taxa, suggesting results obtained from animal models for *miR-200a-b* might be extended to human pathogenesis. In recent literature, *miR-200a* and *miR-200b* were up-regulated simultaneously in a rat model of diet-induced nonalcoholic fatty liver disease (NAFLD) (Alisi, Da Sacco et al. 2011). Since expression pattern of *miR-200a* and *miR-200b* in mouse and human is highly conserved, these two microRNAs also might be involved in the pathogenesis of human NAFLD. Furthermore, *miR-200a* and *miR-200b* were shown to be up-regulated intrarenally in hypertensive nephrosclerosis, in accord with the expression of *miR-200a-b* in kidneys of mice and humans (Wang, Kwan et al. 2010).

mESAdb as a meta-analysis tool and database focuses also on discovery of associations between microRNA sequence and expression. mESAdb is advantageous because it allows interactive analysis of selected subsets of microRNAs in addition to analysis of single microRNA types. The sequence and expression relation is important because microRNAs exert their functions through their seed sequences, generally between 2^{nd} and 7^{th} bases from 5' end of the mature RNA, through binding multiple sequences located in 3' UTR or any other part of mRNA of the target gene to be silenced via base complementation. This base complementation is partial most of the time and same target sequences could be found in thousands of genes. So a motif that describes those target sequences might point to a systemic expression profile change. Such an important relation can be surveyed through mESAdb in two ways: 1) For instance, as mentioned in the result section, *miR-181a* and *miR-181b*,

similar in their mature sequences differing only in 3 nucleotides, (Table 3.2.3) exhibit a common sequence motif (i.e. AACATTCA) in their first 8 nucleotides. The mature sequences *miR-200a* and *miR-200b* are also containing a common motif (i.e. TAA[C][T]ACTG) in their first 8 nucleotides (Table 3.2.3). Those a-b couples are very close to each other in both genomes (Table 3.2.1 and Table 3.2.2). This may be the reason of their expression similarity seen not only in the Figure 3.2.2 but also seen in the Figure 3.2.3.

A broader set of examples of these aforementioned cases and more have been illustrated in section 1.3.5 on CHRNA5 targeting microRNAs. There are increasing number of studies that relate nicotine and estrogen (Biegon, Kim et al. 2010; Lee, Chang et al. 2010; Si, Long et al. 2010; el-Mas, el-Gowilly et al. 2011; Kandi and Hayslett 2011; Wang, Zhao et al. 2011; Yararbas and Pogun 2011). Also there has been a study indicating that a SNP in CHRNA5 gene, *rs16969968*, significantly alters the nicotine dependence of the smoker (Sherva, Wilhelmsen et al. 2008). Furthermore, it has been shown that chromosome 15q24-25.1 is a significant lung cancer susceptibility region in which CHRNA5, CHRNA3 and CHRNB4 genes are allocated. In the results section, the relationship between CHRNA5 targeting microRNAs and estrogen receptor expressing breast cancer samples has been assessed using the GEO dataset GSE15885 (Lowery, Miller et al. 2009).

Correspondence analysis has indicated that *miR-15a*, *miR-15b* and *miR-16*, (Figure 3.5.1, Figure 3.5.2 and Figure 3.5.3) were highly expressed in ER positive samples than they were in ER negative although these three microRNAs did not explain the variability 100%. Indeed, *miR15a*, *miR-5b* and *miR-16*, sharing common seed sequence cluster together, hence potentially target the same genes. All in all, it has been shown that miR-15 / 16 family targets BCL-2 (Cimmino, Calin et al. 2005), which is involved in regulation of apoptosis and thus might generate prognostic differences between ER negative and ER positive breast cancer patients.

In the CHRNA5 example, it has been used two groups of microRNAs, each with a common seed sequence within the group (Figure 3.5.5: First part of the alignment that shows which microRNAs bind to which part of CHRNA5 mRNA. and Tables 1.3.4-5). One of the groups has a binding side starting from 524th

nucleotide of the mRNA, whereas the other group has one starting at 800th base. These two groups did not show a consistent and tight separation in terms of expression variability in the reduced tissue space throughout different datasets (Figure 3.5.7, Figure 3.5.8 and Figure 3.5.9). However three of the microRNAs that target CHRNA5 were found to be consistently divergent in expression based on the correspondence analysis results. Those microRNAs were *miR-214, miR-424* and *miR-519d* (Figure 3.5.7, Figure 3.5.8, Figure 3.5.9, Figure 3.5.10, Figure 3.5.11, Figure 3.5.12 and Figure 3.5.14). In Figure 3.5.10 and Figure 3.5.12, it could be seen that *miR-214* was coupled with ovary and *miR-424* was coupled with testes. Co-inertia analysis based on Ach et al., 2008 and Navon et al., 2009, showed that the above relationship was consistent across datasets (Figure 3.5.18 and Figure 3.5.19). Also, according to the Ach dataset *miR-519d* was found to be placenta specific (Figure 3.5.10). In addition, two other microRNAs, *miR-195* and *miR*-497 were shown to be breast and ovary specific (Figure 3.5.19).

Accordingly, use of mESAdb has allowed identifying the CHRNA5 targeting microRNAs that show tissue specificity. This feature of mESAdb can help researchers to pinpoint a smaller subset of microRNAs among a large number of potential candidates. In fact, a new study suggested that there are 8 microRNA genes regulating p21 on the 19q13 in human genome where the largest human microRNA cluster exist with 46 members (Wu, Huang et al. 2010). They also state that it is the combined action of the eight microRNAs that is regulating p21 gene efficiently. They also identified other microRNAs through screening of the ones predicted by numerous microRNA prediction tools. By summarizing these facts, it has been reported in a review that the combination of microRNA targeting same gene should be studied instead of studying one microRNA-one mRNA connection (Peter 2010). Similarly the consistent tissue pairing shown by these 5 microRNAs targeting CHRNA5 on either of the two regions may reflect this fact. Hence these relations mentioned in the result chapter were surveyed in the literature.

The most studied microRNA among them is *miR214* (Flynt, Li et al. 2007; Li, Flynt et al. 2008; Yang, Kong et al. 2008; Chan, Yue et al. 2009; Juan, Kumar et al. 2009; Yang, Chen et al. 2009; Chen, Shalom-Feuerstein et al. 2010; Jindra,

Bagley et al. 2010; Juan and Sartorelli 2010; Liao, Du et al. 2010; Liu, Luo et al. 2010; Zhang, Ye et al. 2010; Bar-Eli 2011; Denby, Ramdas et al. 2011; Feng, Cao et al. 2011; Narducci, Arcelli et al. 2011; Penna, Orso et al. 2011; Qiang, Wang et al. 2011; Sehic, Risnes et al. 2011) with 19 publications. One study shows that *miR-214* promotes cell survival and resistance to cisplatin by targeting the PTEN hence by activating the Akt pathway in ovarian cancer (Yang, Kong et al. 2008). This evidence is in concordance with the tissue specificity of *miR-214* in ovary and to some extent in breast. Another study has shown that *miR-214* regulates the expression of a protein, lactoferrin (lf), which is abundant in human milk in mammary epithelial cells (Liao, Du et al. 2010). In the same study it has been also shown that as the amount of miR-214 inhibitor is increased in MCF7 cells; activity of caspase3 and lf amount is relatively elevated.

The literature search for the members of newly emerged cluster consisting *miR-195-miR497* produced 11 hits (Soon, Tacon et al. 2009; Xu, Zhu et al. 2009; Gonsalves and Kalra 2010; Liu, Chen et al. 2010; Sekiya, Ogawa et al. 2010; Ujifuku, Mitsutake et al. 2010; Yin, Deng et al. 2010; Li, Zhao et al. 2011; Wang, Wang et al. 2011; Zhu, Yang et al. 2011; Zhu, Zhu et al. 2011). Accordingly, miR-497 targets BCL-2 hence triggers apoptosis (Yin, Deng et al. 2010; Zhu, Zhu et al. 2011). According to literature, *miR-195* has been defined mainly as a tumor suppressor gene (Xu, Zhu et al. 2009; Liu, Chen et al. 2010; Sekiya, Ogawa et al. 2010; Wang, Wang et al. 2011). Both *miR-195* and *miR-497* resides on 17p13.1 as a cluster and they have been declared as potential tumor suppressors previously (Flavin, Smyth et al. 2009). Furthermore, it has been shown that in breast cancer patients this cluster is down regulated (Li, Zhao et al. 2011). This may suggest that, although there is no reported tissue specificity for those microRNAs, they may function in breast tissue as anti-apoptotic fine tuning microRNAs. *miR-424* does not have any known tissue specificity. There are a couple of studies stating that *miR-424* is involved in monocyte differentiation (Rosa, Ballarino et al. 2007; Forrest, Kanamori-Katayama et al. 2010). Also it has been reported that down regulation of *miR-424* contributes to abnormal angiogenesis (Nakashima, Jinnin et al. 2010). Although *miR-424* is not the focus of any study relating it to testes, it has been

reported as differentially expressed between mature and immature testes of porcine (Luo, Ye et al. 2010). Putting everything together, a study that researches the function of *miR-424* in testes would be interesting.

mESAdb is modular. Its each module has a different function so far outputs only from the 'expression & expression' and 'motif & expression' modules have been discussed. Another interesting finding using mESAdb showed the potential of mESAdb to associate a group of micorRNAs with a group of disease using its 'motif & function' module. In the example given in Figure 3.3.3, liver specific microRNAs were searched against the *HuGE* database in terms of their targets. The finding that this set of microRNAs being significantly associated with mainly liver diseases (e.g., Hepatitis A-E) is striking and suggests that microRNA disease connection might be effectively resolved using an approach we used in the mESAdb. In connection with the liver specific microRNAs other diseases also showed up as significant, such as ClubFoot and Cleft palate. What could be the association of liver specific microRNAs with diseases of the cartilage/bone should be further assessed. Interestingly, in the literature, liver steatosis has previsouly been associated with multiple symptoms including club feet (Elliott, Meagher-Villemure et al. 1996).

What makes mESAdb unique is its ability to analyze user data together with default datasets via different modular aspects of expression analysis, such as histograms and multivariate analysis. Previous microRNA databases also have implemented different combinations of barplots and heatmaps (Betel, Wilson et al. 2008; Nam, Kim et al. 2008) but the inclusion of multivariate analysis tools such as correspondence and co-inertia analyses into microRNA expression analysis is novel among microRNA expression databases. For example, *miRGator* (Nam, Kim et al. 2008) represents tissue expression data in the form of barplots and boxplots. Similarly, mESAdb uses barplots but also visually illustrates selected and non-selected microRNAs allowing comparison between their expression patterns. mESAdb enables multiple tissue and microRNA selection prior to the analysis. These selections create relative p-values for the calculated tissue enrichments according to the selections, a case that could create different biological questions. Another commonly used resource *microRNA.org* (Betel, Wilson et al. 2008) enables

the sample addition while viewing the barplots or heatmap based on a java application that represent expression levels for each microRNAs; however, it does not allow sequence based microRNA group selection. *microRNA.org* enables multiple selections if microRNA-target expression comparisons are requested. mESAdb also provides versatile selection of multiple microRNAs via either its intrinsic selection tools or file upload. In mESAdb, graphical representations are not dynamic as it is in *microRNA.org* however they are R based. Accordingly, graphical representation is not limited to barplots and heatmaps, and enables viewing tissues and microRNAs on reduced dimension with two axes using MADE 4 package.

Neither *miRGator* nor *microRNA.org* enables expression comparison across species. This is a feature unique to mESAdb currently. This is made possible by using dimension reduction methods provided by made4 R package (Culhane, Thioulouse et al. 2005). On-the-fly application of K-means algorithm to the mean of the projected coordinates of microRNAs is also a unique feature of mESAdb. Additionally, mESAdb allows for high-throughput user data upload, an advantageous feature allowing for comparison of user specified data with existing datasets. And to our knowledge, mESAdb is currently the only database that has integrated *HuGE phenopedia* and *genopedia* to current microRNA data warehouse.

In summary, versatile input and output methods provided by mESAdb allows production of flexible biological questions.

CHAPTER 5:FUTURE EXTENSIONS

5.1 MESADB

Modular nature of mESAdb allows for incorporation of additional data sets and statistical tools. Regarding this flexibility, future extensions to mESAdb will start with building a function that could automatically add microarray data sets from GEO. It is also planned that client specific on-the-fly class renaming will be possible for uploaded datasets regardless of the source.

Classification of uploaded datasets according to human pathogenesis is also planned. Especially, inclusion of a group header particularly focusing on cancer has a priority in an improvement of this kind. The infrastructure will be based on keywords constructing the MIAME (Brazma, Hingamp et al. 2001) standards. It is also planned to analyze a given set of cancer related microRNA datasets using graph properties and meta-analysis tools as exemplified in the following section on ARC.

One of the main features that makes the mESAdb unique among the microRNA expression databases is the possibility of selecting a subset of genes according to their sequence properties. An additional improvement is intended also in this area. Since there are non-conserved functional microRNA binding sites (Tay, Zhang et al. 2008) and new emphasis put on multiple microRNA control over one mRNA theory (Peter 2010), understanding which set of microRNAs play a role in which particular tissue(s) has become an important concern. In order to do that addition of predicted target specific microRNA lists (e.g., as done for CHRNA5 in the results section) to the microRNA selection list criteria is planned. While providing this, the option of selecting the target source, i.e., PicTar or miRanda will be available. In other words, the set of target microRNA association won't be restricted to Microcosm.

Use of R packages enhances the modular nature of the mESAdb thus future addition of statistical and visual tools for sequence/expression/function analysis of microRNAs also is planned. For example, incorporations of meta-analysis tools such

as metaGEM (Ramasamy, Mondry et al. 2008) and MAMA (Ihnatova 2010) will enable to user to apply meta analysis for more than two datasets.

5.2 AN EXTENSION OF THE FRAMEWORK USED IN MESADB TO OLIGONUCLEOTIDE DEALING WITH CANCERS: ARC

In this section we present an ongoing effort in combining mESAdb with another database, which we name ARC (Annotation and Regulation of Coexpression), focusing on mRNA microarray data analysis. This is important to increase the comprehensive nature of mESAdb and to be able to make associations between microRNAs and mRNA target expression. In this new framework ARC, uses a similar framework as mESAdb and microarray datasets focusing on human cancers, mostly breast and colon cancers, and a series of zebrafish tissues and experiments are used. The advantages of the framework used in mESAdb and ARC are: flexibility in terms of integration of new data types due to its modular structure; power of the statistical environment R in analyzing different types of data together and power of MySQL in storing various types of datasets; use of databases such as ENSEMBL, one of the most important genome browsers, that can be downloaded and installed as mysql databases; use of PHP and JavaScript to create tools easy to use and ones on which on-the-fly high throughput data analyses are possible.

Although an ongoing project, ARC will briefly be introduced herein and then the future perspectives for mESAdb and ARC will be presented. ARC aims to investigate a given set of gene probesets identifying a group of genes in different mRNA microarray datasets, i.e., from cancer patients or cancer cell lines treated with agents and zebrafish tissues and developmental series (e.g., cluster a group of pobesets; annotate a probeset with a function). ARC will serve to integrate similar human and zebrafish datasets and allow for simultaneous analysis across taxa as demonstrated in mESAdb. In the following two sections, the ARC modules shortly will be described.

5.2.1 Clustering module of ARC

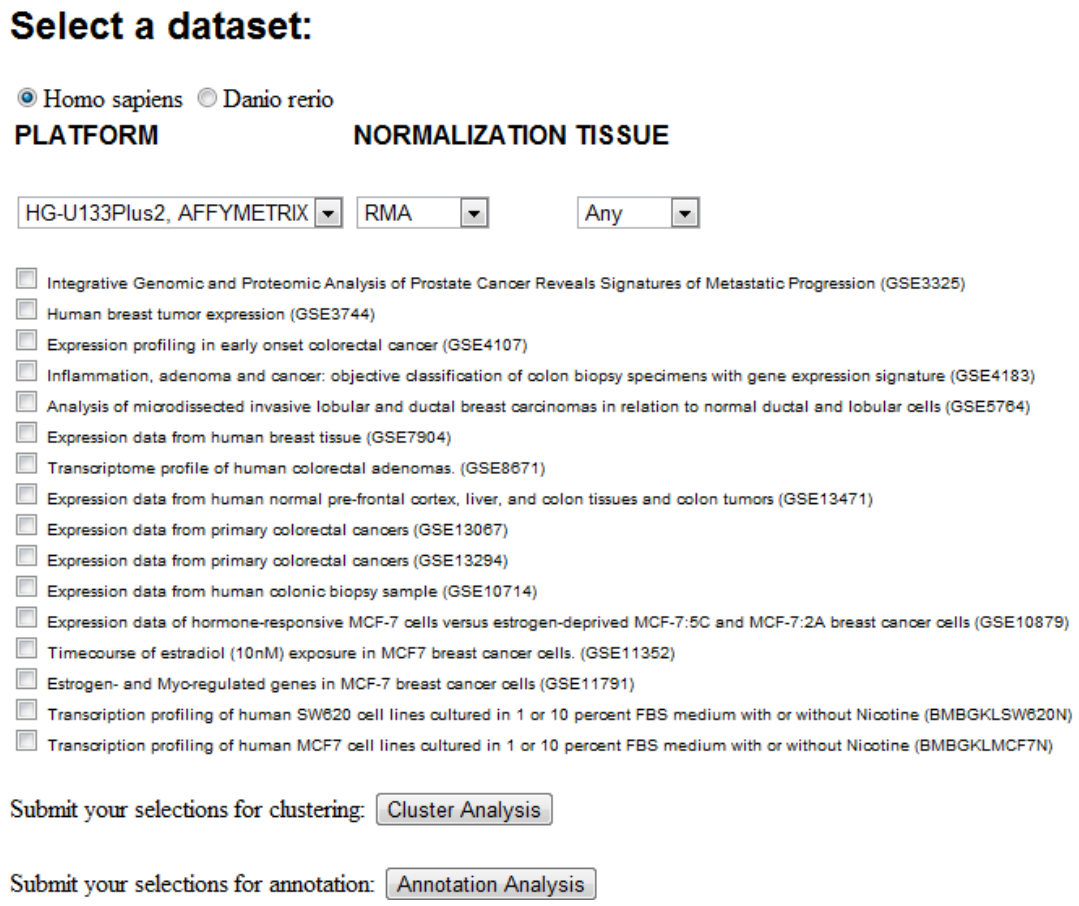


Figure 5.2.1: The view of the main page of tool developed for the analysis of oligo arrays.

The aim of the tool is to analyze/visualize a given set of mRNAs (probesets from Affymetrix platforms from human and zebrafish) selected by user, based on KEGG anotations or Huge Navigator terms. The main page of the tool (Figure 5.2.1**Hata! Başvuru kaynağı bulunamadı.**) allows for dataset selection based on four different criteria: species (i.e., human, zebrafish), array platform (i.e., HGU133A, HGU133 plus 2, Gene Chip Zebrafish), normalization (i.e., MAS5, RMA, GRSN COMBAT) and tissue type (i.e., breast, colon, pancreas, bladder). Selected datasets incorporated into this new database have raw datasets available, thus allowing for user-defined normalization. The selection criteria for normalization types seen on the main page of the tool allow for analysis of different datasets using different normalizations (Figure5.2.1**Hata! Başvuru kaynağı bulunamadı.**). The normalization types listed in the selection list on the main page are MAS5.0

(Affymetrix 2001), RMA (Irizarry, Hobbs et al. 2003), GRSN (Pelz, Kulesz-Martin et al. 2008) and ComBat (Johnson, Li et al. 2007).

The selections made in the main page are directed to two separate pipelines: cluster analysis and annotation analysis (Figure 5.2.1**Hata! Başvuru kaynağı bulunamadı.**). Both have a gene selection/entry bridge pages. For example, when all breast cancer datasets normalized with RMA are directed to cluster analysis pipe the gene entry page below appears.

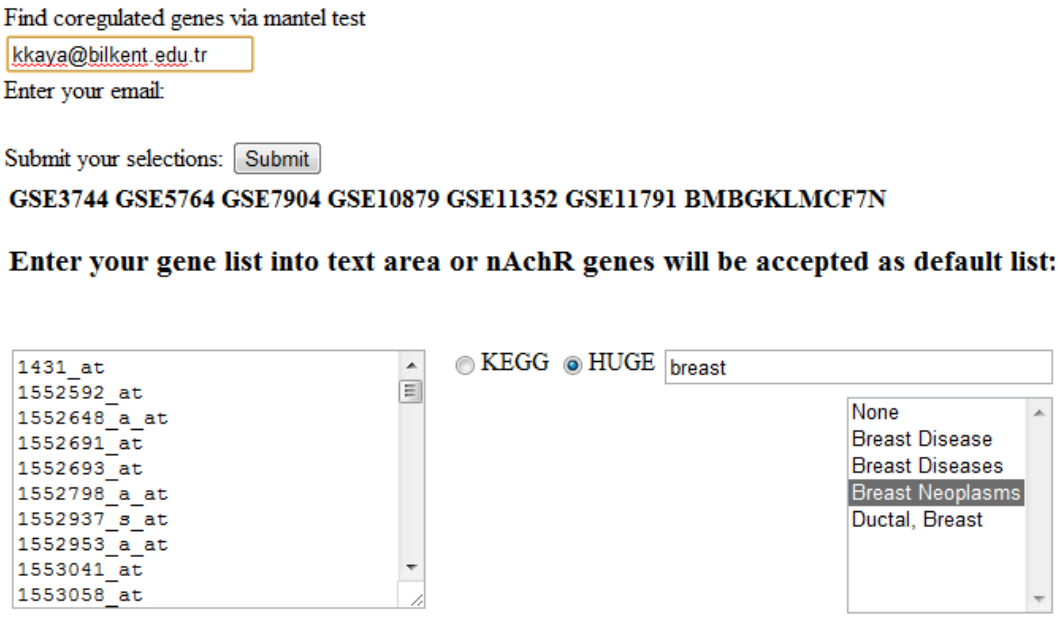


Figure 5.2.2: Snaphot of gene selection page at the beginning of cluster analysis pipe.

In the gene selection page (Figure 5.2.2**Hata! Başvuru kaynağı bulunamadı.**) there are four options to enter genes of interest. On the bottom left side of the page there is a text box, and two of the options are directly entering the genes into it either by their symbols or probeset names representing them. The remaining two options are automatic incorporation of probesets linked with either a *HuGE* term or a *KEGG* pathway into the text box area. Both *HuGE* and *KEGG* terms are long lists. Thus, to ease finding of desired term, a key word entry box has been linked to the lists. When a word is typed; the long list becomes short and this shortened list includes the words bearing only the key one entered. In the Figure 5.2.2**Hata! Başvuru kaynağı bulunamadı.**, this option has been illustrated by using

the keyword “breast”. The keyword eliminated to list to four huge terms, “Breast disease”, “Breast diseases”, “Breast Neoplasms” and “Ductal, Breast”. Among them the term “Breast Neoplasms” has been selected. Hence the gene entry box has automatically filled with the probests representing genes linked to “Breast Neoplasms” according to the *HuGE* database.

The output of the cluster analysis after submission is as illustrated in **Hata! Başvuru kaynağı bulunamadı.**.

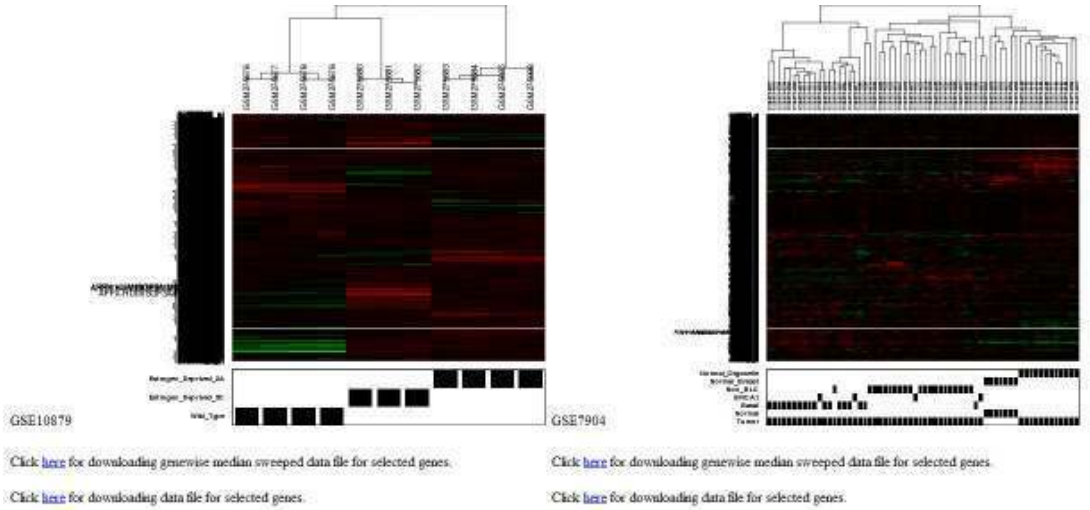


Figure 5.2.3: A snaphot of the cluster analysis output.

The heat maps and cluster figures generated by combination of *Cluster 3* of Eisen lab (Eisen, Spellman et al. 1998) and *Bioconductor* package *Heatplus* (Ploner).

5.2.2 Annotation Module of ARC

There are numerous genes in the human genome with known coordinates and sequences yet with no association to a metabolic or signaling pathway or functional terms (e.g., *KEGG* pathway). One of the powerful and *de novo* strategies to annotate a gene/probeset with a functional keyword has been to find expression neighbours of that particular gene and calculate the functional enrichment of the neighbouring gene set with respect to the whole genome/chip (Trupti Joshi ; Srivastava, Qiu et al. 2010). Examples for novel gene annotation include *PubLiME* using graph based approaches for co-occurring genes (Finocchiaro, Mancuso et al. 2007), *GlobalAncova* where gene sets are tested against others using general linear models (Hummel, Meister et

al. 2008), *IntelliGO* that uses new semantic indices for defining gene function (Benabderrahmane, Smail-Tabbone et al. 2010), disease candidate gene prioritization algorithms based on protein-protein interaction networks (Chen, Aronow et al. 2009; Chen, Yan et al. 2010). In the literature there are several examples: *TOPPgene* allows for association of gene lists with other gene lists using training and a test lists (Chen, Xu et al. 2007; Chen, Bardes et al. 2009). Other novel gene function predictors that use large scale multiple expression studies together with other information resources include *Genemania* (Montojo, Zuberi et al. 2010) where queried gene is associated with other using using a guilt-by-association approach and *TranscriptomeBrowser* that uses a clustering algorithm to extract transcription signatures (Lopez, Textoris et al. 2008), and *mSigDB* where a gene(s) could be associated with another gene list or functional term (Liberzon, Subramanian et al. 2011).

There is still however a need for a meta-analytic approach in gene prioritization and novel gene annotation in the context of conserved cancer genes (e.g., across taxa). Obtaining meta-indices for functional annotation across a series of related experiments increases the power of association obtained from a single study. In ARC, an approach has been used in which a given gene/probeset(s) was associated with a *KEGG* pathway term based on the ranked correlation between the sample similarity matrix of the queried gene(s) and that of the target gene set (i.e., *KEGG* pathway set). A modified *RandProd* (Breitling, Armengaud et al. 2004) algorithm, which is a nonparametric meta-analysis method for microarray datasets (Hong, Breitling et al. 2006), has been adopted to be able to associate an unannotated gene/probeset(s) with a *KEGG* pathway term (*KEGG* Pathway Maps, March 2011) across a given set of related experiments (Appendix 7.2, Table 7.2.1). Original *RankProd* by (Breitling, Armengaud et al. 2004) is based on ranking of fold differences between two classes. The meta-analysis version of the *RankProd* algorithm (Hong, Breitling et al. 2006) uses the products of the ranked fold differences across a given number of studies. A permutation based p-value is then assigned for testing significance.

In our study, the analysis is not restricted to two groups since no fold change needs to be calculated; instead a ranked association index is calculated based on the correlation between two matrices, each of which is made up of eucledian distances between samples based on either the expression values of the queried gene(s) or those of the targeted gene(s) (i.e., *KEGG* pathway set). The correlation values between the expression matrices of the queried and targeted gene sets are then inverse-ranked; the product of these ranks is used as the combined rank of the studies under investigation. A p-value is calculated based on 100 randomly shuffled data matrix (both KEGG indices and the experimental study indices are permuted) that contain the original inverse-ranked correlations. A list of KEGG terms and the associated rank products are provided as the output of this module; only those terms with a p-value less than 0.05 are shown. The correlations between the queried and targeted gene sets as well as their inverse ranks together with the associated p-values obtained from permutations also are provided as .xls files and are downloadable. The tool is now available for looking up a probeset(s)/gene(s) present in several Affymetrix Gene expression Analysis profiles for human and zebrafish with a focus on mostly colon and breast cancer studies (Table 5.2.1). The user can select any number of the available datasets and associate a gene(s) with each of the maximum possible number of different KEGG terms available for the selected platform across selected studies and obtain also a short list of KEGG terms that are significantly associated with the given gene.

For example, it is possible to associate a gene with known functions in one biological process to see what other pathways this gene might be involved in. After selecting all breast cancer datasets pre-processed with RMA normalization, CHRNA5 was entered in to the gene box in the gene selection page for annotation analysis (Figure 5.2.4).

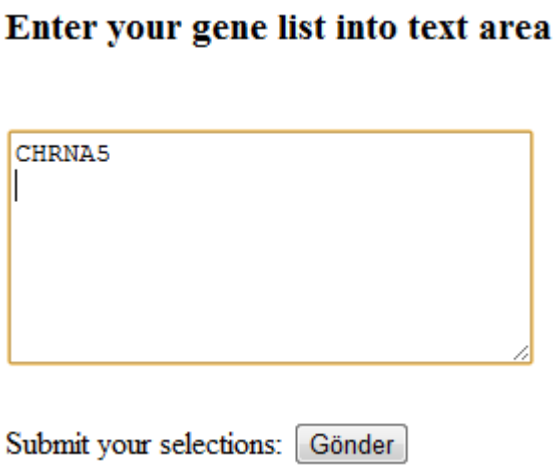
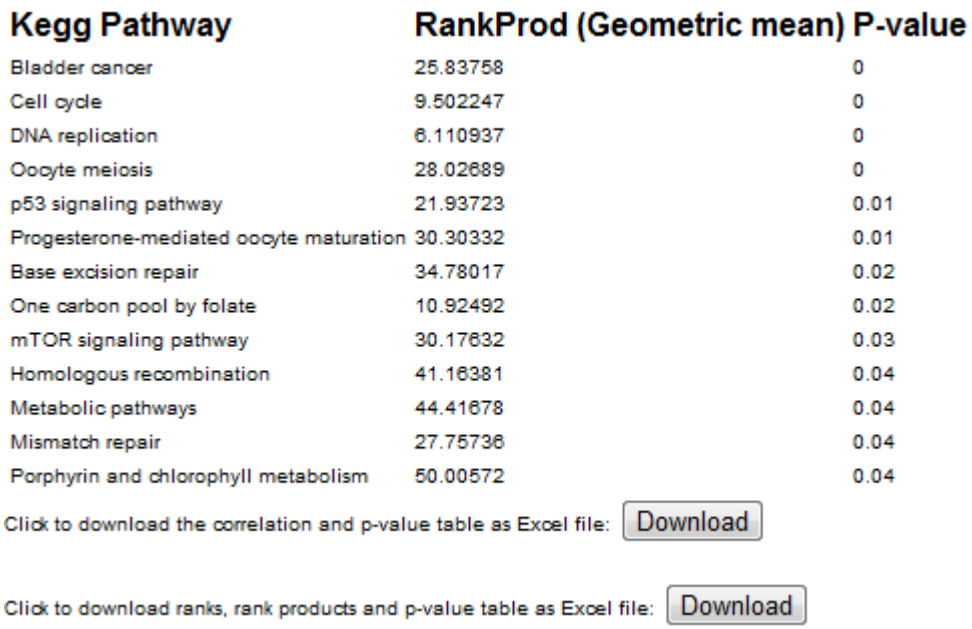


Figure 5.2.4: Gene selection page for annotation analysis.

Upon clicking the submit button, the probeset (or gene) entered is associated with KEGG pathways through a rank based comparison done based on expression values (see above for algorithm) across selected studies. This is a meta-analysis tool that allows for association of an unknown gene with a given set of lists, in this case the KEGG pathways lists. An example is given for CHRNA5 probeset. This gene's expression was mostly associated with ErbB signaling pathway. The output of such a submission has been illustrated in Figure 5.2.5.



Kegg Pathway	RankProd (Geometric mean)	P-value
Bladder cancer	25.83758	0
Cell cycle	9.502247	0
DNA replication	6.110937	0
Oocyte meiosis	28.02689	0
p53 signaling pathway	21.93723	0.01
Progesterone-mediated oocyte maturation	30.30332	0.01
Base excision repair	34.78017	0.02
One carbon pool by folate	10.92492	0.02
mTOR signaling pathway	30.17632	0.03
Homologous recombination	41.16381	0.04
Metabolic pathways	44.41678	0.04
Mismatch repair	27.75736	0.04
Porphyrin and chlorophyll metabolism	50.00572	0.04

Figure 5.2.5: A snapshot of annotation analysis result.

5.3 FUTURE PERSPECTIVES ON COMBINING MESADB WITH ARC

In the future zebrafish and human modules of ARC will be integrated and comparative analysis of zebrafish and human mRNA datasets will be possible for an evolutionarily conserved set of genes. Furthermore, the connection between mESAdb and ARC will be established. The advantages of being ARC and mESAdb interacted will then enable the user to study the mRNA and microRNA expression studies simultaneously and comparatively and also across different taxa. Furthermore, the annotation algorithm of the ARC can be applied to gene lists other than KEGG. One possibility is to include the gene lists that are targeted by a particular microRNA. Accordingly, it will be possible to test whether a set of genes targeted by a microRNA(s) are co-expressed and/or associated with a given gene list or functional term. In addition, the *MADE4* tools used in mESAdb will be applied to the framework in ARC to establish synchronization and potentially analyze microRNA expression patterns with mRNA expression patterns in a user-specified manner. Addition of new datasets into mESAdb and ARC are also planned to

comprehensively analyze cancer datasets. A full integration of zebrafish Affymetrix microarrays is also planned. We also aim to integrate the rank product based annotation module used in ARC and other related meta-analysis tools that exist in the literature into mESAdb framework. This addition will allow to obtain combined statistics on microRNA expression values across different studies for a given tissue or experimental class and will help generalize results from expression profile analyses performed in mESAdb.

CHAPTER 6:REFERENCES

(2003). "The International HapMap Project." Nature **426**(6968): 789-796.

(2004). "Integrating ethics and science in the International HapMap Project." Nat Rev Genet **5**(6): 467-475.

(2005). "A haplotype map of the human genome." Nature **437**(7063): 1299-1320.

Ach, R. A., H. Wang, et al. (2008). "Measuring microRNAs: comparisons of microarray and quantitative PCR measurements, and of different total RNA prep methods." BMC Biotechnol **8**: 69.

Affymetrix (2001). Statistical algorithms reference guide, Technical

report, Affymetrix.

Alexiou, P., M. Maragkakis, et al. (2009). "Lost in translation: an assessment and perspective for computational microRNA target identification." Bioinformatics **25**(23): 3049-3055.

Alimonti, A., A. Carracedo, et al. (2010). "Subtle variations in Pten dose determine cancer susceptibility." Nat Genet **42**(5): 454-458.

Alisi, A., L. Da Sacco, et al. (2011). "Mirnome analysis reveals novel molecular determinants in the pathogenesis of diet-induced nonalcoholic fatty liver disease." Lab Invest **91**(2): 283-293.

Altshuler, D. M., R. A. Gibbs, et al. (2010). "Integrating common and rare genetic variation in diverse human populations." Nature **467**(7311): 52-58.

Alvarez-Garcia, I. and E. A. Miska (2005). "MicroRNA functions in animal development and human disease." Development **132**(21): 4653-4662.

Ambros, V., B. Bartel, et al. (2003). "A uniform system for microRNA annotation." RNA **9**(3): 277-279.

Aravin, A. A., M. Lagos-Quintana, et al. (2003). "The small RNA profile during Drosophila melanogaster development." Dev Cell **5**(2): 337-350.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Bar-Eli, M. (2011). "Searching for the 'melano-miRs': miR-214 drives melanoma metastasis." EMBO J **30**(10): 1880-1881.

Barad, O., E. Meiri, et al. (2004). "MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues." Genome Res **14**(12): 2486-2494.

Bargaje, R., M. Hariharan, et al. (2010). "Consensus miRNA

expression profiles derived from interplatform normalization of microarray data." RNA **16**(1): 16-25.

Barrett, T., D. B. Troup, et al. (2011). "NCBI GEO: archive for functional genomics data sets--10 years on." Nucleic Acids Res **39**(Database issue): D1005-1010.

Barrett, T., D. B. Troup, et al. (2009). "NCBI GEO: archive for high-throughput functional genomic data." Nucleic Acids Res **37**(Database issue): D885-890.

Bartel, D. P. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell **116**(2): 281-297.

Bashirullah, A., A. E. Pasquinelli, et al. (2003). "Coordinate regulation of small temporal RNAs at the onset of Drosophila metamorphosis." Dev Biol **259**(1): 1-8.

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

Basyuk, E., F. Suavet, et al. (2003). "Human let-7 stem-loop precursors harbor features of RNase III cleavage products." Nucleic Acids Res **31**(22): 6593-6597.

Benabderrahmane, S., M. Smail-Tabbone, et al. (2010). "IntelliGO: a new vector-based semantic similarity measure including annotation origin." BMC Bioinformatics **11**: 588.

Betel, D., M. Wilson, et al. (2008). "The microRNA.org resource: targets and expression." Nucleic Acids Res **36**(Database issue): D149-153.

Beuvink, I., F. A. Kolb, et al. (2007). "A novel microarray approach reveals new tissue-specific signatures of known and predicted mammalian microRNAs." Nucleic Acids Res **35**(7): e52.

Biegon, A., S. W. Kim, et al. (2010). "Nicotine blocks brain estrogen synthase (aromatase): in vivo positron emission tomography studies in female baboons." Biol Psychiatry **67**(8): 774-777.

Binns, D., E. Dimmer, et al. (2009). "QuickGO: a web-based tool for Gene Ontology searching." Bioinformatics **25**(22): 3045-3046.

Blake, J. A., C. J. Bult, et al. (2011). "The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics." Nucleic Acids Res **39**(Database issue): D842-848.

Bolstad, B. M., R. A. Irizarry, et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-193.

Boutet, S., F. Vazquez, et al. (2003). "Arabidopsis HEN1: a genetic link between endogenous miRNA controlling development and siRNA controlling transgene silencing and virus resistance." Curr Biol **13**(10): 843-848.

Bracken, C. P., P. A. Gregory, et al. (2008). "A double-negative feedback loop between ZEB1-SIP1 and the microRNA-200 family regulates epithelial-mesenchymal transition." Cancer Res **68**(19): 7846-7854.

Brazma, A., P. Hingamp, et al. (2001). "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." Nat Genet **29**(4): 365-371.

Breitling, R., P. Armengaud, et al. (2004). "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments." FEBS Lett **573**(1-3): 83-92.

Brennecke, J., D. R. Hipfner, et al. (2003). "bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila." Cell **113**(1): 25-36.

Brennecke, J., A. Stark, et al. (2005). "Principles of microRNA-target recognition." PLoS Biol **3**(3): e85.

Chan, L. S., P. Y. Yue, et al. (2009). "Role of microRNA-214 in ginsenoside-Rg1-induced angiogenesis." Eur J Pharm Sci **38**(4): 370-377.

Chen, C. Z., L. Li, et al. (2004). "MicroRNAs modulate hematopoietic lineage differentiation." Science **303**(5654): 83-86.

Chen, H., R. Shalom-Feuerstein, et al. (2010). "miR-7 and miR-214 are specifically expressed during neuroblastoma differentiation, cortical development and embryonic stem cells differentiation, and control neurite outgrowth in vitro." Biochem Biophys Res Commun **394**(4): 921-927.

Chen, J., B. J. Aronow, et al. (2009). "Disease candidate gene identification and prioritization using protein interaction networks." BMC Bioinformatics **10**: 73.

Chen, J., E. E. Bardes, et al. (2009). "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization." Nucleic Acids Res **37**(Web Server issue): W305-311.

Chen, J., S. A. Carney, et al. (2008). "Comparative genomics identifies genes mediating cardiotoxicity in the embryonic zebrafish heart." Physiol Genomics **33**(2): 148-158.

Chen, J., H. Xu, et al. (2007). "Improved human disease candidate gene prioritization using mouse phenotype." BMC Bioinformatics **8**: 392.

Chen, X., G. Y. Yan, et al. (2010). "A novel candidate disease genes prioritization method based on module partition and rank fusion." OMICS **14**(4): 337-356.

Chen, X. M. (2009). "MicroRNA signatures in liver diseases." World J Gastroenterol **15**(14): 1665-1672.

Cherry, J. M., C. Adler, et al. (1998). "SGD: Saccharomyces Genome Database." Nucleic Acids Res **26**(1): 73-79.

Chhabra, R., R. Dubey, et al. (2010). "Cooperative and individualistic functions of the microRNAs in the miR-23a~27a~24-2 cluster and its implication in human diseases." Mol Cancer **9**: 232.

Cimmino, A., G. A. Calin, et al. (2005). "miR-15 and miR-16 induce apoptosis by targeting BCL2." Proc Natl Acad Sci U S A **102**(39): 13944-13949.

Cramér, H. (1946). Mathematical methods of statistics. Princeton,, Princeton university press.

Crescenzi, M. and A. Giuliani (2001). "The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data." FEBS Lett **507**(1): 114-118.

Culhane, A. C., G. Perriere, et al. (2003). "Cross-platform comparison and visualisation of gene expression data using co-inertia analysis." BMC Bioinformatics **4**: 59.

Culhane, A. C., J. Thioulouse, et al. (2005). "MADE4: an R package for multivariate analysis of gene expression data." Bioinformatics **21**(11): 2789-2790.

Denby, L., V. Ramdas, et al. (2011). "miR-21 and miR-214 Are Consistently Modulated during Renal Injury in Rodent Models." Am J Pathol.

Drew, R. E., K. J. Rodnick, et al. (2008). "Effect of starvation on transcriptomes of brain and liver in adult female zebrafish (Danio rerio)." Physiol Genomics **35**(3): 283-295.

Durinck, S., Y. Moreau, et al. (2005). "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis." Bioinformatics **21**(16): 3439-3440.

Duursma, A. M., M. Kedde, et al. (2008). "miR-148 targets human DNMT3b protein coding region." RNA **14**(5): 872-877.

Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-210.

Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-14868.

el-Mas, M. M., S. M. el-Gowilly, et al. (2011). "Estrogen dependence of the renal vasodilatory effect of nicotine in rats: role of alpha7 nicotinic cholinergic receptor/eNOS signaling." Life Sci **88**(3-4): 187-193.

Elliott, A. M., K. Meagher-Villemure, et al. (1996). "Schinzel-Giedion syndrome: further delineation of the phenotype." Clin Dysmorphol **5**(2): 135-142.

Enright, A. J., B. John, et al. (2003). "MicroRNA targets in Drosophila." Genome Biol **5**(1): R1.

Farazi, T. A., H. M. Horlings, et al. (2011). "MicroRNA Sequence and Expression Analysis in Breast Tumors by Deep Sequencing." Cancer Res **71**(13): 4443-4453.

Fellenberg, K., N. C. Hauser, et al. (2001). "Correspondence analysis applied to microarray data." Proc Natl Acad Sci U S A **98**(19): 10781-10786.

Feng, Y., J. H. Cao, et al. (2011). "Inhibition of miR-214 expression represses proliferation and differentiation of C2C12 myoblasts." Cell Biochem Funct **29**(5): 378-383.

Finocchiaro, G., F. M. Mancuso, et al. (2007). "Graph-based identification of cancer signaling pathways from published gene expression signatures using PubLiME." Nucleic Acids Res **35**(7): 2343-2355.

Fish, J. E., M. M. Santoro, et al. (2008). "miR-126 regulates angiogenic signaling and vascular integrity." Dev Cell **15**(2): 272-284.

Flavin, R. J., P. C. Smyth, et al. (2009). "Potentially important microRNA cluster on chromosome 17p13.1 in primary peritoneal carcinoma." Mod Pathol **22**(2): 197-205.

Flicek, P., M. R. Amode, et al. (2011). "Ensembl 2011." Nucleic Acids Res **39**(Database issue): D800-806.

Flynt, A. S., N. Li, et al. (2007). "Zebrafish miR-214 modulates Hedgehog signaling to specify muscle cell fate." Nat Genet **39**(2): 259-263.

Forman, J. J. and H. A. Coller (2010). "The code within the code: microRNAs target coding regions." Cell Cycle **9**(8): 1533-1541.

Forman, J. J., A. Legesse-Miller, et al. (2008). "A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence." Proc Natl Acad Sci U S A **105**(39): 14879-14884.

Forrest, A. R., M. Kanamori-Katayama, et al. (2010). "Induction of microRNAs, mir-155, mir-222, mir-424 and mir-503, promotes monocytic differentiation through combinatorial regulation." Leukemia **24**(2): 460-466.

Frasor, J., A. Weaver, et al. (2009). "Positive cross-talk between estrogen receptor and NF-kappaB in breast cancer." Cancer Res **69**(23): 8918-8925.

Frazer, K. A., D. G. Ballinger, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851-861.

Friedman, R. C., K. K. Farh, et al. (2009). "Most mammalian mRNAs are conserved targets of microRNAs." Genome Res **19**(1): 92-105.

Galamb, O., F. Sipos, et al. (2008). "Diagnostic mRNA expression patterns of inflamed, benign, and malignant colorectal biopsy specimen and their correlation with peripheral blood results." Cancer Epidemiol Biomarkers Prev **17**(10): 2835-2845.

Galamb, O., S. Spisak, et al. (2010). "Reversal of gene expression changes in the colorectal normal-adenoma pathway by NS398 selective COX2 inhibitor." Br J Cancer **102**(4): 765-773.

Gao, W., Y. Yu, et al. (2010). "Deregulated expression of miR-21, miR-143 and miR-181a in non small cell lung cancer is related to clinicopathologic characteristics or patient prognosis." Biomed Pharmacother **64**(6): 399-408.

Gentleman RC, C. V., Bates MD and others (2004). "Bioconductor: Open software development for computational biology and bioinformatics." Genome Biology **5**: R80.

Giraldez, A. J., Y. Mishima, et al. (2006). "Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs." Science **312**(5770): 75-79.

Girard, M., E. Jacquemin, et al. (2008). "miR-122, a paradigm

for the role of microRNAs in the liver." J Hepatol **48**(4): 648-656.

Gonsalves, C. and V. K. Kalra (2010). "Endothelin-1-induced macrophage inflammatory protein-1beta expression in monocytic cells involves hypoxia-inducible factor-1alpha and AP-1 and is negatively regulated by microRNA-195." J Immunol **185**(10): 6253-6264.

Gregory, R. I. and R. Shiekhattar (2005). "MicroRNA biogenesis and cancer." Cancer Res **65**(9): 3509-3512.

Griffiths-Jones, S. (2004). "The microRNA Registry." Nucleic Acids Res **32**(Database issue): D109-111.

Griffiths-Jones, S. (2006). "miRBase: the microRNA sequence database." Methods Mol Biol **342**: 129-138.

Griffiths-Jones, S., R. J. Grocock, et al. (2006). "miRBase: microRNA sequences, targets and gene nomenclature." Nucleic Acids Res **34**(Database issue): D140-144.

Griffiths-Jones, S., H. K. Saini, et al. (2008). "miRBase: tools for microRNA genomics." Nucleic Acids Res **36**(Database issue): D154-158.

Grimson, A., K. K. Farh, et al. (2007). "MicroRNA targeting specificity in mammals: determinants beyond seed pairing." Mol Cell **27**(1): 91-105.

Grun, D., Y. L. Wang, et al. (2005). "microRNA target predictions across seven Drosophila species and comparison to mammalian targets." PLoS Comput Biol **1**(1): e13.

Guilford, J. (1941). "The phi coefficient and chi square as indices of item validity." Psychometrika **6**(1): 11-19.

Gyorffy, B., B. Molnar, et al. (2009). "Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples." PLoS One **4**(5): e5645.

Hartigan, J. A. and M. A. Wong (1979). "Algorithm AS 136: A K-Means Clustering Algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1): 100-108.

He, S., E. Salas-Vidal, et al. (2006). "Genetic and transcriptome characterization of model zebrafish cell lines." Zebrafish **3**(4): 441-453.

Heiden, T. C., C. A. Struble, et al. (2008). "Molecular targets of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) within the zebrafish ovary: insights into TCDD-induced endocrine disruption and reproductive toxicity." Reprod Toxicol **25**(1): 47-57.

Hertel, J., M. Lindemeyer, et al. (2006). "The expansion of the metazoan microRNA repertoire." BMC Genomics **7**: 25.

Hilsenbeck, S. G., W. E. Friedrichs, et al. (1999). "Statistical analysis of array expression data as applied to the problem of tamoxifen resistance." J Natl Cancer Inst **91**(5): 453-459.

Hindorff, L. A., P. Sethupathy, et al. (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proc Natl Acad Sci U S A **106**(23): 9362-9367.

Hofacker, I. L. (2003). "Vienna RNA secondary structure

server." Nucleic Acids Res **31**(13): 3429-3431.

Hong, F., R. Breitling, et al. (2006). "RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis." Bioinformatics **22**(22): 2825-2827.

Hong, Y., K. S. Ho, et al. (2007). "A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis." Clin Cancer Res **13**(4): 1107-1114.

Hotelling, H. (1933). "Analysis of a Complex of Statistical Variables Into Principal Components." Journal of Educational Psychology **24**: 417-441,498-520.

Houbaviy, H. B., M. F. Murray, et al. (2003). "Embryonic stem cell-specific MicroRNAs." Dev Cell **5**(2): 351-358.

Hsu, S. D., F. M. Lin, et al. (2011). "miRTarBase: a database curates experimentally validated microRNA-target interactions." Nucleic Acids Res **39**(Database issue): D163-169.

Hummel, M., R. Meister, et al. (2008). "GlobalANCOVA: exploration and assessment of gene group effects." Bioinformatics **24**(1): 78-85.

Ihnatova, I. (2010). "MAMA: Meta-Analysis of MicroArry." from http://CRAN.R-project.org/package=MAMA.

Irizarry, R. A., B. Hobbs, et al. (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics **4**(2): 249-264.

Irizarry, R. A., C. Ladd-Acosta, et al. (2009). "The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores." Nat Genet **41**(2): 178-186.

Iwama, H., T. Masaki, et al. (2007). "Abundance of microRNA target motifs in the 3'-UTRs of 20527 human genes." FEBS Lett **581**(9): 1805-1810.

Jackson, A. L., J. Burchard, et al. (2006). "Position-specific chemical modification of siRNAs reduces "off-target" transcript silencing." RNA **12**(7): 1197-1205.

Jiang, Q., Y. Wang, et al. (2009). "miR2Disease: a manually curated database for microRNA deregulation in human disease." Nucleic Acids Res **37**(Database issue): D98-104.

Jindra, P. T., J. Bagley, et al. (2010). "Costimulation-dependent expression of microRNA-214 increases the ability of T cells to proliferate by targeting Pten." J Immunol **185**(2): 990-997.

John, B., A. J. Enright, et al. (2004). "Human MicroRNA targets." PLoS Biol **2**(11): e363.

Johnson, A. D., R. E. Handsaker, et al. (2008). "SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap." Bioinformatics **24**(24): 2938-2939.

Johnson, N. L., S. Kotz, et al. (1992). Univariate Discrete Distributions. New York, Wiley.

Johnson, W. E., C. Li, et al. (2007). "Adjusting batch effects in

microarray expression data using empirical Bayes methods." Biostatistics **8**(1): 118-127.

Jollife, I. T. (2002). Principal Component Analysis. New York, Springer.

Jordan, S. D., M. Kruger, et al. (2011). "Obesity-induced overexpression of miRNA-143 inhibits insulin-stimulated AKT activation and impairs glucose metabolism." Nat Cell Biol **13**(4): 434-446.

Jorissen, R. N., L. Lipton, et al. (2008). "DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers." Clin Cancer Res **14**(24): 8061-8069.

Juan, A. H., R. M. Kumar, et al. (2009). "Mir-214-dependent regulation of the polycomb protein Ezh2 in skeletal muscle and embryonic stem cells." Mol Cell **36**(1): 61-74.

Juan, A. H. and V. Sartorelli (2010). "MicroRNA-214 and polycomb group proteins: a regulatory circuit controlling differentiation and cell fate decisions." Cell Cycle **9**(8): 1445-1446.

Kachitvichyanukul, V. and B. Schmeiser (1985). "Computer generation of hypergeometric random variates." Journal of Statistical Computation and Simulation **22**(2): 127 - 145.

Kandi, P. and R. L. Hayslett (2011). "Nicotine and 17beta-estradiol produce an antidepressant-like effect in female ovariectomized rats." Brain Res Bull **84**(3): 224-228.

Kanehisa, M. (1997). "A database for post-genome analysis." Trends Genet **13**(9): 375-376.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res **28**(1): 27-30.

Kanehisa, M., S. Goto, et al. (2010). "KEGG for representation and analysis of molecular networks involving diseases and drugs." Nucleic Acids Res **38**(Database issue): D355-360.

Kawasaki, H. and K. Taira (2003). "Hes1 is a target of microRNA-23 during retinoic-acid-induced neuronal differentiation of NT2 cells." Nature **423**(6942): 838-842.

Kawasaki, H. and K. Taira (2003). "Retraction: Hes1 is a target of microRNA-23 during retinoic-acid-induced neuronal differentiation of NT2 cells." Nature **426**(6962): 100.

Kaya, K. D., G. Karakulah, et al. (2007). "MicroRNA sequence and expression database." BMC Systems Biology **1**(Suppl 1): P29.

Kent, W. J., C. W. Sugnet, et al. (2002). "The human genome browser at UCSC." Genome Res **12**(6): 996-1006.

Kertesz, M., N. Iovino, et al. (2007). "The role of site accessibility in microRNA target recognition." Nat Genet **39**(10): 1278-1284.

Kily, L. J., Y. C. Cowe, et al. (2008). "Gene expression changes in a zebrafish model of drug dependency suggest conservation of neuro-adaptation pathways." J Exp Biol **211**(Pt 10): 1623-1634.

Kozomara, A. and S. Griffiths-Jones (2011). "miRBase: integrating microRNA annotation and deep-sequencing data." Nucleic Acids

Res **39**(Database issue): D152-157.

Krek, A., D. Grun, et al. (2005). "Combinatorial microRNA target predictions." Nat Genet **37**(5): 495-500.

Krichevsky, A. M., K. S. King, et al. (2003). "A microRNA array reveals extensive regulation of microRNAs during brain development." RNA **9**(10): 1274-1281.

Krol, J., I. Loedige, et al. (2010). "The widespread regulation of microRNA biogenesis, function and decay." Nat Rev Genet **11**(9): 597-610.

Lagos-Quintana, M., R. Rauhut, et al. (2001). "Identification of novel genes coding for small expressed RNAs." Science **294**(5543): 853-858.

Lagos-Quintana, M., R. Rauhut, et al. (2003). "New microRNAs from mouse and human." RNA **9**(2): 175-179.

Lagos-Quintana, M., R. Rauhut, et al. (2002). "Identification of tissue-specific microRNAs from mouse." Curr Biol **12**(9): 735-739.

Lai, E. C., P. Tomancak, et al. (2003). "Computational identification of Drosophila microRNA genes." Genome Biol **4**(7): R42.

Lau, N. C., L. P. Lim, et al. (2001). "An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans." Science **294**(5543): 858-862.

Laurent, L. C., J. Chen, et al. (2008). "Comprehensive microRNA profiling reveals a unique human embryonic stem cell signature dominated by a single seed sequence." Stem Cells **26**(6): 1506-1516.

Lee, C. H., Y. C. Chang, et al. (2010). "Crosstalk between nicotine and estrogen-induced estrogen receptor activation induces alpha9-nicotinic acetylcholine receptor expression in human breast cancer cells." Breast Cancer Res Treat.

Lee, R. C. and V. Ambros (2001). "An extensive class of small RNAs in Caenorhabditis elegans." Science **294**(5543): 862-864.

Lee, R. C., R. L. Feinbaum, et al. (1993). "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." Cell **75**(5): 843-854.

Lee, Y., C. Ahn, et al. (2003). "The nuclear RNase III Drosha initiates microRNA processing." Nature **425**(6956): 415-419.

Leung, Y. F., P. Ma, et al. (2007). "Gene expression profiling of zebrafish embryonic retinal pigment epithelium in vivo." Invest Ophthalmol Vis Sci **48**(2): 881-890.

Lewis, B. P., C. B. Burge, et al. (2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." Cell **120**(1): 15-20.

Lewis, B. P., I. H. Shih, et al. (2003). "Prediction of mammalian microRNA targets." Cell **115**(7): 787-798.

Lewis, S. E. (2005). "Gene Ontology: looking backwards and forwards." Genome Biol **6**(1): 103.

Li, D., Y. Zhao, et al. (2011). "Analysis of MiR-195 and MiR-

497 expression, regulation and role in breast cancer." Clin Cancer Res **17**(7): 1722-1730.

Li, L., J. Xu, et al. (2010). "Computational approaches for microRNA studies: a review." Mamm Genome **21**(1-2): 1-12.

Li, N., A. S. Flynt, et al. (2008). "Dispatched Homolog 2 is targeted by miR-214 through a combination of three weak microRNA recognition sites." Nucleic Acids Res **36**(13): 4277-4285.

Li, S. C., W. C. Chan, et al. (2010). "Discovery and characterization of medaka miRNA genes by next generation sequencing platform." BMC Genomics **11 Suppl 4**: S8.

Li, W., L. Xie, et al. (2008). "Diagnostic and prognostic implications of microRNAs in human hepatocellular carcinoma." Int J Cancer **123**(7): 1616-1622.

Liao, Y., X. Du, et al. (2010). "miR-214 regulates lactoferrin expression and pro-apoptotic function in mammary epithelial cells." J Nutr **140**(9): 1552-1556.

Liberzon, A., A. Subramanian, et al. (2011). "Molecular signatures database (MSigDB) 3.0." Bioinformatics **27**(12): 1739-1740.

Lim, L. P., M. E. Glasner, et al. (2003). "Vertebrate microRNA genes." Science **299**(5612): 1540.

Lim, L. P., N. C. Lau, et al. (2005). "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs." Nature **433**(7027): 769-773.

Lim, L. P., N. C. Lau, et al. (2003). "The microRNAs of Caenorhabditis elegans." Genes Dev **17**(8): 991-1008.

Lin, C. Y., V. B. Vega, et al. (2007). "Whole-genome cartography of estrogen receptor alpha binding sites." PLoS Genet **3**(6): e87.

Liu, C. G., G. A. Calin, et al. (2004). "An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues." Proc Natl Acad Sci U S A **101**(26): 9740-9744.

Liu, J., X. J. Luo, et al. (2010). "MicroRNA-214 promotes myogenic differentiation by facilitating exit from mitosis via down-regulation of proto-oncogene N-ras." J Biol Chem **285**(34): 26599-26607.

Liu, L., L. Chen, et al. (2010). "microRNA-195 promotes apoptosis and suppresses tumorigenicity of human colorectal cancer cells." Biochem Biophys Res Commun **400**(2): 236-240.

Long, D., R. Lee, et al. (2007). "Potent effect of target structure on microRNA function." Nat Struct Mol Biol **14**(4): 287-294.

Lopez, F., J. Textoris, et al. (2008). "TranscriptomeBrowser: a powerful and flexible toolbox to explore productively the transcriptional landscape of the Gene Expression Omnibus database." PLoS One **3**(12): e4001.

Lovmar, L., A. Ahlford, et al. (2005). "Silhouette scores for assessment of SNP genotype clusters." BMC Genomics **6**(1): 35.

Lowery, A. J., N. Miller, et al. (2009). "MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor

status in breast cancer." Breast Cancer Res **11**(3): R27.

Lu, J., G. Getz, et al. (2005). "MicroRNA expression profiles classify human cancers." Nature **435**(7043): 834-838.

Lu, M., Q. Zhang, et al. (2008). "An analysis of human microRNA and disease associations." PLoS One **3**(10): e3420.

Lund, E., S. Guttinger, et al. (2004). "Nuclear export of microRNA precursors." Science **303**(5654): 95-98.

Luo, L., L. Ye, et al. (2010). "Microarray-based approach identifies differentially expressed microRNAs in porcine sexually immature and mature testes." PLoS One **5**(8): e11744.

Lynam-Lennon, N., S. G. Maher, et al. (2009). "The roles of microRNA in cancer and apoptosis." Biol Rev Camb Philos Soc **84**(1): 55-71.

MacInnes, A. W., A. Amsterdam, et al. (2008). "Loss of p53 synthesis in zebrafish tumors with ribosomal protein gene mutations." Proc Natl Acad Sci U S A **105**(30): 10408-10413.

Maglott, D., J. Ostell, et al. (2011). "Entrez Gene: gene-centered information at NCBI." Nucleic Acids Res **39**(Database issue): D52-57.

Maragkakis, M., P. Alexiou, et al. (2009). "Accurate microRNA target prediction correlates with protein repression levels." BMC Bioinformatics **10**: 295.

Maragkakis, M., M. Reczko, et al. (2009). "DIANA-microT web server: elucidating microRNA functions through target prediction." Nucleic Acids Res **37**(Web Server issue): W273-276.

Marques, I. J., J. T. Leito, et al. (2008). "Transcriptome analysis of the response to chronic constant hypoxia in zebrafish hearts." J Comp Physiol B **178**(1): 77-92.

Maselli, V., D. Di Bernardo, et al. (2008). "CoGemiR: a comparative genomics microRNA database." BMC Genomics **9**: 457.

Mathew, L. K., S. Sengupta, et al. (2009). "Comparative expression profiling reveals an essential role for raldh2 in epimorphic regeneration." J Biol Chem **284**(48): 33642-33653.

Mathew, L. K., S. S. Sengupta, et al. (2008). "Crosstalk between AHR and Wnt signaling through R-Spondin1 impairs tissue regeneration in zebrafish." FASEB J **22**(8): 3087-3096.

Mathews, D. H., J. Sabina, et al. (1999). "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure." J Mol Biol **288**(5): 911-940.

Maves, L., A. J. Waskiewicz, et al. (2007). "Pbx homeodomain proteins direct Myod activity to promote fast-muscle differentiation." Development **134**(18): 3371-3382.

Maziere, P. and A. J. Enright (2007). "Prediction of microRNA targets." Drug Discov Today **12**(11-12): 452-458.

Megraw, M., P. Sethupathy, et al. (2007). "miRGen: a database for the study of animal microRNA genomic organization and function." Nucleic Acids Res **35**(Database issue): D149-155.

Meiri, E., A. Levy, et al. (2010). "Discovery of microRNAs and other small RNAs in solid tumors." Nucleic Acids Res **38**(18): 6234-6246.

Meister, G. and T. Tuschl (2004). "Mechanisms of gene silencing by double-stranded RNA." Nature **431**(7006): 343-349.

Min, H. and S. Yoon (2010). "Got target? Computational methods for microRNA target prediction and their extension." Exp Mol Med **42**(4): 233-244.

Montojo, J., K. Zuberi, et al. (2010). "GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop." Bioinformatics **26**(22): 2927-2928.

Morin, R. D., M. D. O'Connor, et al. (2008). "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells." Genome Res **18**(4): 610-621.

Murchison, E. P. and G. J. Hannon (2004). "miRNAs on the move: miRNA biogenesis and the RNAi machinery." Curr Opin Cell Biol **16**(3): 223-229.

Musgrove, E. A., C. M. Sergio, et al. (2008). "Identification of downstream targets of estrogen and c-myc in breast cancer cells." Adv Exp Med Biol **617**: 445-451.

Nakashima, T., M. Jinnin, et al. (2010). "Down-regulation of mir-424 contributes to the abnormal angiogenesis via MEK1 and cyclin E1 in senile hemangioma: its implications to therapy." PLoS One **5**(12): e14334.

Nam, S., B. Kim, et al. (2008). "miRGator: an integrated system for functional annotation of microRNAs." Nucleic Acids Res **36**(Database issue): D159-164.

Narducci, M. G., D. Arcelli, et al. (2011). "MicroRNA profiling reveals that miR-21, miR486 and miR-214 are upregulated and involved in cell survival in Sezary syndrome." Cell Death Dis **2**: e151.

Navon, R., H. Wang, et al. (2009). "Novel rank-based statistical methods reveal microRNAs with differential expression in multiple cancer types." PLoS One **4**(11): e8003.

Panico, R., W. H. Powell, et al., Eds. (1993). A Guide to IUPAC Nomenclature of Organic Compounds, Blackwell Scientific publications.

Pasquinelli, A. E., B. J. Reinhart, et al. (2000). "Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA." Nature **408**(6808): 86-89.

Pearson, K. (1901). "On lines and planes of closest fit to systems of points in space." Phliosophical Magazine **2**(11): 559-572.

Pelz, C. R., M. Kulesz-Martin, et al. (2008). "Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data." BMC Bioinformatics **9**: 520.

Penna, E., F. Orso, et al. (2011). "microRNA-214 contributes to melanoma tumour progression through suppression of TFAP2C." EMBO J **30**(10): 1990-2007.

Peter, M. E. (2010). "Targeting of mRNAs by multiple miRNAs: the next step." Oncogene **29**(15): 2161-2164.

Ploner, A. "Heatplus: A heat map displaying covariates and coloring clusters.".

Qiang, R., F. Wang, et al. (2011). "Plexin-B1 is a target of miR-214 in cervical cancer and promotes the growth and invasion of HeLa cells." Int J Biochem Cell Biol **43**(4): 632-641.

R (2010). "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria."

Ramasamy, A., A. Mondry, et al. (2008). "Key issues in conducting a meta-analysis of gene expression microarray datasets." PLoS Med **5**(9): e184.

Raychaudhuri, S., J. M. Stuart, et al. (2000). "Principal components analysis to summarize microarray experiments: application to sporulation time series." Pac Symp Biocomput: 455-466.

Reinhart, B. J., F. J. Slack, et al. (2000). "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans." Nature **403**(6772): 901-906.

Richardson, A. L., Z. C. Wang, et al. (2006). "X chromosomal abnormalities in basal-like human breast cancer." Cancer Cell **9**(2): 121-132.

Riley, M. (1993). "Functions of the gene products of Escherichia coli." Microbiol Rev **57**(4): 862-952.

Robert, P. and Y. Escoufier (1976). "A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient." Applied Statistics **25**(3): 9.

Robison, B. D., R. E. Drew, et al. (2008). "Sexual dimorphism in hepatic gene expression and the response to dietary carbohydrate manipulation in the zebrafish (Danio rerio)." Comp Biochem Physiol Part D Genomics Proteomics **3**(2): 141-154.

Rosa, A., M. Ballarino, et al. (2007). "The interplay between the master transcription factor PU.1 and miR-424 regulates human monocyte/macrophage differentiation." Proc Natl Acad Sci U S A **104**(50): 19849-19854.

Rossum, G. v. (May 1995). Python tutorial, Technical Report CS-R9526. Amsterdam, Centrum voor Wiskunde en Informatica (CWI).

Sabates-Bellver, J., L. G. Van der Flier, et al. (2007). "Transcriptome profile of human colorectal adenomas." Mol Cancer Res **5**(12): 1263-1275.

Saini, H. K., S. Griffiths-Jones, et al. (2007). "Genomic analysis of human microRNA transcripts." Proc Natl Acad Sci U S A **104**(45): 17719-17724.

Sean, D. and P. S. Meltzer (2007). "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." Bioinformatics **23**(14): 1846-1847.

Sehic, A., S. Risnes, et al. (2011). "Effects of in vivo

transfection with anti-miR-214 on gene expression in murine molar tooth germ." Physiol Genomics **43**(9): 488-498.

Sekiya, Y., T. Ogawa, et al. (2010). "Down-regulation of cyclin E1 expression by microRNA-195 accounts for interferon-beta-induced inhibition of hepatic stellate cell proliferation." J Cell Physiol.

Sempere, L. F., S. Freemantle, 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 Biol **5**(3): R13.

Sethupathy, P., B. Corda, et al. (2006). "TarBase: A comprehensive database of experimentally supported animal microRNA targets." RNA **12**(2): 192-197.

Severin, J., K. Beal, et al. (2010). "eHive: an artificial intelligence workflow system for genomic analysis." BMC Bioinformatics **11**: 240.

Shen, W. F., Y. L. Hu, et al. (2008). "MicroRNA-126 regulates HOXA9 by binding to the homeobox." Mol Cell Biol **28**(14): 4609-4619.

Sherva, R., K. Wilhelmsen, et al. (2008). "Association of a single nucleotide polymorphism in neuronal acetylcholine receptor subunit alpha 5 (CHRNA5) with smoking status and with 'pleasurable buzz' during early experimentation with smoking." Addiction **103**(9): 1544-1552.

Shi, L., Z. Cheng, et al. (2008). "hsa-mir-181a and hsa-mir-181b function as tumor suppressors in human glioma cells." Brain Res **1236**: 185-193.

Shi, Y. and Y. Jin (2009). "MicroRNA in cell differentiation and development." Sci China C Life Sci **52**(3): 205-211.

Si, M. L., C. Long, et al. (2010). "Estrogen prevents beta-amyloid inhibition of sympathetic alpha7-nAChR-mediated nitrergic neurogenic dilation in porcine basilar arteries." Acta Physiol (Oxf).

Slack, F. J., M. Basson, et al. (2000). "The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor." Mol Cell **5**(4): 659-669.

Soon, P. S., L. J. Tacon, et al. (2009). "miR-195 and miR-483-5p Identified as Predictors of Poor Prognosis in Adrenocortical Cancer." Clin Cancer Res **15**(24): 7684-7692.

Srivastava, G. P., J. Qiu, et al. (2010). "Genome-wide functional annotation by integrating multiple microarray datasets using meta-analysis." Int J Data Min Bioinform **4**(4): 357-376.

Stabenau, A., G. McVicker, et al. (2004). "The Ensembl core software libraries." Genome Res **14**(5): 929-933.

Subramanian, A., H. Kuehn, et al. (2007). "GSEA-P: a desktop application for Gene Set Enrichment Analysis." Bioinformatics **23**(23): 3251-3253.

Sun, Y., S. Koo, et al. (2004). "Development of a micro-array to detect human and mouse microRNAs and characterization of expression in

human organs." Nucleic Acids Res **32**(22): e188.

Taccioli, C., E. Fabbri, et al. (2009). "UCbase & miRfunc: a database of ultraconserved sequences and microRNA function." Nucleic Acids Res **37**(Database issue): D41-48.

Tamasi, V., K. Monostory, et al. (2011). "Role of xenobiotic metabolism in cancer: involvement of transcriptional and miRNA regulation of P450s." Cell Mol Life Sci **68**(7): 1131-1146.

Tay, Y., J. Zhang, et al. (2008). "MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation." Nature **455**(7216): 1124-1128.

Thomson, J. M., J. Parker, et al. (2004). "A custom microarray platform for analysis of microRNA gene expression." Nat Methods **1**(1): 47-53.

Thorisson, G. A., A. V. Smith, et al. (2005). "The International HapMap Project Web site." Genome Res **15**(11): 1592-1593.

Trupti Joshi, Y. C., Chao Zhang, Guan Ning Lin, Zhao Song, Dong Xu Gene Function Annotation System. Columbia, Digital Biology Laboratory, University of Missouri

Tsang, J. S., M. S. Ebert, et al. (2010). "Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures." Mol Cell **38**(1): 140-153.

Turashvili, G., J. Bouchal, et al. (2007). "Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis." BMC Cancer **7**: 55.

Tweedie, S., M. Ashburner, et al. (2009). "FlyBase: enhancing Drosophila Gene Ontology annotations." Nucleic Acids Res **37**(Database issue): D555-559.

Ujifuku, K., N. Mitsutake, et al. (2010). "miR-195, miR-455-3p and miR-10a(*) are implicated in acquired temozolomide resistance in glioblastoma multiforme cells." Cancer Lett **296**(2): 241-248.

Varambally, S., J. Yu, et al. (2005). "Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression." Cancer Cell **8**(5): 393-406.

Vilella, A. J., J. Severin, et al. (2009). "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates." Genome Res **19**(2): 327-335.

Volkman, H. E., T. C. Pozos, et al. (2010). "Tuberculous granuloma induction via interaction of a bacterial secreted protein with host epithelium." Science **327**(5964): 466-469.

Wang, G., B. C. Kwan, et al. (2010). "Intrarenal expression of miRNAs in patients with hypertensive nephrosclerosis." Am J Hypertens **23**(1): 78-84.

Wang, L. L., J. L. Zhao, et al. (2011). "Estradiol pretreatment attenuated nicotine-induced endothelial cell apoptosis via estradiol functional membrane receptor." Int Immunopharmacol **11**(6): 675-682.

Wang, X. (2008). "miRDB: a microRNA target prediction and

functional annotation database with a wiki interface." RNA **14**(6): 1012-1017.

Wang, X., J. Wang, et al. (2011). "Downregulation of miR-195 correlates with lymph node metastasis and poor prognosis in colorectal cancer." Med Oncol.

Wang, Y., Y. Yu, et al. (2011). "Transforming growth factor-beta regulates the sphere-initiating stem cell-like feature in breast cancer through miRNA-181 and ATM." Oncogene **30**(12): 1470-1480.

Watanabe, T., T. Kobunai, et al. (2007). "Gene expression signature and the prediction of ulcerative colitis-associated colorectal cancer by DNA microarray." Clin Cancer Res **13**(2 Pt 1): 415-420.

Watanabe, T., T. Kobunai, et al. (2006). "Distal colorectal cancers with microsatellite instability (MSI) display distinct gene expression profiles that are different from proximal MSI cancers." Cancer Res **66**(20): 9804-9808.

Watanabe, Y., M. Tomita, et al. (2007). "Computational methods for microRNA target prediction." Methods Enzymol **427**: 65-86.

Wheeler, B. M., A. M. Heimberg, et al. (2009). "The deep evolution of metazoan microRNAs." Evol Dev **11**(1): 50-68.

Wienholds, E., W. P. Kloosterman, et al. (2005). "MicroRNA expression in zebrafish embryonic development." Science **309**(5732): 310-311.

Wu, L., J. Fan, et al. (2006). "MicroRNAs direct rapid deadenylation of mRNA." Proc Natl Acad Sci U S A **103**(11): 4034-4039.

Wu, S., S. Huang, et al. (2010). "Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3' untranslated region." Oncogene **29**(15): 2302-2308.

Wuchty, S., W. Fontana, et al. (1999). "Complete suboptimal folding of RNA and the stability of secondary structures." Biopolymers **49**(2): 145-165.

Xie, X., J. Lu, et al. (2005). "Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals." Nature **434**(7031): 338-345.

Xu, P., S. Y. Vernooy, et al. (2003). "The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism." Curr Biol **13**(9): 790-795.

Xu, T., Y. Zhu, et al. (2009). "MicroRNA-195 suppresses tumorigenicity and regulates G1/S transition of human hepatocellular carcinoma cells." Hepatology **50**(1): 113-121.

Yang, H., W. Kong, et al. (2008). "MicroRNA expression profiling in human ovarian cancer: miR-214 induces cell survival and cisplatin resistance by targeting PTEN." Cancer Res **68**(2): 425-433.

Yang, J. H., J. H. Li, et al. (2011). "starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data." Nucleic Acids Res **39**(Database issue): D202-209.

Yang, Z., S. Chen, et al. (2009). "MicroRNA-214 is aberrantly

expressed in cervical cancers and inhibits the growth of HeLa cells." IUBMB Life **61**(11): 1075-1082.

Yararbas, G. and S. Pogun (2011). "Tamoxifen and mifepriston modulate nicotine induced conditioned place preference in female rats." Brain Res Bull **84**(6): 425-429.

Yi, R., Y. Qin, et al. (2003). "Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs." Genes Dev **17**(24): 3011-3016.

Yin, K. J., Z. Deng, et al. (2010). "miR-497 regulates neuronal death in mouse brain after transient focal cerebral ischemia." Neurobiol Dis **38**(1): 17-26.

Yu, W., M. Clyne, et al. (2010). "Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations." Bioinformatics **26**(1): 145-146.

Yu, W., M. Gwinn, et al. (2008). "A navigator for human genome epidemiology." Nat Genet **40**(2): 124-125.

Yu, W., R. Ned, et al. (2009). "The need for genetic variant naming standards in published abstracts of human genetic association studies." BMC Res Notes **2**: 56.

Yu, W., A. Wulf, et al. (2008). "Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases." BMC Bioinformatics **9**: 528.

Yu, W., A. Wulf, et al. (2008). "HuGE Watch: tracking trends and patterns of published studies of genetic association and human genome epidemiology in near-real time." Eur J Hum Genet **16**(9): 1155-1158.

Yu, W., A. Yesupriya, et al. (2007). "An open source infrastructure for managing knowledge and finding potential collaborators in a domain-specific subset of PubMed, with an example from human genome epidemiology." BMC Bioinformatics **8**: 436.

Zhang, X. J., H. Ye, et al. (2010). "Dysregulation of miR-15a and miR-214 in human pancreatic cancer." J Hematol Oncol **3**: 46.

Zhang, Y., L. X. Yan, et al. (2011). "miR-125b is Methylated and Functions as A Tumor Suppressor by Regulating the ETS1 proto-oncogene in Human Invasive Breast Cancer." Cancer Res.

Zhu, H., Y. Yang, et al. (2011). "MicroRNA-195 promotes palmitate-induced apoptosis in cardiomyocytes by down-regulating Sirt1." Cardiovasc Res.

Zhu, W., D. Zhu, et al. (2011). "miR-497 modulates multidrug resistance of human cancer cell lines by targeting BCL2." Med Oncol.

CHAPTER 7:APPENDIX

7.1 TUTORIALS ON HOW TO USE MESADB

7.1.1 Protocols:

7.1.1.1 Motif & Expression

1) Select a dataset from the list
2) Select a group of microRNAs either by manual entry, or through uploading a list of microRNAs or entering a sequence motif in IUPAC alphabet to search for.
3) Select one of the three analysis options: expression analysis, correspondence analysis, or co-inertia analysis
    a. If expression analysis is selected, the resulting output page will contain a bar graphic in which the selected microRNA expression level is indicated with respect to that of the unselected microRNAs in the different tissues. Each bar is color coded by using a phi-coefficient indicating the enrichment (red) or depletion (green) of the selected microRNAs in the given tissue. The expression data can be retrieved as a table.
    b. If correspondence analysis is selected, a page with three tabs will come up. The tabs are 'Tissues', 'miRNAs' and 'Correspondence'. The figure represented in 'tissues' tab shows projections of tissues onto a new reduced common covariance space of tissues and microRNAs. The tissues are represented by one to three letters. To see the full names of the tissues, putting the mouse arrow on the letters is necessary. If the 'miRNAs' tab is selected, the figure show projections of microRNAs on to the same space. Finally 'Correspondence' tab is for visualizing the co-localization of the previous two figures. Closer the microRNAs or tissues in space to each other the closer they are in expression patterns.

c. If co-inertia button is clicked the output will be a two-tabbed result page. Two figures that appear in 'Tissues' tab shows the correspondence between the tissues and the seed motifs obtained from MEME, a motif finding algorithm.

7.1.1.2 Expression &Expression

1) Select a dataset from the left box and send it to the box on the right with the arrow button between two boxes. Do it for a second dataset. Then click the submit button.
2) Select a group of microRNAs either by manual entry, or through uploading a list of microRNAs or entering a sequence motif in IUPAC alphabet to search for. Click on 'Co-inertia Analysis'. Two tabs are available: Tissues and miRNAs. In tissues, how two datasets compare to each other with respect to expression in tissues will be shown using MADE 4.0 program. The other tab, 'miRNAs' shows how projected microRNAs diverged between two datasets on the common space. The tails of the arrows represent projections from the first dataset and the heads of the arrows represent the projections from second dataset. The miRNA tab has three sub tabs. They are 'Labeled', 'Unlabeled' and 'Clustered'. 'Labeled' and 'Unlabeled' sub tabs are for showing or hiding the labels holding the names of microRNAs represented by the arrows. 'Clustered' tab is the output of K-means clustering of projections with minimum silhouette, for k in 2 to 10. The cluster number that shows the lowest silhouette will be the default number of clusters chosen automatically. A select box exists on the left top site of the page shown in this sub tab for selecting the different clustering for each K. The resulting page will have an extra tab called 'Heatmap' that shows the hierarchical clustering of the two datasets in terms of genes and classes i.e., tissues, accompanied with the heatmap representing relative expression values that are mean subtracted. Red color means that the value above 0 whereas green color means that the value is below 0.

7.1.1.3 Motif & Function

1) Select an organism. There are three possibilities: *Homo sapiens*, *Mus musculus* and *Danio rerio*.
2) Select the method of microRNAs from selection box 'select miRNA upon' by specifiying dinucleotide motif, or through uploading a list of microRNAs or entering a sequence motif in IUPAC alphabet to search for. Except for the file upload option, determine a region where the motif should be present. Then click Submit.
3) Resulting page will show you 'GO terms', 'HUGE terms' and 'KEGG pathways' buttons. 'HUGE terms' button is only available for *Homo sapiens*. In the next line a selection bar for GO term type is present. If GO term analysis is considered, selection of the desired ontology type is required. Last line of this page is a tab called 'Click to see miRNAs'. If this tab is clicked, the selected list of microRNAs will appear, with a download option for getting their mature sequences in FASTA format. Each microRNAs in the list is clickable to direct the client to a page where the targets of it could be seen. This option is extended in following relevant section.
4) The resulting page will show four columns. First column is a Term ID column depends on the source selection for the term, i.e., HuGE ID. These IDs are clickable and directs the user for the original source page. The second column is 'Term Name' column. The list and the header of the column are specific to the term source. Third and forth columns are representing the genes targeted by the selected microRNAs and P-values that show the significance of the terms determined by the target genes respectively. All in all, the result page shows that whether the selected microRNAs are associated with significant GO terms, HUGE terms or KEGG pathways. The results are downloadable as tab delimited text file format.

7.1.1.4 MicroRNA search

1) Enter a mature microRNA ID.
2) Select an organism.
3) In this section one can analyze whether the selected microRNA is associated with a functional term (i.e., GO, HUGE or KEGG) or which targets it has or what kind of expression it has in a selected expression dataset. The figures and results pages are as described before in the motif-expression, expression-expression, and motif-function modules.

7.2 ARC TABLES

Table 7.2.1: Datasets used in ARC.

Name of the Study	Reference	Organism/Tissue	# of Samples	Normalization
Integrative Genomic and Proteomic Analysis of Prostate Cancer Reveals Signatures of Metastatic Progression (GSE3325)	(Varambally, Yu et al. 2005)	Human/Prostate	19	RMA
Molecular Marker for predicting development of cancer in ulcerative colitis (GSE3629)	(Watanabe, Kobunai et al. 2007)	Human/Colon	121	MAS5
Human breast tumor expression (GSE3744)	(Richardson, Wang et al. 2006; Alimonti, Carracedo et al. 2010)	Human/Breast	47	RMA
Expression profiling in early onset colorectal cancer (GSE4107)	(Hong, Ho et al. 2007)	Human/Colon	22	RMA

Inflammation, adenoma and cancer: objective classification of colon biopsy specimens with gene expression signature (GSE4183)	(Gyorffy, Molnar et al. 2009; Galamb, Spisak et al. 2010)	Human/Colon	53	RMA
Gene expression signature of colorectal cancer with microsatellite instability (GSE4554)	(Watanabe, Kobunai et al. 2006)	Human/Colon	84	MAS5
Analysis of microdissected invasive lobular and ductal breast carcinomas in relation to normal ductal and lobular cells (GSE5764)	(Turashvili, Bouchal et al. 2007)	Human/Breast	30	RMA
Expression data from human breast tissue (GSE7904)	(Richardson, Wang et al. 2006)	Human/Breast	62	RMA
Transcriptome profile of human colorectal adenomas. (GSE8671)	(Sabates-Bellver, Van der Flier et al. 2007)	Human/Colon	64	RMA
embryo (GSE4201)	(Giraldez, Mishima et al. 2006)	Zebrafish/Embryo	15	RMA
heart (GSE4989)	(Marques, Leito et al. 2008)	Zebrafish/Hearth	10	RMA

endothelial (GSE12039)	(Fish, Santoro et al. 2008)	Zebrfish/Embryo	12	RMA
nerve (GSE11493)	(MacInnes, Amsterdam et al. 2008)	Zebrafish/Nerve	11	RMA
brain (GSE11107)	(Drew, Rodnick et al. 2008)	Zebrafish/Brain	8	RMA
liver (GSE11107)	(Drew, Rodnick et al. 2008)	Zebrafish/Brain	10	RMA
fin (GSE10188)	(Mathew, Sengupta et al. 2009)	Zebrafish/Fin	24	RMA
fin (GSE10184)	(Mathew, Sengupta et al. 2008)	Zebrafish/Fin	12	RMA
heart (GSE9020)	(Chen, Carney et al. 2008)	Zebrafish/Hearth	48	RMA
embryo (GSE9020)	(Chen, Carney et al. 2008)	Zebrafish/Hearth	48	RMA
retina (GSE8874)	(Chen, Carney et al. 2008)	Zebrafish/Retina	24	RMA
embryo (GSE8874)	(Chen, Carney et al. 2008)	Zebrafish/Retina	24	RMA

liver (GSE8856)	(Robison, Drew et al. 2008)	Zebrafish/Liver	17	RMA
embryo (GSE8428)	(Maves, Waskiewicz et al. 2007)	Zebrafish/Embryo	12	RMA
larvae (GSE8327)	(Volkman, Pozos et al. 2010)	Zebrafish/Embryo	9	RMA
embryo (GSE7658)		Zebrafish/Embryo	8	RMA
embryo (GSE6127)		Zebrafish/Embryo	6	RMA
retina (GSE5048)	(Leung, Ma et al. 2007)	Zebrafish/Ovary	8	RMA
ovary (GSE4859)	(Heiden, Struble et al. 2008)			RMA
Expression data from human normal pre-frontal cortex, liver, and colon tissues and colon tumors (GSE13471)	(Irizarry, Ladd-Acosta et al. 2009)	Human/Colon	13	RMA
Expression data from primary colorectal cancers (GSE13067)	(Jorissen, Lipton et al. 2008)	Human/Colon	74	RMA

Expression data from primary colorectal cancers (GSE13294)	(Jorissen, Lipton et al. 2008)	Human/Colon	155	RMA
Expression data from human colonic biopsy sample (GSE10714)	(Galamb, Sipos et al. 2008; Galamb, Spisak et al. 2010)	Human/Colon	33	RMA
Expression data of hormone-responsive MCF-7 cells versus estrogen-deprived MCF-7:5C and MCF-7:2A breast cancer cells (GSE10879)		Human/Breast	11	RMA
Timecourse of estradiol (10nM) exposure in MCF7 breast cancer cells. (GSE11352)	(Lin, Vega et al. 2007)	Human/Breast	18	RMA
Crosstalk between Estrogen and TNFalpha in MCF-7 Breast Cancer Cells (GSE11467)	(Frasor, Weaver et al. 2009)	Human/Breast	12	RMA
Estrogen- and Myc-regulated genes in MCF-7 breast cancer cells (GSE11791)	(Musgrove, Sergio et al. 2008)	Human/Breast	15	RMA
Transcription profiling of brain from zebrafish treated with ethanol or nicotine suggests conservation of neuro-adaptation	(Kily, Cowe et al. 2008)	Zebrafish/Brain	7	RMA

pathways (EMEXP1301)				
Transcription profiling of zebrafish ZF4 and PAC2 cell lines cultured in medium with or without FCS (EMEXP736)	(He, Salas-Vidal et al. 2006)	Zebrafish/Fibroblast	9	RMA

D170–D180 *Nucleic Acids Research, 2011, Vol. 39, Database issue*
doi:10.1093/nar/gkq1256

mESAdb: microRNA Expression and Sequence Analysis Database

Koray D. Kaya[1], Gökhan Karakülah[2], Cengiz M. Yakıcıer[1,3], Aybar C. Acar[4,*] and Özlen Konu[1,5,*]

[1]Department of Molecular Biology and Genetics, Bilkent University, 06800 Ankara, Turkey, [2]Department of Medical Informatics, Health Sciences Institute, Dokuz Eylül University, 35340 Inciralti, Izmir, [3]Department of Medical Biology, Acibadem University, 34848 Maltepe, Istanbul, [4]Department of Computer Engineering and [5]BilGen, Bilkent University Genetics and Biotechnology Research Center, Bilkent University, 06800 Ankara, Turkey

Received October 1, 2010; Revised November 19, 2010; Accepted November 20, 2010

ABSTRACT

microRNA expression and sequence analysis database (http://konulab.fen.bilkent.edu.tr/mirna/) (mESAdb) is a regularly updated database for the multivariate analysis of sequences and expression of microRNAs from multiple taxa. mESAdb is modular and has a user interface implemented in PHP and JavaScript and coupled with statistical analysis and visualization packages written for the R language. The database primarily comprises mature microRNA sequences and their target data, along with selected human, mouse and zebrafish expression data sets. mESAdb analysis modules allow (i) mining of microRNA expression data sets for subsets of microRNAs selected manually or by motif; (ii) pair-wise multivariate analysis of expression data sets within and between taxa; and (iii) association of microRNA subsets with annotation databases, HUGE Navigator, KEGG and GO. The use of existing and customized R packages facilitates future addition of data sets and analysis tools. Furthermore, the ability to upload and analyze user-specified data sets makes mESAdb an interactive and expandable analysis tool for microRNA sequence and expression data.

INTRODUCTION

microRNAs are small (19–22 nt) RNAs that play crucial roles in many cellular processes via targeting mRNAs for translational repression or cleavage thus regulating gene expression (1). microRNAs, through their compatible 5′-seed sequences, exert regulatory functions primarily on the 3′-untranslated regions (UTRs) of targeted mRNAs (2–4). microRNAs that target the same mRNAs may share common motifs due to duplication events and/or common evolutionary ancestry (5–7). Previous studies using genome search and target prediction algorithms have provided lists of common genomic regulatory nucleotide motifs, some of which are also shared by microRNA sequences (8). However, whether microRNAs that are similar in sequence exhibit similarities also in function and/or expression is not yet well understood and warrants further study.

Based on large-scale studies, microRNAs have been annotated for their specificity for particular tissues, developmental stages and/or pathologies such as cancer (9–12). For example, Bargaje *et al.* (13) compiled and normalized multiple data sets from different sources to determine the tissue-specific and tissue-invariant consensus expression profiles. Others have surveyed microRNA expression profiles in large numbers of normal and cancerous tissues to decipher microRNA networks and conserved expression clusters in disease (14). There also is evidence suggesting that expression patterns of microRNAs are conserved at the species level (5). Nevertheless, the conserved associations between microRNA sequence motifs and expression profiles across taxa still remain relatively unexplored (15–17). Similarly, there is a growing need for tools developed for multivariate comparison of expression patterns between different data sets (18).

In recent years, several databases and analysis tools also have been published that feature high-throughput analysis results of microRNA sequence or expression. Among these, miRBase functions as a central repository for microRNA genomics for a variety of organisms and thus serves the community with up-to-date microRNA sequence, chromosome location and transcript information (19). mSigDB, using motif lists from Xie *et al.* (8), provides microRNA target gene lists that could be tested

*To whom correspondence should be addressed. Tel: +90 312 290 2123; Fax: +90 312 266 5097; Email: konu@fen.bilkent.edu.tr
Correspondence should also be addressed to Aybar C. Acar. Tel: +90 312 290 2094; Fax: +90 312 266 4047; Email: aacar@cs.bilkent.edu.tr

for enrichment with Gene Ontology (GO) functional terms, KEGG signaling pathways or other gene lists (20). Similarly, a manually curated database, called Mir2DiseaseBase, can be used for extracting associations between diseases and microRNAs (21). Most recently, mirBridge has been developed to predict microRNA function and link microRNAs with cellular pathways using network algorithms (22). Among the expression analysis focused databases, miRGator is a comprehensive repository and analysis tool for microRNA expression, target and ontology data providing a graphical transcriptional evaluation of selected microRNA types for mice or humans (23). microRNA.org is another source of microRNA expression and functional data for understanding microRNA expression regulation through target prediction and examination of tissue transcript abundance (24). Accordingly, using a database approach has been fruitful in allowing the users to interactively query and make associations among large-scale data sets and thus is highly suited for exploring the association between sequence motifs and expression profiles of microRNAs and meta-analysis of microRNA expression data sets.

In the present study, we have developed the microRNA Expression and Sequence Analysis Database, mESAdb, to provide a series of interactive analysis tools for testing the association of microRNA sequence characteristics with target gene function, human diseases and microRNA expression patterns using multivariate analyses. mESAdb is also a meta-analysis tool for comparative analysis of function and expression for microRNA lists across different taxa, including human, mouse and zebrafish. Complementing existing databases, mESAdb takes advantage of the available sequence information to search for microRNAs with common motifs (e.g. dinucleotide frequencies or conserved seed sequences) after which these microRNA sets can be analyzed for determination of the extent of coordinate expression and target gene enrichment using terms from GO, KEGG and HUGE Navigator databases (25–27). mESAdb is compatible and periodically updated in an automated way with data from related large external repositories. It also allows upload and analysis of user-specified data sets, and makes extensive use of existing and customized R packages (28). Overall, we believe that mESAdb, by specifically addressing the need for comparative and multivariate analysis of microRNA sequence and expression profiles, is likely to significantly enhance our understanding of the role of microRNAs in biological processes.

DATABASE DESIGN AND STRUCTURE

mESAdb enables access and retrieval of microRNAs with specified motifs to associate and analyze them functionally as well as based on expression profiles (Figure 1). An initial version of this work was presented in abstract form in BioSysBio 2007: Systems Biology, Bioinformatics, Synthetic Biology (29).

Data collection and storage

Data used in mESAdb are obtained periodically from multiple sources and processed for integration into the underlying MySQL database using a series of routines which download, parse and integrate these data from relevant sources (Ensembl, miRBase, microCosm, HUGE, KEGG and GO) either directly or through the Biomart integration service (Figures 1 and 2) (19,25–27,30).

Mature microRNA ID and sequences were downloaded from miRBase Release 15 (31); each microRNA was associated with a species-specific sequence and stored in a table. microRNA microarray experiment data sets for human, mouse and zebrafish, primarily focusing on expression from different tissues and developmental stages, were stored separately as default data sets (14,32–38) (Table 1; Supplementary Data). Tables containing the normalized expression values were associated with sequence data linked with the corresponding mirRBase names for these microRNAs (Figure 2). Where available, the probe sequences printed on microarrays that match exactly with the species-specific reverse complementary sequences in miRBase were included resulting in increased stringency; thus the number of microRNAs from each microarray study incorporated into mESAdb might be smaller than that reported in the original study. Expression data were logarithmically transformed where necessary, and quantile normalized (39). To link sequence and expression properties with functional information, the predicted human targets were retrieved from MicroCosm Targets (Figure 2) (19). These targets were further processed on the R environment (Version 2.11.1) (www.bioconductor.org); transcript IDs were matched with Ensembl Gene IDs (Ensembl Relese 59) using the package *biomaRt* (30). Only a single Ensembl ID was retrieved for each target gene with multiple transcript entries. Species-specific microRNAs were paired with target gene IDs associated with ontology terms and these matched pairs were stored in mESAdb's underlying DBMS (Figure 2; MySQL). KEGG and Gene Ontology terms associated with microRNA targets were extracted and matched with the corresponding microRNA IDs (25,26). The disease terms associated with microRNA targets were obtained from the phenopedia view of HUGE Navigator, an integrated knowledge base of genetic associations and human genome epidemiology (27). These terms were parsed and matched with microRNA targets and stored in the MySQL tables underlying mESAdb. Target and associated terms are updated periodically (Figure 2).

User-specified expression data set management

mESAdb incorporates a tool for the upload of user-specified expression data sets provided as comma separated files (Figure 3). The user is free to add, view and remove expression data sets having expression data for arbitrary numbers of microRNAs against arbitrary number of expression classes (e.g. tissues, developmental stages, disease states). The format for the input file is straightforward: a comma delimited file with the first row giving the names of the classes, the subsequent lines

Figure 1. Screenshot of the mESAdb main page. The modules, 'motif-expression', 'expression-expression', 'motif-function' and 'microRNA search', are shown.

of the file each beginning with the name of the microRNA (e.g. the miRBase ID) and, optionally, the probe sequence, followed by the measured expression for each of the classes given in the header line. The file uploaded is preprocessed line-by-line; for each, if the reverse complement of the probe sequence given for that line contains the mature sequence for the corresponding microRNA as given in the latest miRBase, the line is verified. For lines that cannot be thus verified, the system searches for a match in the miRBase sequences for the relevant species. If found, the microRNA is renamed to its miRBase standard name, if not, the line is discarded. Subsequently, the lines for duplicate microRNAs are averaged. The upload module generates a downlodable text file listing the actions performed while parsing and processing the .csv file. mESAdb uses nomenclature by miRBase for cross taxa comparisons performed in 'expression–expression' module where microRNAs with the same name from two different species are matched. Most microRNAs carrying the same name exhibit high sequence similarity across species while ~5% are relatively divergent. Data Upload utility also warns users for such cases. The module also provides utilities to log transform, center and scale or quantile normalize the expression data upon verification by miRbase (Figure 3).

A user-uploaded data set is tied to the specific user account that creates it and may be retrieved from another source or may be the product of the users' own research. We protect privacy of proprietary data by keeping uploaded sets visible only to the account that owns it and no data is retained once a user removes a data set. An exemplary data set was provided in the current version of the mESAdb (40) (i.e. GSE2564NORMAL_seq.csv; mESAdb supporting material; http://konulab.fen.bilkent.edu.tr/mirna/supplementary_files.php). Accordingly, we downloaded GSE2564 expression series matrix from GEO (41). This data set includes normal tissues from stomach ($n = 5$), colon ($n = 5$), pancreas ($n = 1$), liver ($n = 3$), kidney ($n = 3$), bladder ($n = 2$), prostate ($n = 8$), uterus ($n = 9$), lung ($n = 4$), breast ($n = 3$) and brain ($n = 2$), together with cancer samples for different tissues. For the example used herein, only the expression data on the normal tissues were obtained; linked with microRNA IDs and probe

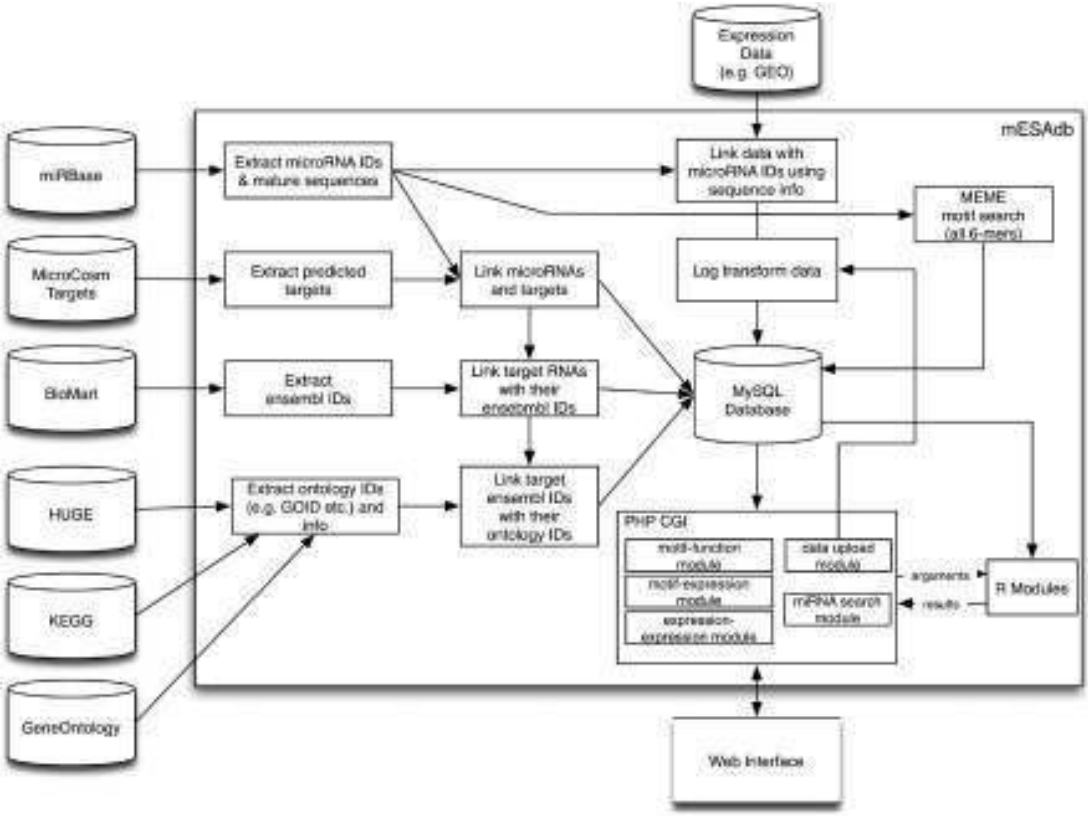


Figure 2. Workflow diagram of mESAdb. MESAdb combines data from a variety of external data sources. For example, microRNA mature sequences and IDs are retrieved from miRBase and matched with microRNA data sets (e.g. from GEO). microRNA sequences are processed by the MEME motif finder for conserved motifs. The microRNA targets are fetched from EBI's MicroCosm Targets for each species; BioMart is used to get the ENSEMBL Gene IDs of the targets' transcript IDs. These ENSEMBL Gene IDs are then linked to HUGE Navigator Disease IDs, KEGG Pathway IDs and GO IDs. A user-friendly interface has been developed in PHP for accessing data in the system and allowing versatile analysis via various R scripts (http://php.net; http://www.r-project.org./; http://www.mysql.com/).

Table 1. Default data sets provided in mESAdb (processed data sets; supporting material at http://konulab.fen.bilkent.edu.tr/mirna/supplementary_files.php)

References	Species	GSE No.	Platform	Pubmed ID	Tissues
Meiri *et al.* (36)	*Homo sapiens*	GSE20414	GPL10067	20483914	Ly, K, En, Lu, Bl, B, H, Li
Navon *et al.* (14)	*Homo sapiens*	GSE14985	GPL8227	19946373	Br, Pr, Ly, O, Co, Li, Te, Lu
Ach *et al.* (32)	*Homo sapiens*	GSE11806	GPL6955	18783629	Pl, B, Br, H, Th, Li, O, SM, Te
Barad *et al.* (34)	*Homo sapiens*	–	MOE-ER array	15574827	HeLa, B, Li, Th, Te, Pl
Baskerville and Bartel (33)	*Homo sapiens*	–	MWG biotech	15701730	BM, B, H, K, Li, Lu, Pa, Pr, SM, Sp, Th, FC, Ly, Co, HelaS3, Ce, Bl, Te, A, U, Br, F, SI, Pl, O
Wienholds *et al.* (38)	*Danio rerio*	GSE2628	GPL2023	15919954	B, Ey, SM, H, Gi, Fi, Sk, G, Li, Te, O
Thomson *et al.* (37)	*Mus musculus*	GSE1635	GPL1391	15782152	Li, K, Lu, O, H, B, Th, ES, EBD3, EBD28, E7, E11, E15, E17
Beuvink *et al.* (35)	*Mus musculus*	–	Custom	17355992	B, Li, Lu, SI, SM, H, K, Sp

F, Fallopian tube; U, Uterus; Ly, Lymph node; Pl, Placenta; Br, Breast; Pa, Pancreas; Li, Liver; B, Brain; Th, Thymus; H, Heart; Lu, Lungs; Sp, Spleen; Te, Testicle; O, Ovary; K, Kidney; SM, Skeletal muscle; SI, Small intestine; Co, Colon; Pr, Prostate; Bl, Bladder; Ce, Cervix; A, Adrenal gland; St, Stomach; BM, Bone Marrow; FC, Frontal Cortex; Ey, Eye; Gi, Gill; Fi, Fin; Sk, Skin; G, Gut; HeLa, HeLa Cells; HL3S, HeLa S3; En, Endometrium. (http://php.net; http://www.r-project.org./; http://www.mysql.com/)

sequences in the GPL1986 description file; and a comma separated file was formed for upload. An account of processing of the microRNAs in the .csv file was generated by mESAdb. The data set, called GSE2564_normal, could be uploaded using the 'Manage Datasets' facility of mESAdb (Figure 3) and compared with the existing data sets listed in Table 1.

Integration of R-packages

mESAdb uses a hybrid of PHP and R as a computational environment. The basic operations and the web interface elements are coded in PHP whereas more significant statistical analyses are performed in R (Figure 2). The web interface has been made as responsive and user-friendly as

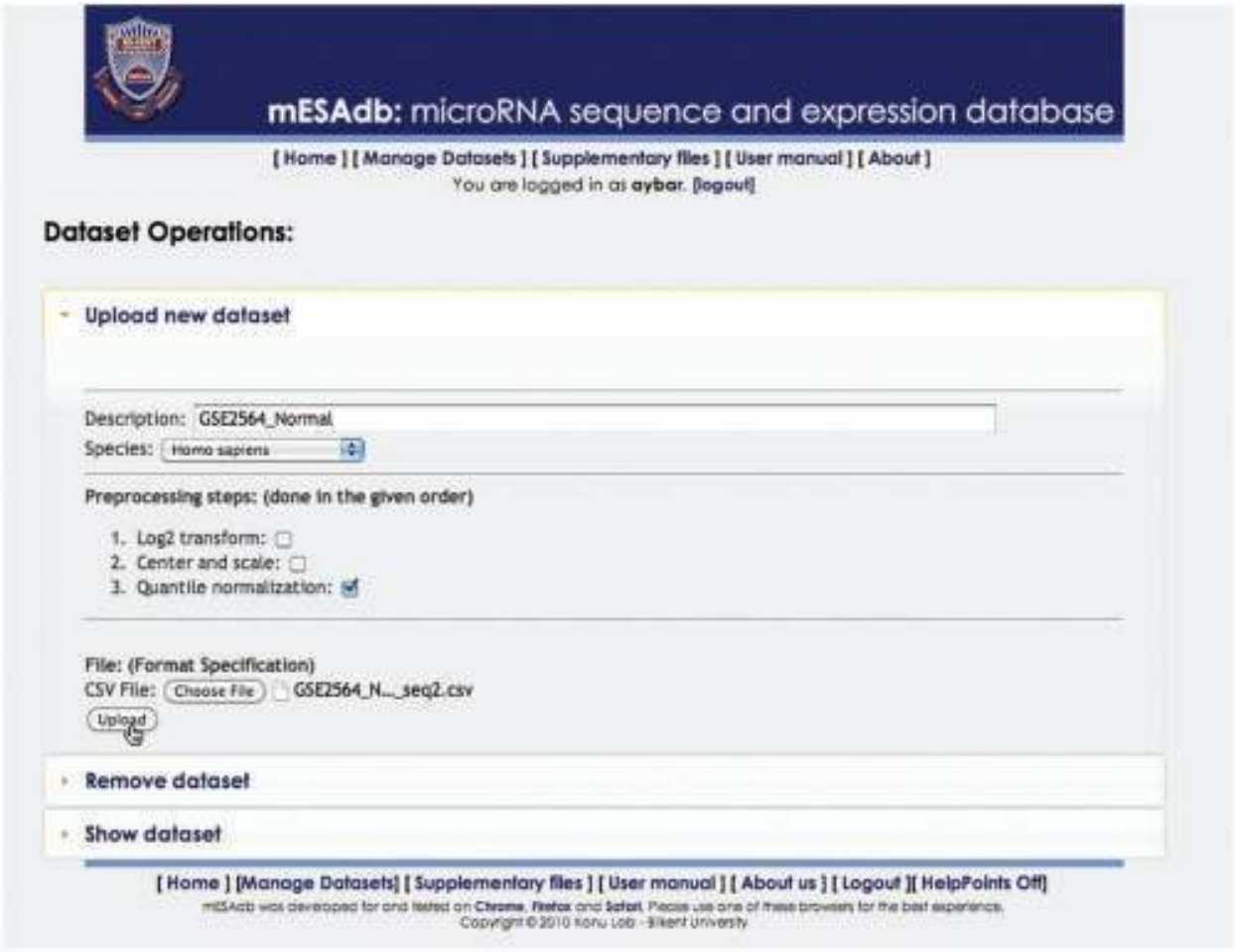


Figure 3. Screenshot of the data upload module. User can select from species and microarray normalization options and then browse to upload a data set.

possible with the addition of dynamic elements created with Javacript and the JQuery UI (http://jqueryui.com/) library. The communication between the PHP and R environments is performed using the common underlying MySQL database and Unix pipes. Briefly, a PHP script creates a child R process to which command line arguments are passed onto. The R process uses this information to retrieve the relevant information from the MySQL database and subsequently prepares the output (e.g. graphics; the bar plot, correspondence plots) which it passes onto the calling PHP script to display on the page. If the output is mostly textual (e.g. tabular data), it is passed on the output stream of the R program. If it is a larger result like an image, the R program saves it under a predetermined filename in a temporary location, which the PHP script retrieves from once the child R process is finished. This two-way communication between the PHP code and its R child processes has been implemented as a simple but effective API, which allows new R scripts to be easily integrated into the mESAdb tool as needed. This enables mESAdb to build on well-designed and verified analysis packages such as MADE4 (28) available for the R environment and use them to leverage its analysis tasks.

DESCRIPTIONS OF ANALYSIS MODULES

Motif-expression

mESAdb has a motif selection tool with a pulldown menu in which users might select from different options to group retreived microRNAs with a given motif, i.e. dinucleotide motifs or motifs up to 6 nt long using the IUPAC code (42). It is also possible to upload user-specified microRNA lists. 'Motif-expression' module integrates the motif selection tool with default microarray data sets found in mESAdb as well as those uploaded by the user (Figures 1 and 3; Table 1). Accordingly, mESAdb provides a platform for visualization of microRNA expression in humans, mouse and zebrafish. Once a microRNA list is selected, expression of this set of microRNAs can be investigated using three different analysis options: 'expression analysis', 'correspondence analysis' and 'co-intertia analysis'.

The 'expression analysis' option enables the user to compare, using bar plots, the amount of mean expression of the selected microRNAs with those of the remaining microRNAs across the studied expression classes, i.e. tissues or developmental stages. Expression data (Table 1) for the selected microRNAs and those for the

unselected microRNAs are extracted from the quantile normalized log transformed expression tables that have been generated by the mESAdb Data Upload facility. The class (e.g. tissue) specific mean values for the selected and unselected microRNAs then are plotted separately for each column of the data set (e.g. each tissue) using a bar plot. Bars are color coded by the value of the ϕ-coefficient to assess the association of the selected microRNAs with the tissue in consideration (also called the Yule-ϕ; Supplementary Data) (43). A dynamic hover feature has been implemented for user to see exact information about each column by hovering the mouse pointer over it in the barplot. Expression data sets are accessible in the html format, and the χ^2 and *P*-values for the ϕ-coefficient also are generated. Help boxes are made available for data plots and analysis tools.

mESAdb performs multivariate analysis of expression using the R package *MADE4* customized for visualization and analysis in mESAdb (28). 'Correspondence analysis' of the selected set of microRNAs produce three graphical outputs, allowing for visualization of the expression patterns across classes (e.g. tissues), or microRNAs, or both the classes and microRNAs. 'Co-inertia analysis' (28) of the selected set of microRNAs helps visualize the similarities between microRNA expression and occurrence of common 6-mer MEME motifs (44) found among the microRNA sequences housed in mESAdb (45). Users can link from a motif to back to the 'expression analysis' module explained above to visualize the expression data as bar plots per expression class (e.g. tissue), of the group of microRNAs used in the coinertia plot containing the specified motif . MEME motif outputs we generated for the human, mouse and zebrafish microRNAs can be accessed from the supporting material (http://konulab.fen.bilkent.edu.tr/mirna/supplementary_files.php) found at the mESAdb.

Expression-expression

This module provides a tool for meta-analysis of microRNA expression data sets. Selected sets of microRNAs can be investigated with regard to the data sets listed in Table 1 in a pair-wise fashion; other user-defined data sets can be uploaded and analyzed as well (Figure 1). 'Expression–expression module' outputs coinertia graphics for (i) classes (e.g. tissues) and (ii) microRNAs, and also a heatmap of both data sets using customized MADE4 (28) and heatplus (http://bioconductor.org/packages/2.6/bioc/vignettes/Heatplus/inst/doc/Heatplus.pdf) packages in R (www.bioconductor.org). The output has been customized for better visualization; and the degree of association, indicated by the RV coefficient (28,46) between two different microarray data sets also is provided. A high RV score suggests better correlation among data sets. For the microRNA oriented coinertia graph, several utilities are provided in order to facilitate the visualization of potentially high numbers of datapoints. It is possible to visualize the microRNA datapoints with or without labels on the coinertia graph. The coinertia tool also provides an automatic clustering of the microRNAs based on the similarity of their expressions in both data sets using *k*-means clustering (47); the default clustering displayed is the clustering with the maximum silhouette coefficient (48). Since *k*-means clustering is not deterministic, for each *k*-value the module performs 20 runs of the algorithm and the best clustering for each *k* is selected using highest

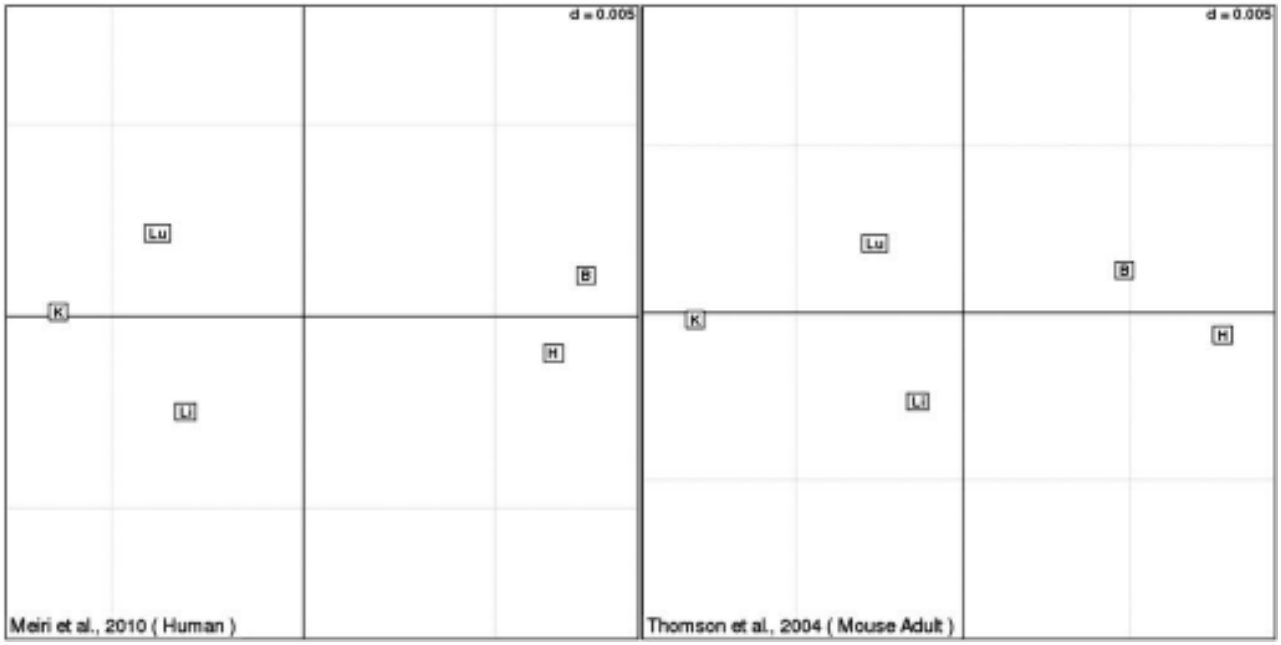


Figure 4. Coinertia plot of Meiri and Thomson expression data sets for a set of microRNA clusters with sequence similarity (mESAdb supporting material; http://konulab.fen.bilkent.edu.tr/mirna/supplementary_files.php). Similarity of microRNA expression patterns between mice and humans are shown for brain (B), heart (H), kidney (K), liver (Li), and lung (Lu).

silhouette. The clustering with the overall best silhouette is displayed by default. The user can manually set a cluster number between 2 and 10 clusters (i.e. $2 \leq k \leq 10$) if desired. These clusters can further be investigated to visualize the expression profiles for the given data sets using expression bar plots of in-cluster and out-of-cluster microRNAs, by clicking on the cluster centroids.

Motif-function

This function may be useful for functional analysis of, for example, a set of differentially expressed microRNAs (Figures 1 and 2). In the present study, information from HUGE Navigator, in addition to GO and KEGG databases can be associated with the selected microRNAs (25–27). For any selected subset of microRNAs, mESAdb then can be used to retrieve the mappings of the selected functional terms, with the targets of these microRNAs and subsequently to calculate a probability value based on the hypergeometric distribution (49).

microRNA search

Functional and expression correlates of a single microRNA can be assessed using this module to enable a quick search involving multiple modules of mESAdb (Figure 1). Terms from GO, HUGE, KEGG and target genes associated with the given microRNA can be extracted; and the observed and expected counts as well as hypergeometric *P*-values can be downloaded. Expression profile of the selected microRNA also can be visualized using the aforementioned bar plots and downloaded as .txt files.

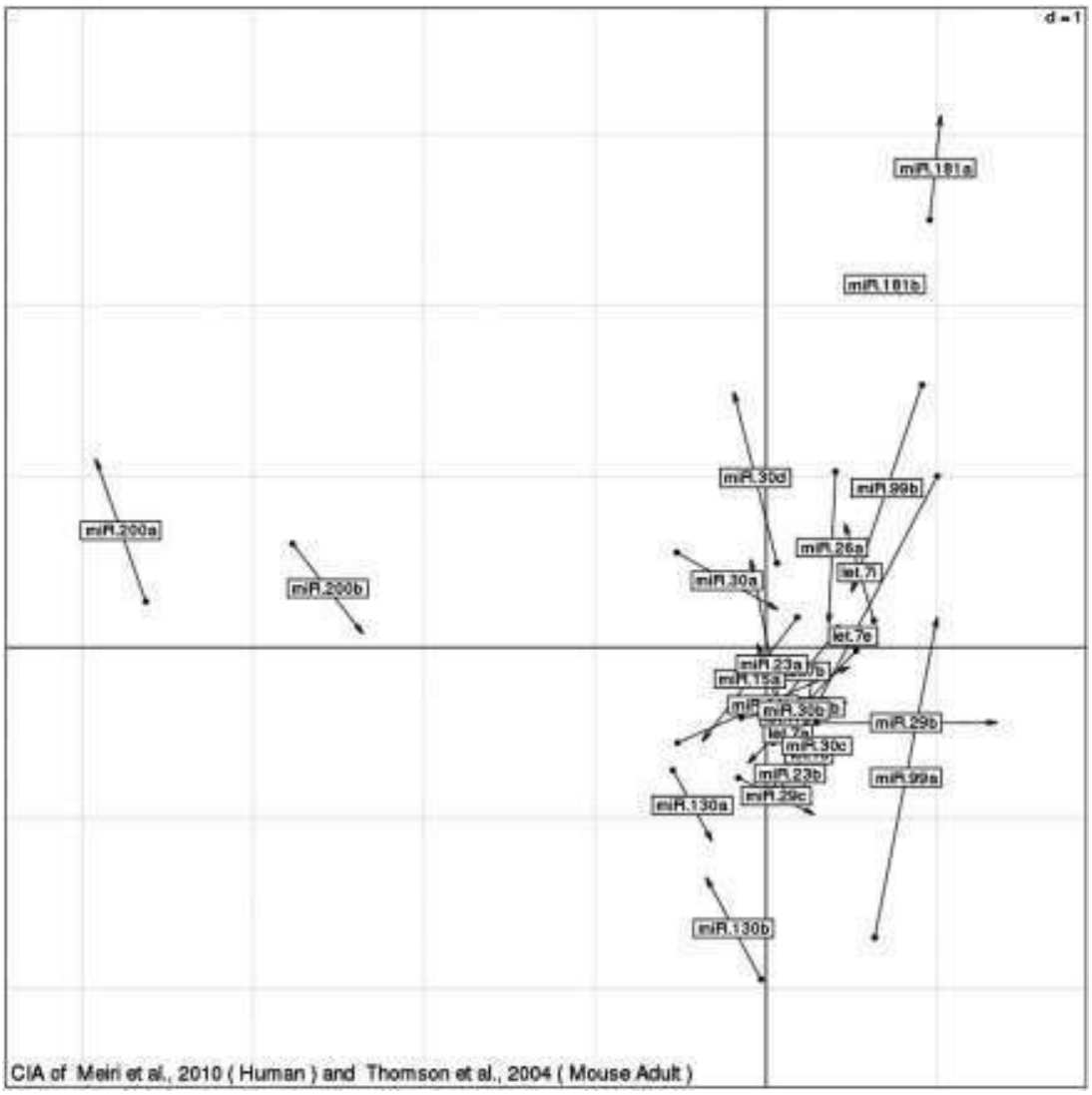


Figure 5. Distribution of microRNAs after dimension reduction by co inertia analysis. microRNAs related in expression clustered together. The length of an arrow correlates with the amount of expression divergence for a particular microRNA between the two data sets, i.e. human versus mouse.

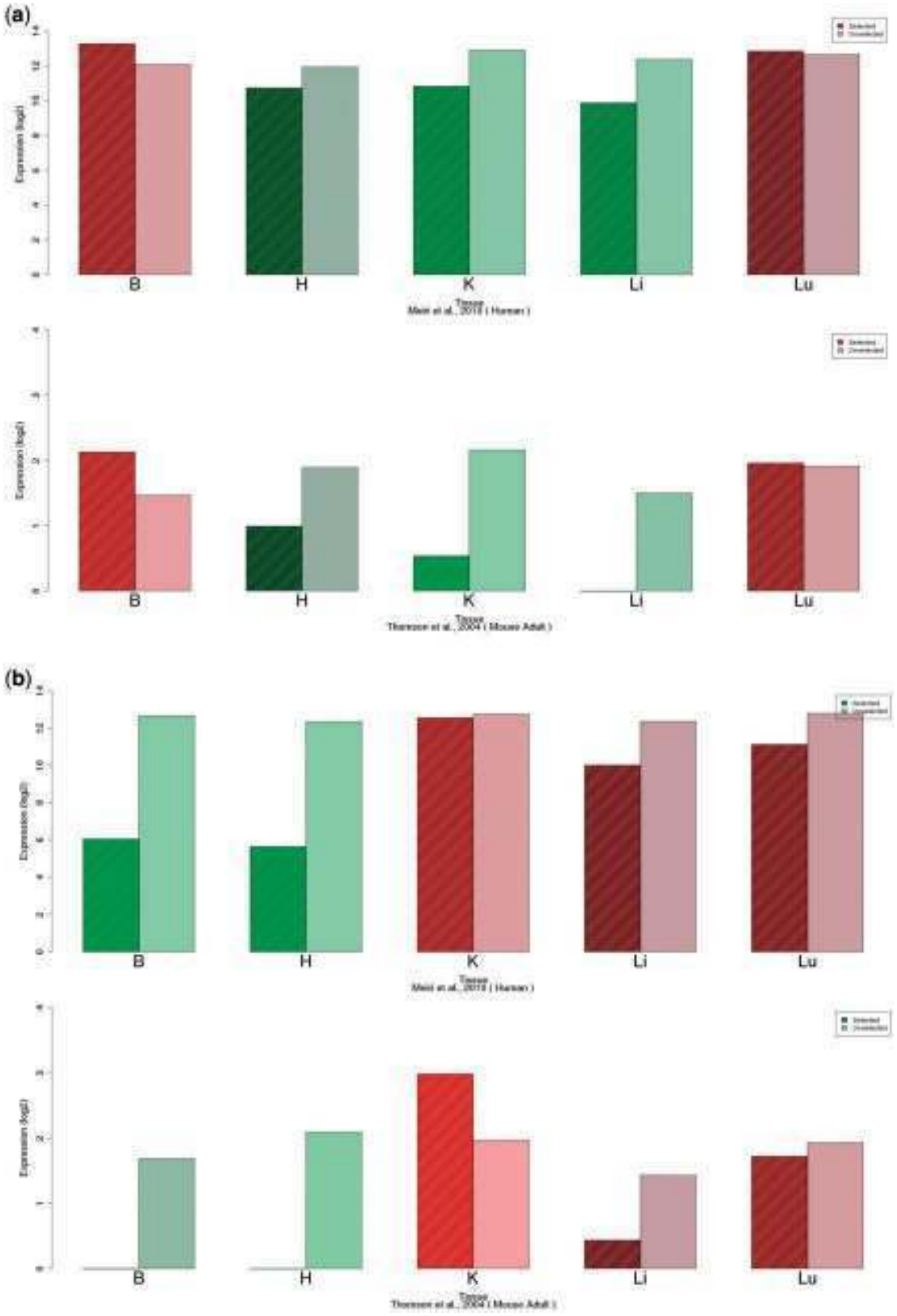


Figure 6. Similarity of expression of microRNA expression from Meiri and Thomson. Expression bar plot of (**a**) *mir-181a* and *mir-181b* cluster (**b**) *miR-200a* and *miR-200b* for five different tissues [i.e. brain (B), heart (H), kidney (K), liver (Li), and lung (Lu), respectively.]. For each tissue, the bar on the left indicates the mean expression of the members of the cluster and the right hand bar indicates the mean expression of the remainder of the data set.

DATABASE USAGE

mESAdb is a highly interactive and flexible database with an ability to analyze and visualize selected expression profiles for a given subset of microRNAs in a multivariate manner using correspondence and co-inertia analyses. One can also study a single microRNA of interest using bar plots associated with a gene expression enrichment index, based on the ϕ-coefficient. This index provides a significance value for the relative enrichment of a microRNA(s) in a particular class with respect to others (Supplementary Data). Furthermore, the user can obtain information about the functional enrichment of a microRNA or a group of microRNAs using different databases, including GO, KEGG and HUGE Navigator.

The default expression data sets currently focus on tissue- and stage-specificity; however, users can add any microarray data containing other types of expression classes, e.g. cancer versus normal, treatment versus control (Figure 3). This allows for great flexibility in analyzing one's own research data.

As an example, we demonstrate that the user can compare two data sets with respect to a list of microRNA clusters that are common to both mice and humans. Using the 'expression–expression' module of mESAdb, we have chosen a human (36) and a mouse (37) data set (Table 1) and uploaded a microRNA list (mESAdb supporting material; http://konulab.fen.bilkent.edu.tr/mirna/supplementary_files.php; the list included *let-7a-i*, *mir-130a-b*, *mir-15a-b*, *mir-181a-b*, *mir-200a-b*, *mir-23a-b*, *mir-26a-b*, *mir-29a-c*, *mir-30a-d* and *mir-99a-b* clusters). We then performed the coinertia analysis using only the tissues common to both data sets, namely, brain (B), liver (Li), lung (Lu), kidney (K) and heart (H) (Figure 4). mESAdb through coinertia analysis allows for comparison and visualization of two expression data sets by plotting them side by side in terms of the expression of selected microRNAs for the given tissues. Accordingly, we found that microRNAs in our list were expressed similarly in human and mouse data sets because the location of the projected tissues closely corresponded between the two plots (Figure 4). mESAdb also enables visualization of the expression of selected microRNAs from both data sets by simultaneously overlaying them on a two-dimensional plot. In this microRNA-oriented view, similarly expressed microRNAs are found closer in space. The analysis of our microRNA list indicated that several microRNAs formed clusters based on their expression, in particular, *mir-181a* and *mir-181b*, and *mir-200a* and *mir-200b* (Figure 5 and Supplementary Data). Indeed, *mir-181a* and *mir-181b* that are similar in sequence and diverging only with 3 nt exhibit a common sequence motif (i.e. AACATTCA) in their first 8 nt. Similarly, *mir-200a* and *mir-200b* are similar in their sequences containing a common motif (i.e. TAA[C][T]ACTG) in their first 8 nt. Using the 'expression analysis' module, *miR-181a* and *miR-181b* were found to be expressed primarily in the brain and lung (Figure 6a) whereas the *miR-200a-b* cluster was clearly expressed mostly in the kidney and lung both in mice and humans (Figure 6b). Our findings suggested that expression patterns of *mir-181a-b* and *mir-200a-b* were highly conserved between human and mice.

In conclusion, mESAdb focuses on providing a meta-analysis tool/database to enhance our understanding in an important field in microRNA biology, i.e. discovery of associations between microRNA sequence and expression. mESAdb is advantageous because it allows interactive analysis of selected subsets of microRNAs in addition to analysis of single microRNA types. Its modular and expandable nature makes mESAdb a unique and functional database for comparative analysis of microRNA sequence and expression.

AVAILABILITY

mESAdb is freely available at http://konulab.fen.bilkent.edu.tr/mirna/. mESAdb is located on a Linux server (Apache/2.2.4; Ubuntu 8.04 LTS, Kernel: 2.6.24-24-server; PHP 5.2.3-1; R-2-11.1) equipped with four Intel® Xeon® CPU E5335, 2.00GHz processors and 8 GB RAM. Microarray data sets incorporated into the mESAdb, as well as R codes used in correspondence analysis are available for download at the mESAdb site.

FUTURE EXTENSIONS

Modular nature of mESAdb allows for incorporation of additional data sets and statistical tools. Future extensions to mESAdb will include addition of microarray data sets from GEO particularly focusing on different aspects of human pathogenesis. Use of R packages enhances the modular nature of the mESAdb thus future addition of statistical and visual tools for sequence/expression/function analysis of microRNAs is planned.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Alper Tolga Kocatas for help in optimizing MySQL queries for faster execution, Sergen Eren for proofreading microarray data set processing, Rengul Cetin-Atalay for providing rack space for the server and Michelle Adams for her helpful comments on the article.

FUNDING

The Scientific and Technological Research Council of Turkey (TUBITAK) and Bilkent University, Ankara. Funding for open access charge: Partially waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **23**, 281–297.
2. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
3. Grimson,A., Farh,K.K., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
4. Iwama,H., Masaki,T. and Kuriyama,S. (2007) Abundance of microRNA target motifs in the 3′-UTRs of 20527 human genes. *FEBS Lett.*, **581**, 1805–1810.
5. Hertel,J., Lindemeyer,M., Missal,K., Fried,C., Tanzer,A., Flamm,C., Hofacker,I.L. and Stadler,P.F. (2006) The expansion of the metazoan microRNA repertoire. *BMC Genomics*, **7**, 25.
6. Bentwich,I., Avniel,A., Karov,Y., Aharonov,R., Gilad,S., Barad,O., Barzilai,A., Einat,P., Einav,U., Meiri,E. *et al.* (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.*, **37**, 766–770.
7. Yu,J., Wang,F., Yang,G.H., Wang,F.L., Ma,Y.N., Du,Z.W. and Zhang,J.W. (2006) Human microRNA clusters: genomic organization and expression profile in leukemia cell lines. *Biochem. Biophys. Res. Commun.*, **349**, 59–68.
8. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
9. Sun,Y., Koo,S., White,N., Peralta,E., Esau,C., Dean,N.M. and Perera,R.J. (2004) Development of a micro-array to detect human and mouse microRNAs and characterization of expression in human organs. *Nucleic Acids Res.*, **32**, e188.
10. Sempere,L.F., Freemantle,S., Pitha-Rowe,I., Moss,E., Dmitrovsky,E. and Ambros,V. (2004) Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome Biol.*, **5**, R13.
11. Houbaviy,H.B., Murray,M.F. and Sharp,P.A. (2003) Embryonic stem cell-specific MicroRNAs. *Dev. Cell*, **5**, 351–358.
12. Liu,C.G., Calin,G.A., Meloon,B., Gamliel,N., Sevignani,C., Ferracin,M., Dumitru,C.D., Shimizu,M., Zupo,S., Dono,M. *et al.* (2004) An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. *Proc. Natl Acad. Sci. USA*, **101**, 9740–9744.
13. Bargaje,R., Hariharan,M., Scaria,V. and Pillai,B. (2010) Consensus miRNA expression profiles derived from interplatform normalization of microarray data. *RNA*, **16**, 16–25.
14. Navon,R., Wang,H., Steinfeld,I., Tsalenko,A., Ben-Dor,A. and Yakhini,Z. (2009) Novel rank-based statistical methods reveal microRNAs with differential expression in multiple cancer types. *PLoS One*, **4**, e8003.
15. Smith,D.D., Saetrom,P., Snove,O. Jr, Lundberg,C., Rivas,G.E., Glackin,C. and Larson,G.P. (2008) Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation. *BMC Bioinformatics*, **9**, 63.
16. Shalgi,R., Lieber,D., Oren,M. and Pilpel,Y. (2007) Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS Comput. Biol.*, **3**, e131.
17. Sood,P., Krek,A., Zavolan,M., Macino,G. and Rajewsky,N. (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc. Natl Acad. Sci. USA*, **103**, 2746–2751.
18. Madden,S.F., Carpenter,S.B., Jeffery,I.B., Bjorkbacka,H., Fitzgerald,K.A., O'Neill,L.A. and Higgins,D.G. (2010) Detecting microRNA activity from gene expression data. *BMC Bioinformatics*, **11**, 257.
19. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
20. Subramanian,A., Kuehn,H., Gould,J., Tamayo,P. and Mesirov,J.P. (2007) GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics*, **23**, 3251–3253.
21. Jiang,Q., Wang,Y., Hao,Y., Juan,L., Teng,M., Zhang,X., Li,M., Wang,G. and Liu,Y. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**, D98–D104.
22. Tsang,J.S., Ebert,M.S. and van Oudenaarden,A. (2010) Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. *Mol. Cell*, **38**, 140–153.
23. Nam,S., Kim,B., Shin,S. and Lee,S. (2008) miRGator: an integrated system for functional annotation of microRNAs. *Nucleic Acids Res.*, **36**, D159–D164.
24. Betel,D., Wilson,M., Gabow,A., Marks,D.S. and Sander,C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res.*, **36**, D149–D153.
25. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
26. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
27. Yu,W., Gwinn,M., Clyne,M., Yesupriya,A. and Khoury,M.J. (2008) A navigator for human genome epidemiology. *Nat. Genet.*, **40**, 124–125.
28. Culhane,A.C., Thioulouse,J., Perriere,G. and Higgins,D.G. (2005) MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics*, **21**, 2789–2790.
29. Kaya,K.D., Karakulah,G., Yakicier,C. and Konu,O. (2007) MicroRNA sequence and expression database. *BMC Syst. Biol.*, **1**, P29.
30. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
31. Griffiths-Jones,S. (2006) miRBase: the microRNA sequence database. *Methods Mol. Biol.*, **342**, 129–138.
32. Ach,R.A., Wang,H. and Curry,B. (2008) Measuring microRNAs: comparisons of microarray and quantitative PCR measurements, and of different total RNA prep methods. *BMC Biotechnol.*, **8**, 69.
33. Baskerville,S. and Bartel,D.P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, **11**, 241–247.
34. Barad,O., Meiri,E., Avniel,A., Aharonov,R., Barzilai,A., Bentwich,I., Einav,U., Gilad,S., Hurban,P., Karov,Y. *et al.* (2004) MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues. *Genome Res.*, **14**, 2486–2494.
35. Beuvink,I., Kolb,F.A., Budach,W., Garnier,A., Lange,J., Natt,F., Dengler,U., Hall,J., Filipowicz,W. and Weiler,J. (2007) A novel microarray approach reveals new tissue-specific signatures of known and predicted mammalian microRNAs. *Nucleic Acids Res.*, **35**, e52.
36. Meiri,E., Levy,A., Benjamin,H., Ben-David,M., Cohen,L., Dov,A., Dromi,N., Elyakim,E., Yerushalmi,N., Zion,O. *et al.* (2010) Discovery of microRNAs and other small RNAs in solid tumors. *Nucleic Acids Res.*, **38**, 6234–6246.
37. Thomson,J.M., Parker,J., Perou,C.M. and Hammond,S.M. (2004) A custom microarray platform for analysis of microRNA gene expression. *Nat. Methods*, **1**, 47–53.
38. Wienholds,E., Kloosterman,W.P., Miska,E., Alvarez-Saavedra,E., Berezikov,E., de Bruijn,E., Horvitz,H.R., Kauppinen,S. and Plasterk,R.H. (2005) MicroRNA expression in zebrafish embryonic development. *Science*, **309**, 310–311.
39. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
40. Lu,J., Getz,G., Miska,E.A., Alvarez-Saavedra,E., Lamb,J., Peck,D., Sweet-Cordero,A., Ebert,B.L., Mak,R.H., Ferrando,A.A. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.
41. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A. *et al.* (2009) NCBI GEO: archive for

high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.
42. Panico,R., Powell,W.H. and Richer,J.C. (eds), (1993) *A Guide to IUPAC Nomenclature of Organic Compounds*. Blackwell Scientific Publications, Oxford.
43. Guilford,J. (1941) The phi coefficient and chi square as indices of item validity. *Psychometrika*, **6**, 11–19.
44. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
45. Hennig,C. and Hausdorf,B. (2004) Distance-based parametric bootstrap tests for clustering of species ranges. *Comput. Stat. Data Anal.*, **45**, 875–895.
46. Robert,P. and Escoufier,Y. (1976) A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Appl. Stat.*, **25**, 257–265.
47. Hartigan,J.A. and Wong,M.A. (1979) Algorithm AS 136: a K-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.*, **28**, 100–108.
48. Lovmar,L., Ahlford,A., Jonsson,M. and Syvanen,A.C. (2005) Silhouette scores for assessment of SNP genotype clusters. *BMC Genomics*, **6**, 35.
49. Kachitvichyanukul,V. and Schmeiser,B. (1985) Computer generation of hypergeometric random variates. *J. Stat. Comput. Simul.*, **22**, 127–145.

Oxford Journals

Access & Purchase

- Oxford Journals
- Access & Purchase
- Publication Rights Policies

Publication Rights Policies

What is our policy?

For the majority of journals[1] published by Oxford University Press, we have a policy of acquiring a sole and exclusive licence for all published content, rather than asking authors to transfer ownership of their copyright, which has been common practice in the past. We believe this policy more carefully balances the interests of our authors with our need to maintain the viability and reputation of the journals through which our authors are accorded status, recognition and widespread distribution. In developing this policy we have been guided by the following principles:

- As a university press and not-for-profit academic publisher, we rely heavily on the good relationships we have with our authors. Having a licensing policy which enables an author to be identified as the owner of the copyright in an article is one of the key ways of demonstrating how highly we value these relationships.
- An exclusive licence enables the centralised and efficient management of permissions and licencing, ensuring the widest dissemination of the content through intermediaries;
- Exclusive rights also enable OUP to take measures on behalf of our authors against infringement, inappropriate use of an article, libel or plagiarism;
- At the same time, by maintaining exclusive rights, in all media for all published content, we can monitor and uphold the integrity of an article once refereed and accepted for publication to be maintained;

Where to get a copy of the Licence to Publish

OUP cannot publish your article until a completed licence form has been received. You should receive a form as soon as your article is accepted for publication.

Footnotes to this section

1. A small number of OUP Journals still have a policy of requesting a full Assignment of Copyright. If unclear about the policy of the Journal concerned, please contact the Editorial office to clarify.

Government employees

- If you are or were a UK Crown servant and the article has been written in that capacity, we have an arrangement with HMSO to enable us to publish it while acknowledging that it is Crown Copyright. Please inform the Editorial office or Oxford University Press at the time of acceptance or as soon as possible that the article is Crown Copyright, so that we can ensure the appropriate acknowledgement and copyright line are used, as required by our arrangement with HMSO.

- If you are a US Government employee and the article has been written in that capacity, we acknowledge that the Licence to Publish applies only to the extent allowable by US law.

Re-use of third party content as part of your Oxford Journals article

- As part of your article, you may wish to reuse material sourced from third parties such as other publishers, authors, museums, art galleries etc. To assist with this process, we have a Permission Request form and accompanying Guidelines that specifies the rights required in order for third party material to be published as part of your Article. For a copy of this form, please email.
- Responsibility for clearing these third party permissions must be borne by the Author, and this process completed as soon as possible - preferably before acceptance of the manuscript, but if not possible, before the Article reaches the Production stage of the process.

Rights retained by ALL Oxford Journal Authors

- The right, after publication by Oxford Journals, to use all or part of the Article and abstract, for their own personal use, including their own classroom teaching purposes;
- The right, after publication by Oxford Journals, to use all or part of the Article and abstract, in the preparation of derivative works, extension of the article into book-length or in other works, provided that a full acknowledgement is made to the original publication in the journal;
- The right to include the article in full or in part in a thesis or dissertation, provided that this not published commercially;

For the uses specified here, please note that there is no need for you to apply for written permission from Oxford University Press in advance. Please go ahead with the use ensuring that a full acknowledgment is made to the original source of the material including the journal name, volume, issue, page numbers, year of publication, title of article and to Oxford University Press and/or the learned society.

The only exception to this is for the re-use of material for commercial purposes, as defined in the information available via the above url. Permission for this kind of re-use is required and can be obtained by using Rightslink:

With Copyright Clearance Center's Rightslink ® service it's faster and easier than ever before to secure permission from OUP titles to be republished in a coursepack, book, CD-ROM/DVD, brochure or pamphlet, journal or magazine, newsletter, newspaper, make a photocopy, or translate.

- Simply visit: www.oxfordjournals.org and locate your desired content.
- Click on (Order Permissions) within the table of contents and/ or at the bottom article's abstract to open the following page:
- Select the way you would like to reuse the content
- Create an account or login to your existing account
- Accept the terms and conditions and permission is granted

For questions about using the Rightslink service, please contact Customer Support via phone 877/622-5543 (toll free) or 978/777-9929, or email Rightslink customer care.

Preprint use of Oxford Journals content

- For the majority of Oxford Journals, prior to acceptance for publication, authors retain the right to make a pre-print [*A preprint is defined here as un-refereed author version of the article*] version of the article available on your own personal website and/or that of your employer and/or in free public servers of preprints and/or articles in your subject area, provided that where possible.
    - You acknowledge that the article has been accepted for publication in [Journal Title] ©: [year] [owner as specified on the article] Published by Oxford University Press [on behalf of

xxxxxx]. All rights reserved.
    - Once the article has been published, we do not require that preprint versions are removed from where they are available. However, we do ask that these are not updated or replaced with the finally published version. Once an article is published, a link could be provided to the final authoritative version on the Oxford Journals Web site. Where possible, the preprint notice should be amended to:
    - This is an electronic version of an article published in [include the complete citation information for the final version of the Article as published in the print edition of the Journal.]
- Once an article is accepted for publication, an author may not make a pre-print available as above or replace an existing pre-print with the final published version. **NB**There are some Oxford Journals such as the Journal of the National Cancer Institute, which do not permit any kind of preprint use. For clarification of the preprint policy for any journal please contact the Rights and New Business Development Department.

Postprint use of Oxford Journals content:

[*A postprint is defined here as being the final draft author manuscript as accepted for publication, following peer review, BUT before it has undergone the copyediting and proof correction process*].

We have detailed policies on the use of postprints for all of our journals. To view these for individual journals please refer to the author self archiving policies on journal homepages. If you require further information please contact the Rights and New Business Development Department.

Other uses by authors should be authorized by Oxford Journals through the Rights and New Business Development Department.

Additional Rights retained by the Author when publishing in an Oxford Open participating journal

Please note that these rights only apply to content published in an Oxford Journal on an Open Access basis in exchange for payment of an author charge. For more details about how Oxford Open works please click here.

The right to reproduce, disseminate or display articles published under this model for educational purposes, provided that:

- the original authorship is properly and fully attributed;
- the Journal and OUP are attributed as the original place of publication with the correct citation details given;
- if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated
- the right to deposit the postprint and/or URL or PDF of the finally published version of the article into an institutional or centrally organized repository, immediately upon publication

Commercial Use of Open Access version

For permission to make any kind of commercial use of the material from the Open Access version of an Oxford Journal (ie.the online version), please contact the Rights and New Business Development department: you want to use and a brief description of the intended use.

Commercial re-use guidelines for open access content

Definition of commercial use: any re-use of material from the Open Access part of an Oxford Journal for the commercial gain of the user and/or their employing institution. In particular,

- re-use by a non-author/third party/other publisher of parts of or all of an article or articles in another publication (journal or book) to be sold for commercial purposes. Permission to reproduce selected figures will generally be granted free of charge, although OUP reserves the right to levy a fee for the use of these and/or the full text of an article/articles
- the proactive supply of multiple print or electronic copies of items taken from the Journal to third parties on a systematic basis for marketing purposes. Permission for this kind of reuse should be obtained from the publisher, who retains the right to levy an appropriate fee
- re-use by an author of parts of or all of an article in other publications from commercial organizations. Permission for this kind of reuse should be obtained from the publisher. We would consider this to be commercial reuse but would not normally charge a permission fee if the author is involved.

NB: Please note that any income generated from permissions granted for this kind of use will be returned directly to the journal itself in order to help minimise the costs of making content from it available on an Open Access basis.

Permissions

- All requests to reuse the article, in whole or in part, in another publication will be handled by Oxford Journals. Unless otherwise stated, any permission fees will be retained by the Journal concerned. Where possible, any requests to reproduce substantial parts of the article (including in other Oxford University Press publications) will be subject to your approval (which is deemed to be given if we have not heard from you within 4 weeks of the permission being granted).
- If copyright of the article is held by someone other than the Author, e.g. the Author's employer, Oxford Journals requires non-exclusive permission to administer any requests from third parties. Such requests will be handled in accordance with Notes 6 above.
- The Journal is registered with the Copyright Licensing Agency (London) and the Copyright Clearance Center (Danvers, Massachusetts), and other Reproduction Rights Organizations. These are non-profit organizations which offer centralised licensing arrangements for photocopying on behalf of publishers such as Oxford University Press.
- Please forward requests to re-use all or part or your article, or to use figures contained within it, to the Rights and New Business Development Department.

My Account

Log in here to manage your account

Services

Keeping you Updated

Illuminea
Read our quarterly newsletter for librarians & information professionals

Changes to our list
New launches, new acquisitions, and titles changing frequency

Frequently asked questions

What is your policy on **perpetual access**?
What is your policy on **digital preservation**?

Why am I having **registration problems**?
Who should I contact about **online access problems**?

Full list of FAQs

Contact

Technical & customer support

Collection sales enquires

Oxford Journals *Oxford University Press*

- Site Map
- Privacy Policy
- Frequently Asked Questions

Other Oxford University Press sites: Oxford University Press

however, as the journal would not be sustainable without income from author charges to cover its costs. The *NAR* Editors hope that contributors to *NAR* will support the journal's open access model by paying the publication charges if they are able to do so.

In general, requests for waivers from authors with funding sources outlined in the 'Funds to pay publication charges' section above cannot be considered.

A waiver request form is available here. Any contributor wishing to submit a waiver or partial waiver request is required to submit the form, endorsed by a senior financial administrator or head of department, directly to Oxford University Press, after his or her manuscript has been accepted for publication. As outlined above, we have a special publication charge policy for authors from developing countries included in the OUP Developing Countries Initiative; for example, all Corresponding authors based in List A countries will automatically be eligible for a full waiver.

NAR OPEN ACCESS LICENCE AGREEMENT

NAR authors are asked to sign an Open Access license agreement which reflects the Open Access model outlined below.

Articles published under this *NAR* Open Access model are made freely available online immediately upon publication, as part of a long-term archive, without subscription barriers to access. We have chosen to implement the Creative Commons Attribution-Non Commercial licence for articles published under the *NAR* Open Access model. This means that users of *NAR* articles published under this Open Access model are entitled to use, reproduce, disseminate or display these articles provided that:

1. the original authorship is properly and fully attributed;
2. the journal and publisher are attributed as the original place of publication with correct citation details given;
3. if an original work is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this is clearly indicated;
4. no articles are reproduced for commercial purposes without the prior consent of OUP (see rights and permissions) and payment to OUP of any appropriate fee.

In our 2004 author survey, we asked *NAR* users to tell us what criteria they felt were important for an Open Access model. Of the 1052 individuals who responded, 75% felt that the unrestricted right to re-use Open Access content for educational and research purposes was important. In contrast, 8% felt that unrestricted re-use by others for commercial purposes was important.

Under *NAR*'s Open Access model, all users have unrestricted rights to re-use Open Access content for educational and research purposes; however, those wishing to re-use content for commercial purposes must continue to obtain permission from OUP as the original publisher, and pay the appropriate fee. We believe that this provision will have several benefits - OUP can continue to act as a central point of contact for commercial re-use requests and will seek to protect the original author and the journal from misuse of published content; revenue resulting from such permission requests will be used by *NAR* to supplement publication charges for authors, thus helping to keep these charges as low as possible; and OUP will be able to monitor commercial re-use that could directly harm the business interests of the journal.

AUTHOR SELF-ARCHIVING/PUBLIC ACCESS POLICY

All new *NAR* content is automatically deposited in PubMed Central and UK PubMed Central, and made freely available via these resources upon publication in the journal. This means that publishing in *NAR* is fully compliant with e.g. the National Institutes of Health (NIH) Public Access policy and the HHMI, UK

MRC and Wellcome Trust policies on open access. Authors wishing to comply with these policies need not take further action.

For more information about *NAR*'s policy, please visit our Author Self Archiving Policy page.

FAQS

FAQ Authors FAQ Librarians and Subscription Agents

CONTACT US

We would welcome any comments or questions you may have - please email us at Open Access.

The Journal

- About this journal
- NAR Methods online
- 2011 Database Issue
- 2011 Web Server Issue
- NAR Special Collections
- Referee Information
- Rights & Permissions
- Dispatch date of the next issue
- This journal is a member of the Committee on Publication Ethics (COPE)
- view Recent Comments on articles

Impact factor: 7.836

5-Yr impact factor: 7.314

Senior Executive Editors

Keith Fox, Southampton, UK

Barry Stoddard, Seattle, WA, USA

- View full editorial board

For Authors

- Instructions to authors
- Scope and Criteria for Consideration
- Online submission instructions
- Submit a manuscript now
- Self-archiving policy