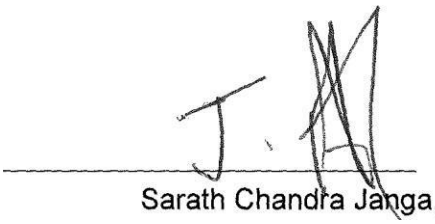
Role of post-transcriptional regulation in human liver

Praneet Chaturvedi

Submitted to the faculty of the Bioinformatics Graduate Program in partial fulfillment of the requirements for the degree Master of Science in Bioinformatics in the School of Informatics and Computing Indiana University February 2015

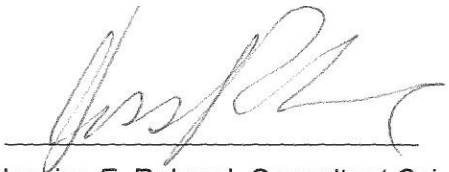
Accepted by the Graduate faculty, Indiana University, in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics

Master's Thesis Committee

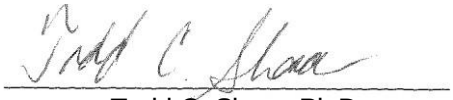


Sarath Chandra Janga, Ph.D.

Lang Li, Ph.D.



Jessica F. Rehmel, Consultant Scientist



Todd C. Skaar, Ph.D.

To my parents, Mr. Anurag Chaturvedi and Mrs. Preeti Chaturvedi and my family. They raised me, supported me, taught me and loved me. This thesis is dedicated to them.

Acknowledgement

This thesis was made possible due to the guidance and encouragement from many people. So it gives me a great pleasure to thank these people and acknowledge their contribution. I owe a sincere appreciation to my thesis advisor Dr. Sarath Chandra Janga for his motivation and immense knowledge. Without his thoughtful guidance, energy and constructive criticism this thesis would have never been possible. Besides my lab members, I would also like to thank the other members of my committee Dr. Lang Li, Jessica F. Rehmel and Dr. Skaar for supporting my work, reading my thesis and providing helpful suggestions.

I wish to express sincere appreciation to School of Informatics at Indiana University Purdue University Indianapolis for providing me an opportunity to pursue a bright carrier in bioinformatics.

I thank my family for the constant support, love and blessing. Their teachings have made me the person I am today.

Praneet Chaturvedi

Role of Post-transcriptional regulation in human liver

Abstract

My thesis comprises of two individual projects which revolve around the importance of post-transcriptional regulation in liver. My first project is studying the integrated miRNA – mRNA network in NAFLD. For fulfillment of the study we conducted a genome-wide study to identify microRNAs (miRs) as well as the miR-mRNA regulatory network associated with hepatic fat and NAFLD. Hepatic fat content (HFC), miR and mRNA expression were assessed in 73 human liver samples. Liver histology of 49 samples was further characterized into normal (n=33) and NAFLD (n=16). Liver miRNome and transcriptome were significantly associated with HFC and utilized to (a) build miR-mRNA association networks in NAFLD and normal livers separately based on the potential miR-mRNA targeting and (b) conduct pathway enrichment analyses. We identified 62 miRs significantly correlated with HFC ($p < 0.05$ with $q < 0.15$), with miR-518b and miR-19b being most positively and negatively correlated with HFC, respectively ($p < 0.008$ for both). Integrated network analysis showed that six miRs (miRs-30b*, 612, 17*, 129-5p, 204 and 20a) controlled ~ 70% of 151 HFC-associated mRNAs ($p < 0.001$ with $q < 0.005$). Pathway analyses of these HFC-associated mRNA revealed their key effect ($p<0.05$) in inflammation pathways and lipid metabolism. Further, significant ($p<2.47e-4$, Wilcoxon test) reduction in degree of negative associations for HFC-associated miRs with HFC-associated mRNAs was observed in NAFLD as compared to normal livers, strongly suggesting highly dysfunctional miR-mRNA post-transcriptional regulatory network in NAFLD. Our study makes several novel observations which provide clues to better understand the pathogenesis and potential treatment targets of NAFLD.

My second project is based on uncovering important players of post-transcriptional regulation (RBPs) and how they are associated with age and gender during healthy liver development. For this study, we performed an association analysis focusing on the expression changes of 1344 RNA Binding proteins (RBPs) as a function of age and gender in human liver. We identify 88 and 45 RBPs to be significantly associated with age and gender respectively. Experimental verification of several of the predicted associations in the mouse model confirmed our findings. Our results suggest that a small fraction of the gender-associated RBPs (~40%) are likely to be up-regulated in males. Altogether, these observations show that several of these RBPs are important developmentally conserved regulators. Further analysis of the protein interaction network of RBPs associated with age and gender based on the centrality measures like degree, betweenness and closeness revealed that several of these RBPs might be prominent players in liver development and impart gender specific alterations in gene expression via the formation of protein complexes. Indeed, both age and gender-associated RBPs in liver were found to show significantly higher clustering coefficients and network centrality measures compared to non-associated RBPs. The compendium of RBPs and this study will help us gain insight into the role of post-transcriptional regulatory molecules in aging and gender specific expression of genes.

Table of contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Understanding the post-transcriptional regulation

Post-transcriptional regulation is mainly carried out by two important players 1) miRNAs 2) RBPs. Post-transcriptional control is important when degradation of transcripts is required, providing support or translocation ability to transcripts, translation to proteins etc. also it is very important in organ development and when studying causes of disease in an organism. Let's see what miRNAs are : MicroRNAs (miRs) are small (~20nt) non-coding RNA molecules that regulate gene expression through either transcript degradation [1] or translational repression [2]. miRs have been predicted to regulate 30% of human genes [3] with bioinformatics analysis indicating that thousands of genes could be controlled by single miR [4] forming a dense network of post-transcriptional interactions [5]. In particular, miRs are documented as key players in the regulation of genes involved in lipid metabolism [6], glucose metabolism [7], energy homeostasis [8] and many other related biological processes like cell proliferation, apoptosis, lymphocyte development, adipocyte differentiation, and insulin secretion [9]. Several studies also reveal that miRs could act as oncogenes [10] or tumor suppressors [11] with their dysregulation playing significant roles in promoting hepatocellular carcinoma [12]. On the other hand RBPs play different functions and are very important at post-transcriptional control/regulation. RNA Binding Proteins (RBPs) bind to RNA molecules to control different post transcriptional processes such as pre-mRNA splicing, mRNA cytoplasmic export, turnover, storage, and translation [13, 14]. Thus, the capacity of these proteins to influence gene expression at post-transcriptional level is extremely important especially during the developmental

process to give rise to complex organs and tissues[15, 16].

1.2 What is NAFLD

Non-Alcoholic Fatty Liver Disease (NAFLD) has emerged as a global problem with an incidence of ~ 30% in adults and ~ 10% in children and adolescents [17]. NAFLD can be broadly categorized into simple steatosis and steatohepatitis (NASH). Simple steatosis is generally thought to be benign whereas NASH can progress to cirrhosis, liver failure and liver cancer [18-20]. Over the last decade, there have been many important advances which expanded our understanding of the pathogenesis of NAFLD and NASH, but significant knowledge gaps remain.

1.3 Why miRNAs are of importance in NAFLD

A number of miRs have been identified to be associated with NAFLD and/or NASH in human livers [21, 22] as well as animal models [23, 24]. However, these studies have generally focused on miRs in a unidimensional fashion. Given the complexity of the interactions between miRs and mRNAs and the dependence of the function of miRs on mRNAs, it is of critical importance to understand the miR-mRNA regulatory network simultaneously in the same set of samples.

1.4 Why RBPs are of importance in organ development

RBPs play a substantial role in mediating developmental changes of a mammalian cell. For instance, PTBP1 (polypyrimidine tract binding protein), a ubiquitous protein known to be important in mammalian development at early stages of gastrulation [25, 26] and ELAVL1 (HuR) - a protein that acts as an mRNA stability factor, is also

known for its role in placental branching, embryonic and neuronal development. [27, 28]. Likewise, CRD-BP (IGF2BP1), a member of the insulin-like growth factor 2 mRNA-binding protein family, is the first example of a putative mammalian mRNA-binding proteins that is abundant in fetal tissue but absent in the adult tissue [29]. Furthermore, a study on understanding the regulation of HNF4alpha in liver development revealed that the expression of HNF4aplha is widely regulated by the sequential promoter usage and alternative splicing in the 3' end to produce different isoforms important for the liver development[30]. Yet another study identified UPF2, one of the key players of the non-sense mediated mRNA decay (NMD) machinery, as a critical regulator of the liver development [31]. Thus, although specific RBPs have been studied for their role in mediating developmental processes of liver, no global association analysis has been performed in humans to uncover the repertoire of RBPs contributing to liver development.

1.5 What we have done

For the first study that is understanding integrated miR – mRNA network in NAFLD we have conducted a study with following objectives; **(a)** to assess genome-wide miRs and mRNAs that are significantly associated with hepatic fat concentration (HFC) in human liver tissue; **(b)** to build miR-mRNA association networks based on the potential miR-mRNA targeting that distinguish NAFLD from normal liver tissue; and **(c)** to conduct a pathway enrichment analysis of NAFLD.

For the second study that is uncovering the RBPs associated as a function of age and gender during liver development we did a genome-wide association analysis of their expression patterns with age and gender was conducted. To achieve this, we performed an association analysis of RBPs expression levels in human liver tissues with respect to age and gender, by integrating a dataset of 1344 genes

experimentally known to encode for RBPs. This allowed us to catalogue for the first time age and gender-related associations for RBPs as well as to study the wiring patterns of RBPs and their protein complexes (protein-protein interaction networks) with liver development in the human genome.

Chapter 2 Integrated Network Analysis Uncovers Key microRNAs, microRNA-mRNA Regulatory Networks and Related Pathways Associated with Human Hepatic Fat Accumulation

2.1 Introduction

Non-Alcoholic Fatty Liver Disease (NAFLD) has emerged as a global problem with an incidence of ~ 30% in adults and ~ 10% in children and adolescents [17]. NAFLD can be broadly categorized into simple steatosis and steatohepatitis (NASH). Simple steatosis is generally thought to be benign whereas NASH can progress to cirrhosis, liver failure and liver cancer [18-20]. Over the last decade, there have been many important advances which expanded our understanding of the pathogenesis of NAFLD and NASH, but significant knowledge gaps remain.

MicroRNAs (miRs) are small (~20nt) non-coding RNA molecules that regulate gene expression through either transcript degradation [1] or translational repression [2]. miRs have been predicted to regulate 30% of human genes [3] with bioinformatics analysis indicating that thousands of genes could be controlled by single miR [4] forming a dense network of post-transcriptional interactions [5]. In particular, miRs are documented as key players in the regulation of genes involved in lipid metabolism [6], glucose metabolism [7], energy homeostasis [8] and many other related biological processes like cell proliferation, apoptosis, lymphocyte development, adipocyte differentiation, and insulin secretion [9]. Several studies also reveal that miRs could act as oncogenes [10] or tumor suppressors [11] with their dysregulation playing significant roles in promoting hepatocellular carcinoma [12]. To date, a number of miRs have been identified to be associated with NAFLD and/or NASH in human livers [21, 22] as well as animal models [23, 24]. However, these studies have generally focused on miRs in a unidimensional fashion. Given the complexity of the interactions between miRs and mRNAs and the dependence of the function of miRs on mRNAs, it is of critical importance to understand the miR-mRNA regulatory network simultaneously in the

same set of samples.

In order to further our understanding of pathogenesis of fat accumulation and NAFLD in humans, we have conducted a study with following objectives; **(a)** to assess genome-wide miRs and mRNAs that are significantly associated with hepatic fat concentration (HFC) in human liver tissue; **(b)** to build miR-mRNA association networks based on the potential miR-mRNA targeting that distinguish NAFLD from normal liver tissue; and **(c)** to conduct a pathway enrichment analysis of NAFLD.

2.2 Materials and Methods

2.2.1 Samples and genomic datasets

Seventy three human liver tissue samples collected from transplant donors were used in this study. The sample procurement procedure and related information were described in our previous study [32, 33]. In brief, these transplant donors lacked heavy alcohol consumption and tested negative for viral hepatitis B and C. Their demographics and available clinical characteristics were shown in **Table S1**. The mRNA expression was analyzed using the Agilent expression arrays (Agilent-014850 4×44k arrays, GPL4133) [32], and expression of 850 miRNAs was analyzed using the Exiqon miRCURY™ LNA Array v10.0 (Exiqon, Inc., Denmark) [33]. Purdue University IRB has approved this study.

2.2.2 Measurement of hepatic fat content

Total hepatic fat content (HFC) of 73 human liver samples was quantified using a modified hexane/isopropanol method according to our previously established protocol [34]. Total hepatic protein content among these samples was also measured by the standard Bradford method. Total HFC was then normalized to the protein content and expressed as fat/protein weight ratio.

2.2.3 Histopathological evaluation

Sufficient liver tissue was available from 49 liver samples for histopathological assessment. Approximately, 10 mg of their liver tissue was fixed in 10% formalin and embedded in paraffin. The tissue blocks were then sectioned and stained with H&E and Masson's trichrome stains. An expert hepatopathologist in a blinded fashion classified samples into normal (n=33) and NAFLD (n=16). The latter was further sub-classified into fatty liver (n=2), borderline NASH (n=9), and NASH (n=5) categories based on the NASH CRN criteria [35] (Table S1). Additional histological characterization included estimation for the degree of steatosis (%), hepatocytes ballooning, lobular inflammation, Mallory bodies, NAFLD Activity Score (NAS) and fibrosis. Histological examination revealed no other incidental causes for chronic liver disease (e.g., iron overload, alpha-1 antitrypsin inclusions, autoimmune hepatitis, granulomas or biliary tract disease).

2.2.4 Association of miRs and mRNA expression profiles with hepatic fat content

Expression profiles of 850 miRs of 73 liver samples were compared with their HFC to identify the miRs that are most correlated with HFC. Spearman correlation was calculated between miR expression and HFC, to identify miRs significantly ($p < 0.05$) correlated with HFC with q-value less than 0.15 using the fdrtool [36].

In a separate analysis, spearman correlations were calculated for the complete transcriptomic data of 73 liver samples to identify the transcripts that are most correlated with HFC. Significantly correlated ($p < 0.001$ and q-value less than 0.005 using fdrtool [36]) mRNAs (genes) were analyzed by Panther-GO ontology bioinformatics tool to categorize the functional classes of these genes in biological processes [37]. These HFC-associated genes were considered for further pathway enrichment analysis by using bioinformatic tools including Panther [38], DAVID [39] and G-profiler [40].

2.2.5 Prediction of genes targeted by miRs

Both miRs and mRNAs which were found correlated with HFC values across 73 samples were further analyzed for their possible miR–target mRNA association using miR target prediction tools, TargetScan [4] and miRanda [41, 42]. TargetScan prediction is based on the occurrence of conserved 7-8mer seed region of a miRNA in 3'-untranslated region (UTR) of its biological targets, followed by their rank based scoring of conserved overlapping probability [43]. Similarly, miRanda is another algorithm which allows the prediction of genome-wide targets of miRs and is largely complementary to TargetScan as it does not use conservation information. In particular, miRanda is a machine learning method for ranking microRNA target sites by a down-regulation score by integrating gene expression and sequence level features. The algorithm trains a regression model on sequence and contextual features extracted from high confidence miRanda-predicted target sites and hence can predict non-canonical and non-conserved sites as well [42]. To predict the mRNA targets of each of the miRNAs in this study, 3'-UTR for the complete set of genes in the human genome were downloaded from Ensembl biomart [44, 45] and fasta sequences for miRs were downloaded from miRBase [46]. miRNA target sites on each of the mRNAs correlated with HFC were predicted using both the methods at default thresholds. A gene was considered to be targeted by a miRNA if either of the methods predicted it as a target. This allowed the construction of a HFC associated miRNA-mRNA interaction network which was visualized using cytoscape [47]. Association between each miR and mRNA correlated with HFC represented as a matrix, was also shown as a heatmap using Java Tree view [48]. Also, a subset of this interaction network was further generated to show most important genes and associated miRs that could be involved in various biological pathways associated with HFC.

2.2.6 Investigating the relationship between HFC and other variables

Using a nonparametric test (Wilcoxon test), the HFC was compared between normal and NAFLD (and its subclasses) and are represented as box plots. Similar approach was adopted to investigate the influence of additional variables like age, gender and race on HFC. Since gender and race were not found to be associated with HFC, the expression profiles of the 62 miRNAs significantly correlated with HFC across the 73 liver samples together with the age were used for a Principal Component Analysis (PCA) to test the influence of age in explaining the association of miRNAs with HFC.

2.2.7 Dynamics of miR-mRNA association network in NAFLD

We constructed miR-mRNA association networks for NAFLD (n=16) and normal (n=33) groups separately by calculating the correlation between miRNA and mRNA levels across the samples. This was achieved by calculating the Spearman correlation between each pair of miRNA-mRNA expression levels across the 16 NAFLD samples and 33 normal samples and identifying those pairs which exhibited a negative correlation with a p-value < 0.001 threshold. The degree of association of each miR with mRNAs was defined for each group as the number of mRNAs which are negatively correlated in expression at a significant correlation ($p < 0.001$). Additionally, degree ratio in miR-mRNA association networks for these two groups for each miR was calculated. We also employed the miRNA target predictions identified by both the methods (TargetScan and miRanda) to further filter the above association networks to contain only the predicted targets and obtain the ratio of the number of mRNA targets between the two groups. This analysis allowed us to visualize the global effect of the dysregulation of miRNAs independently found to be associated with HFC in two groups separately.

2.3 Results

Overall study scheme and sample size for various analyses are shown in Figure 1. In brief, HFC, miR and mRNA analyses were performed on 73 human liver samples whereas histological characterization was performed on 49 liver samples.

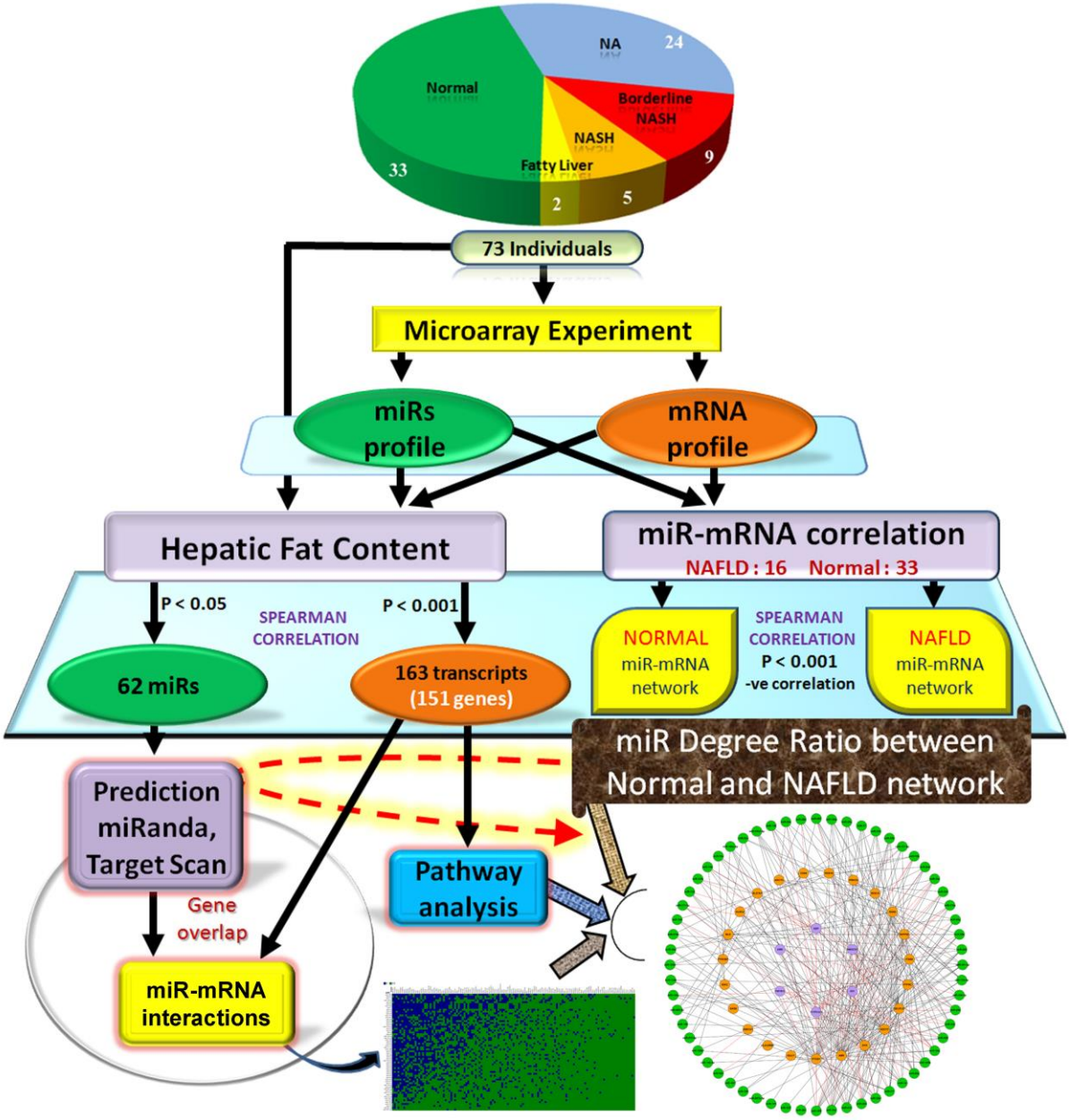


Figure 1. Work flow. Liver samples were collected from 73 organ transplant donors for our study. The miRnome and transcriptome profiles for these 73 samples were generated by

microarray experiments. Subsequently, miRnome and transcriptome data were correlated (Spearman correlation) separately with hepatic fat content (HFC) of 73 individuals. HFC associated 163 mRNAs (151 genes) were analyzed for their enrichment in several biological pathways. Additionally 129 targets, predicted by miRanda or TargetScan, miRNA target prediction algorithms, for HFC associated 62 miRs were found as an overlapped subset of HFC associated mRNAs (151 genes). HFC affected miR-mRNA interaction network (62 miRs connected to 129 genes with 1972 edges) was generated and a subset of this network was studied for HFC influenced dysregulated biological pathways. Similarly, in a subset of 49 samples (33 Normal and 16 NAFLD), miR and mRNA expression data were correlated ($p < 0.001$) separately for each group and degree ratio between negatively correlated miR-mRNA interaction for normal and NAFLD groups was calculated and interpreted to understand the dynamics of post transcriptional regulatory network in NAFLD.

2.3.1 A significant number of miRs were associated with hepatic fat accumulation

Our analysis identified 62 miRs significantly correlated with HFC levels ($p < 0.05$ with $q < 0.15$). Table S2 shows a complete list of these miRNAs along with their correlation coefficient and significance. Notably, many miRs previously described to be functionally important for liver disorders (e.g., fatty liver, hepatitis, cirrhosis and liver neoplasia) were among these 62 significant miRs (miR-34a, miR-142-3p, miR-150 and miR-122 were found to be positively correlated while miR-26a/b, miR-215-5p, miR-194, miR-101, miR-30b/e were negatively correlated with HFC). Selected set of these positively and negatively correlated miRNAs with experimental support from the literature are shown in Table 1. Besides these miRs with known function in liver diseases, we also identified a few new miRs including miR-518b and miR-19b which exhibited most significant positive and negative correlations with HFC, respectively ($p < 0.009$ for both) (Table 1).

Table 1: Selected set of HFC associated miRs with their documented roles in various liver disorders. (NA refers to data/ study not available)

miR	Correlation indices	p value	Liver Disorders
miR-518b	0.681965	8.64E-03	NA
miR-497	0.529583	5.16E-02	NA

miR-204	0.467266	3.02E-02	NA
miR-339-5p	0.433866	1.21E-02	NA
miR-150	0.397664	4.60E-02	Liver Neoplasms [49] Liver Diseases [50] Liver Cirrhosis [51] Hepatoblastoma [52]
miR-142-3p	0.347483	3.78E-03	Liver Diseases [53]
miR-129-5p	0.339155	4.00E-02	NA
miR-34a	0.337861	3.11E-03	Alcoholic liver disease [54] Fatty Liver [55] Liver Diseases [55] Liver Cirrhosis [54] Hepatitis [56] Liver Neoplasms [56] Hepatitis C [57]
miR-381	0.33767	1.01E-02	NA
miR-30b*	0.32129	5.47E-03	Hepatitis [58]
miR-26b	-0.38147	7.77E-04	Liver Neoplasms [59]
miR-30b	-0.38153	7.53E-04	Hepatitis [58]
miR-30e	-0.38812	5.41E-04	Hepatitis [58]
miR-19a	-0.38933	5.10E-04	NA
miR-378	-0.39209	4.25E-04	NA
miR-101	-0.39378	4.45E-04	Hepatitis B [60] Liver Cirrhosis [61] Liver Neoplasm [62]
miR-194	-0.39651	3.70E-04	Hepatitis [63] Liver Cirrhosis [51]
miR-215-5p	-0.41136	1.98E-04	Hepatitis [64] Liver Cirrhosis [64]
miR-26a	-0.41603	2.69E-04	Hepatitis B [65] Hepatitis Chronic [65] Liver Neoplasms [65] Liver Diseases [66] Biliary Cirrhosis [67]
miR-19b	-0.42251	1.41E-04	NA

2.3.2 HFC associated dysregulation of mRNAs

We calculated the correlation between HFC of 73 liver samples and their hepatic transcriptomes which allowed the identification of mRNAs most correlated with fat accumulation. A total of 163 probes corresponding to 151 genes were significantly correlated ($p < 0.001$ with $q < 0.005$) with HFC and are listed in Table S3. A selected set of highly

correlated mRNAs with HFC (10 positively and 10 negatively) are shown in Table 2.

Table 2: Selected set of mRNAs (genes) highly (10 positively and 10 negatively) correlated with hepatic fat concentration (with a brief description)

mRNA probe ID	UniProt Gene name	Correlation indices	p-value	Description (Using ensembl.org)
5052	STMN2	0.465248	3.36E-05	Stathmin-like 2
20785	LAMA1	0.46079	4.09E-05	Laminin, alpha 1
1488	MYO1F	0.460682	4.11E-05	Myosin IF
24614	PTGS1	0.450886	6.25E-05	Prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cyclooxygenase)
3320	GAS6	0.444646	9.11E-05	Growth arrest-specific 6
10883	CLSTN1	0.444546	8.14E-05	Calsyntenin 1
975	IGKC	0.438066	1.06E-04	Immunoglobulin kappa constant
14674	CD86	0.43404	1.25E-04	CD86 molecule
19605	EEF1A2	0.432143	1.35E-04	Eukaryotic translation elongation factor 1 alpha 2
16262	DKFZP434G032	0.429165	1.52E-04	Keratin 23 (histone deacetylase inducible)
3472	FBXO5	-0.42076	2.10E-04	F-box protein 5
25839	IPO5	-0.42378	1.87E-04	Importin 5
7496	EFCAB3	-0.42458	1.81E-04	EF-hand calcium binding domain 3
6656	CYP20A1	-0.42482	2.00E-04	Cytochrome P450, family 20, subfamily A, polypeptide 1
22079	C16ORF35	-0.42893	1.53E-04	Nitrogen permease regulator-like 3 (S. cerevisiae)
6102	CYP51A1	-0.42998	1.47E-04	Cytochrome P450, family 51, subfamily A, polypeptide 1
12276	KRTCAP3	-0.43128	1.39E-04	Keratinocyte associated protein 3
19215	ZNF518B	-0.446	7.67E-05	Zinc finger protein 518B
18862	EIF4E2	-0.45567	5.10E-05	Eukaryotic translation initiation factor 4E family member 2
1325	CXORF57	-0.48472	1.39E-05	Chromosome X open reading frame 57

2.3.3 Potential targets that could be controlled by combined effect of HFC associated miRs

Target prediction for 62 HFC associated miRNAs was performed by using two publicly available prediction tools TargetScan and miRanda as described in Materials and Methods

(Table S4). The list of target genes for these miRNAs which overlapped with the list of mRNAs also significantly correlated with HFC (from Table S3), were used to generate miRNA-mRNA interactions. This is represented as a heatmap of miR-mRNA interactions showing 62 miRs targeting among 129 protein coding genes (mRNA), each also significantly correlated with HFC (Figure 2 and Table S5). Interestingly, six miRs (miRs-30b*, 612, 17*, 129-5p, 204 and 20a) were found to control ~ 70% of 151 HFC-associated mRNAs in this network, with an average connectivity of ~32 targets for each miRNA indicating the dense networking between these HFC-associated miRNAs and mRNAs. The degree of association of each of the miRs in this resulting interaction network is shown in Table S6, among which miR-30b* was found to be highly associated and miR-487b-3p was least connected.

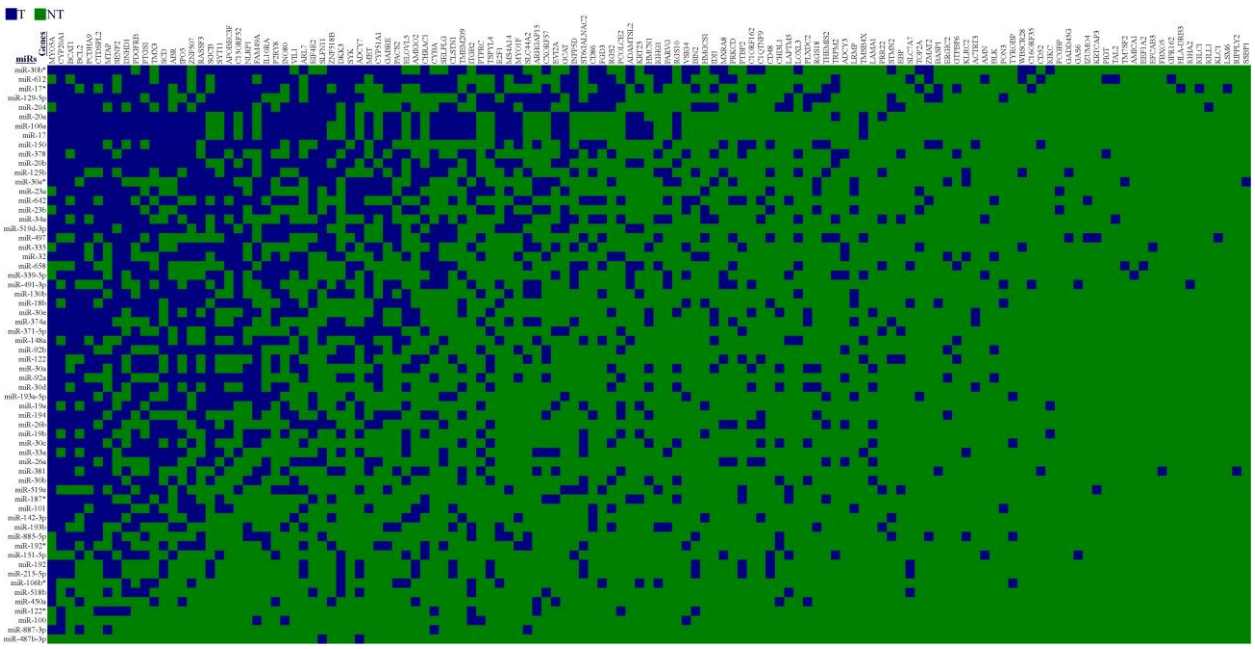


Figure 2. HFC associated miR-mRNA (gene) interaction network. miRnome and transcriptome expression profiles were correlated separately with HFC of 73 individuals and each miR predicted target mRNA (gene) is shown as miR-mRNA heatmap. It includes 62 HFC associated miRs and 129 HFC associated target genes with blue colored box as target and green colored box as non-target.

2.3.4 Role of HFC associated mRNAs in biological processes

Panther-Gene Ontology analysis provided an initial snapshot of the involvement of HFC-associated 151 genes in various classes of biological processes. Figure 3a shows the various biological processes and their contribution to the list of 151 genes identified to be associated with HFC, with metabolic processes (45.3%), cellular processes (40.2%), and cell communication (32.5%), immune system processes (18.0%) forming the core set of processes associated with fat accumulation (Table S7 shows all the biological processes identified in this analysis).

HFC associated 151 genes (mRNAs) were further analyzed independently by three different pathway enrichment tools, namely Panther [38], DAVID [39] and G-profiler [40] functional analysis systems to have a comprehensive and unbiased overview of the enriched functional themes. This analysis revealed that inflammation pathways mediated by chemokine and cytokine signaling, Wnt signaling, B cell activation, NK cell mediated cytotoxicity and lipid metabolism pathways were significantly enriched ($p < 0.05$). List of enriched pathways identified by each of these tools are documented in Tables S8, S9 and S10 respectively and also shown as pie charts (Figure 3b, 3c and 3d). One of the highly enriched pathways is shown in Figure 3e with several HFC associated genes marked in red. This pathway comprises of several cell adhesion molecules involved in various cascades of immunity and signaling pathways like T-cell receptor, complement and coagulation cascade and leukocyte transendothelial migration of dendritic cells, macrophages, TH cell, B cell, etc. Additionally, some of affected pathways and the HFC associated genes identified in them are shown in Figures S1, S2 and S3.

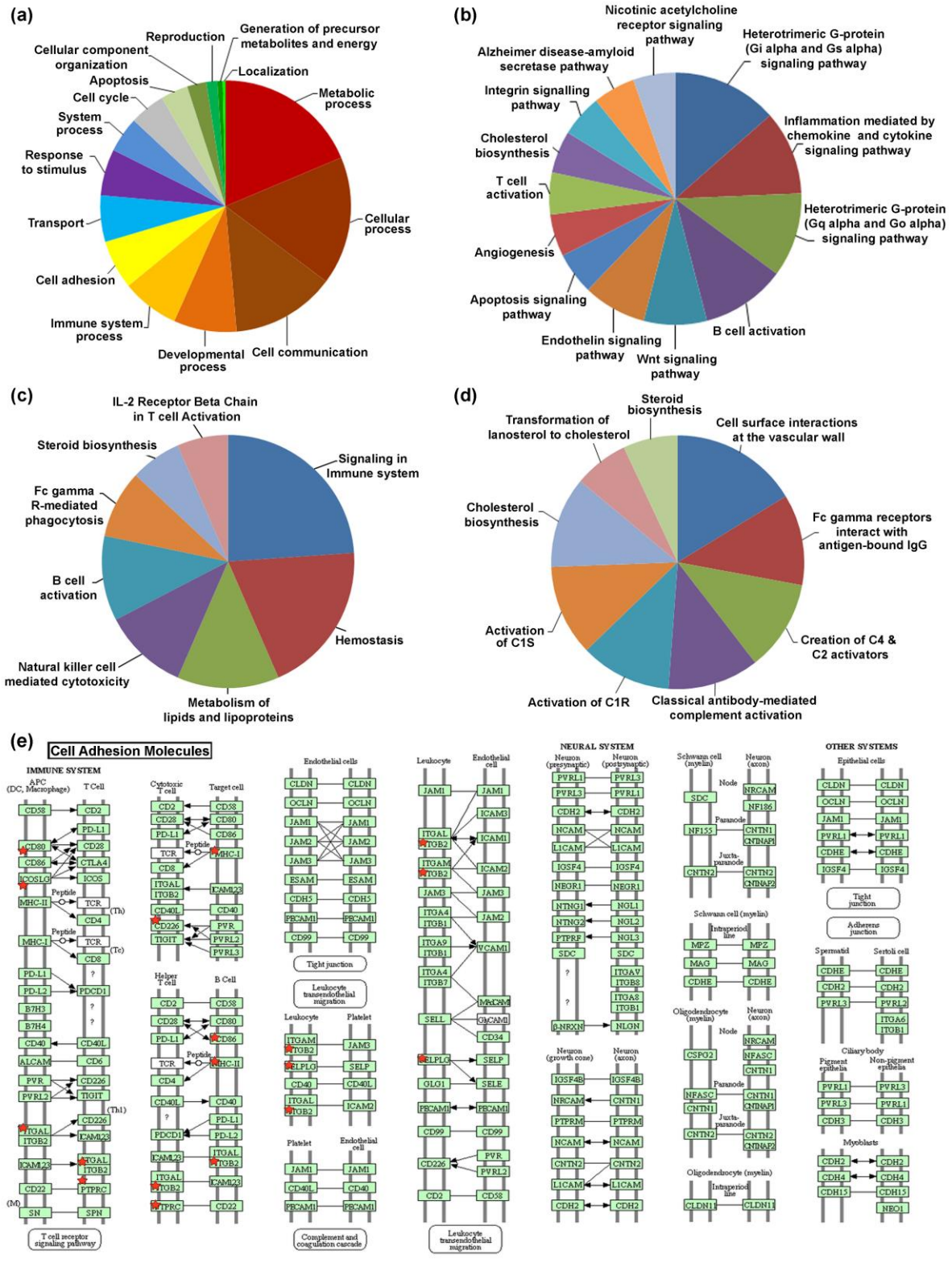


Figure 3. Role of HFC associated genes (mRNA) in several biological pathways. HFC associated 151 genes were analyzed for possible biological significance. These genes were analyzed for their role in various biological processes using (a) PANTHER-Gene ontology

tool and their highly enriched pathways predicted by (b) PANTHER (c) DAVID and (d) g-Profiler tools are shown as pie charts respectively; (e) One of the pathways predicted by DAVID, includes cell adhesion molecules, that could be highly affected by dysregulation of associated genes (red asterisk) is shown in figure 3(e).

Two major classes of biological processes, lipid metabolism and immune response and signaling were identified in pathway analyses as recurring functional themes with high significance and gene count. Figure 4 shows a subset of miR-mRNA interactions, where the genes interacting with miRs involved in lipid metabolism (6 genes) and immune response & signaling (23 genes) are highlighted.

2.3.5 Relationship between HFC and NAFLD

Hepatic fat content was examined for notable change among the normal and NAFLD groups, as well as the NAFLD subclasses, fatty liver, borderline NASH and NASH. As expected, compared to normal liver samples, HFC was significantly higher in the NAFLD group, but HFC did not significantly differ among different NAFLD subgroups (Figure 5a, Figure S4).

2.3.6 Relationship between HFC and selected co-variables

We also explored for the confounding factors like age, gender and race, which might contribute to the variation of HFC across samples. We found no significant effect of race and gender on HFC content across samples as shown in Supplementary Figure S5. However, age was found to be associated with HFC especially when liver samples from individuals' ≥ 45 years were compared to samples ≤ 18 years ($p < 0.002$, Figure 5b). In order to test whether age contributes to the variance in expression levels of the 62 identified HFC-associated miRNAs, PCA (Principle component analysis) was performed for examining the additive influence of age on the dysregulation of miRs across samples. This analysis

unambiguously revealed that more than 60% of the variance was explained by the first two principal components and age had small contribution to the loadings of either of the two components, as shown in Figure S6. Indeed, age did not have significant loading values in the first four principal components which together were able to explain more than 70% of the variance (Table S11).

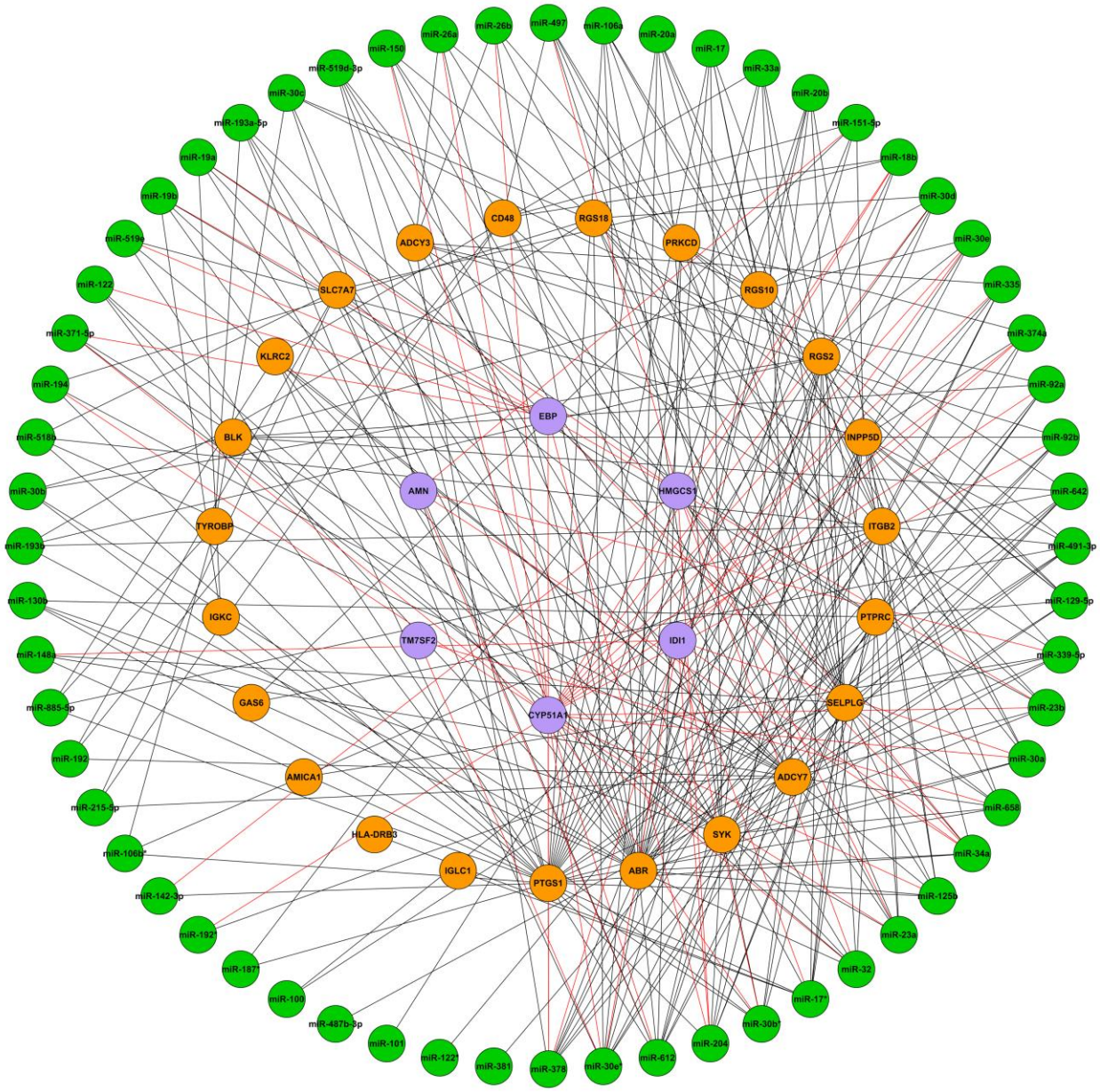


Figure 4. Pathway specific post transcriptional network**.** HFC associated 62 miRs targeting genes, involved in important pathways viz. 23 genes in immune response and signaling (middle circle- orange) and 6 genes in lipid metabolism (inner circle- purple), was shown in an miR-target gene interaction network corresponding to a total of 327 connections (edges).

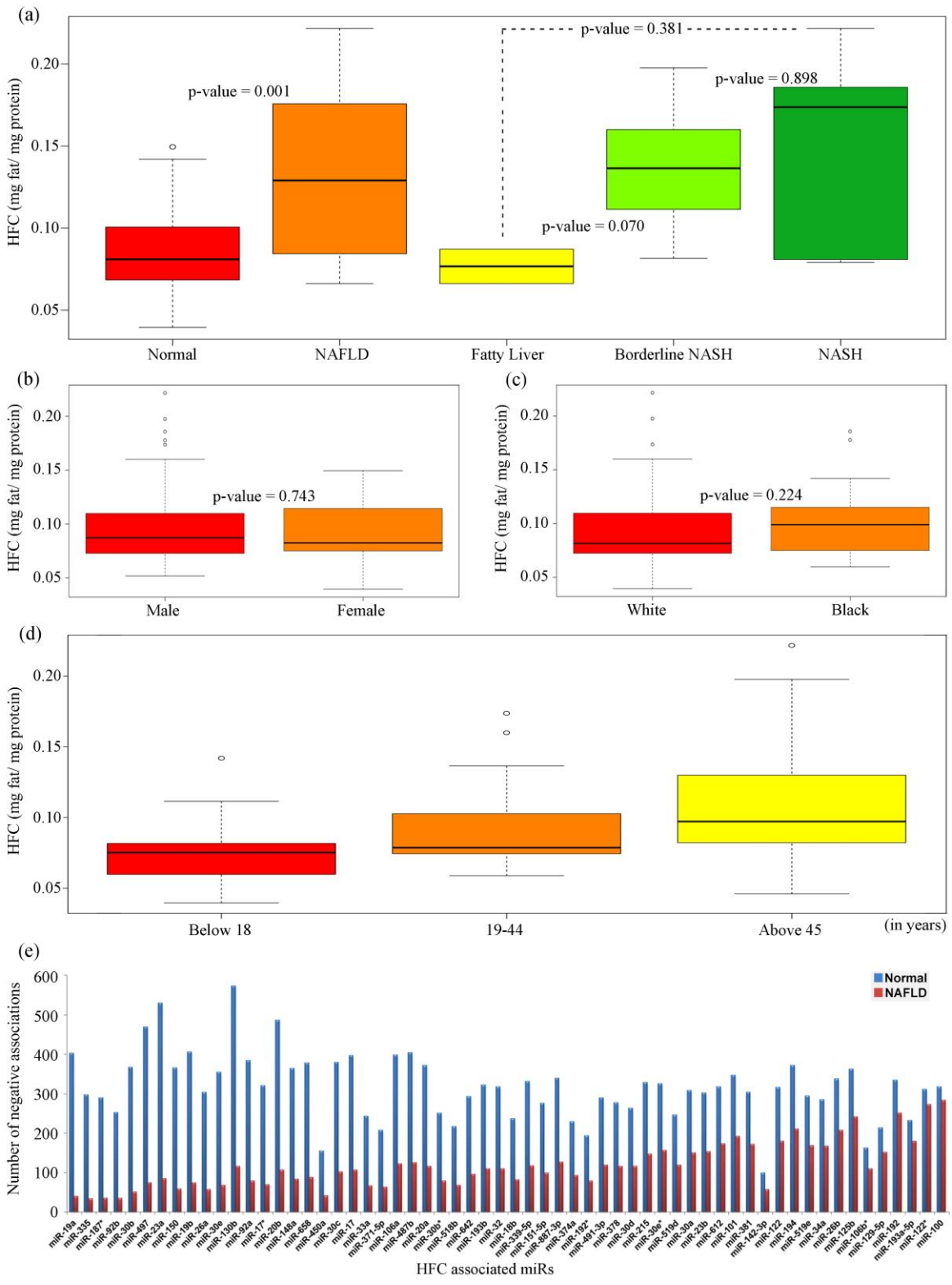


Figure 5. HFC associated miRs direct post transcriptional network changes in NAFLD: (a) Comparison of hepatic fat content between liver samples categorized to have normal liver histology and NAFLD; (b) Comparison of hepatic fat content according to different age

groups; (c) MiR-mRNA (gene) association network with higher significance ($p < 0.001$) was generated and number of negatively correlated mRNA with HFC-associated miRs in each case, normal (33 samples) and NAFLD (16 samples) is shown as a bar graph. The degree of negative associations for HFC-associated miRs was significantly decreased in NAFLD as compared to normal livers (median [IQR] 110 [32-188] vs. 318 [223-411], $p < 2.2e-16$ by Wilcox test).

2.3.7 Rewiring of post transcriptional network in NAFLD

MiR-mRNA association network, where the expression levels of these molecules were negatively correlated ($p < 0.001$), was generated in NAFLD (16 samples) and normal (33 samples) groups separately (Table S12). We observed that degree of negative associations for HFC-associated miRs was significantly decreased in NAFLD as compared to normal livers (median [IQR] 110 [32-188] vs. 318 [223-411], $p < 2.2e-16$ by Wilcox test) (Figure 5c). Similarly, as shown in Table S13, even after filtering these miRNA-mRNA associations for predicted miRNA targets by miRanda TargetScan algorithms, we found significantly fewer associations for NAFLD group compared to normal liver samples (median [IQR) 30 [4-56] vs. 94 [36-152], $p < 4.59e-10$). These data strongly suggest that miR-mRNA post transcriptional regulatory network is highly dysfunctional in NAFLD.

Sub-networks derived from Table S12, where the mRNAs were the predicted targets of miRNAs (Table S4), were constructed in normal and NAFLD samples. This allowed the calculation of degree ratio for 62 HFC associated miRs in normal and NAFLD samples allowing a comparison of the number of targets for each HFC associated miRNA in healthy and disease states (Table S13). We found that most HFC-associated miRs had differing number of associations in NAFLD compared to normal livers. In particular, a significant alteration in miR-mRNA interaction network was observed for miR-335, miR-19a, miR-30b, miR-26a and miR-20b (at least five-fold) with a decrease in the number of potential mRNA

targets in NAFLD samples. Functional analysis of the targets of these miRNAs indicated enrichment for immune response, phagocytosis, focal adhesion and signaling pathways. It is worth noting all of these miRNAs were found to be negatively correlated with HFC and hence could act as protective players in inflammation and immune response. Indeed, mRNAs like LAMA1, MYO1F, PTGS1, GAS6, CLSTN1, IGKC and CD86 which were identified to be positively correlated with HFC (Table 2) were predicted to be controlled by several of these rewiring miRNAs. While a subset of miRNAs including miR-129-5p, mir-125b, miR-106b* and miR-100 exhibited intermediate level (between 2 to 5 fold decrease in the number of interactions) of rewiring. We also observed a slight increase for miR-122* in terms of the number of interactions (Table S13). Additionally, we tested whether the number of miRNA-mRNAs interactions between normal and NAFLD groups differ when we filter the networks by including only the HFC mRNAs. Although the size of the networks were significantly smaller with degree information available in both normal and NAFLD groups for a subset of HFC miRNAs (marked in Table S13), we still found significantly fewer associations for NAFLD group compared to normal liver samples ($p < 2.47e-4$).

Chapter 3 Uncovering RNA binding proteins associated with age and gender during liver maturation

3.1 Introduction

Gene expression changes dynamically throughout the lifetime of an organism and the sub set of proteins expressed at each point in time allows cells to carry out important functions such as response to external stimuli, cell differentiation and development. These age related expression changes would influence the functioning of an organism. A study of post-mortem human brain tissue from 30 individuals aged 26 to 106 years showed that approximately 4% of the 11,000 genes analysed show a significant age related expression change[68]. Another independent study examined healthy renal tissue removed at nephrectomy from 74 patients ranging in age from 27 to 92 years to identify ~1000 genes to be differentially expressed with age[69]. In addition, a significant difference in the expression of several genes encoding for antioxidant and detoxifying enzymes was seen in aged livers of both rats and human[70]. More recently, a study on age-dependent gene expression changes in 5 different tissues showed skin to have the most age related gene expression changes[71]. Similar to the age-related expression changes, it is also seen that genes express differently in the same organs of male and female. Recently, it was reported that though male and females share high similarity at genome level, most of the dimorphic traits are constrained to happen based on sex-biased gene regulation[72]. Another study that studied the sex based differences in the transcriptome of the human blood identified a gender specific expression in 582 autosomal genes of which 57.2% were up regulated in females [73]. It is also suggested that the gender based differences in epigenetic mechanisms may have profound consequences on brain development [74].

Although, the expression of genes varies based on age and gender, molecular mechanisms causing these differences still remain unclear. Transcriptome changes could be largely attributed to difference in the levels of regulators participating at various stages of gene

expression. One such class of regulatory molecules are the RNA Binding Proteins (RBPs)-that bind RNA molecules to control different post transcriptional processes such as pre-mRNA splicing, mRNA cytoplasmic export, turnover, storage, and translation [13, 14]. Thus, the capacity of these proteins to influence gene expression at post-transcriptional level is extremely important especially during the developmental process to give rise to complex organs and tissues[15, 16]. For instance, PTBP1 (polypyrimidine tract binding protein), a ubiquitous protein known to be important in mammalian development at early stages of gastrulation[25, 26] and ELAVL1 (HuR) - a protein that acts as an mRNA stability factor, is also known for its role in placental branching, embryonic and neuronal development. [27, 28]. Likewise, CRD-BP (IGF2BP1), a member of the insulin-like growth factor 2 mRNA-binding protein family, is the first example of a putative mammalian mRNA-binding proteins that is abundant in fetal tissue but absent in the adult tissue [29].

As is evident from the above examples, RBPs play a substantial role in mediating developmental changes of a mammalian cell. Furthermore, a study on understanding the regulation of HNF4alpha in liver development revealed that the expression of HNF4aplha is widely regulated by the sequential promoter usage and alternative splicing in the 3' end to produce different isoforms important for the liver development[30]. Yet another study identified UPF2, one of the key players of the non-sense mediated mRNA decay (NMD) machinery, as a critical regulator of the liver development[31]. Thus, although specific RBPs have been studied for their role in mediating developmental processes of liver, no global association analysis has been performed in humans to uncover the repertoire of RBPs contributing to liver development. Hence, to complement this gap in our global understanding about the functions of RBPs as developmental regulators in liver, in this study a genome-wide association analysis of their expression patterns with age and gender was conducted. To achieve this, we performed an association analysis of RBPs expression levels in human liver tissues with respect to age and gender, by integrating a dataset of 1344 genes experimentally known to encode for RBPs. This

allowed us to catalogue for the first time age and gender-related associations for RBPs as well as to study the wiring patterns of RBPs and their protein complexes (protein-protein interaction networks) with liver development in the human genome.

3.2 Materials and Methods

3.2.1 Data set of RNA-binding proteins and expression profiles for human liver tissues

In the present study, we catalogued a set of 1344 genes encoding for RBPs in the human genome. This compendium comprised of proteins identified as RBPs in several recent experimental screens, including Castello et. al, [75], Baltz et. al. [76], Ray et. al, [77], human orthologs of RBPs identified in mouse embryonic stem cells by Kwon et. al, [78] and those reported in RBPDB [79].

In this study, we employed two different microarray-based liver expression cohorts for performing the association analysis. The first, a study by Innocenti et. al, which profiled 206 normal human livers (183 European Americans and 23 African Americans) of which 74 were females and 132 males with age ranging from 1 to 81 years to map expression quantitative trait loci through genome wide association mapping (GSE25935) [32]. The second study by Schroder et. al, profiled 149 liver samples of Caucasian origin to identify expression quantitative loci (eQTL) in human liver of 71 females and 78 males ranging from 7 to 85 years of age (GSE32504) [80]. Data processing and normalization were carried out using the packages available in the R statistical framework. Raw expression data from GSE25935, generated using Agilent array was available in single colour format. This raw data was processed using limma package. For normalization we used a three step approach 1) for background correction (method = “normexp” and offset = 50) 2) for normalization between arrays (method = “quantile”) and finally 3) log transformation of the normalized data was performed. Secondly, in case of

GSE32504, raw data was generated using Illumina Bead Studio version 3 and thus beadarray [81] and lumi [82] packages from Bioconductor biocLite were used for processing of the raw data. Normalization of the raw data was carried out in a single step using quantile normalization approach. Figure 6A shows a generic flowchart summarizing the various steps involved in the data processing and analysis of the two microarray datasets.

3.2.2 Associating the expression of RBPs with age and gender

To test the association of expression with age and gender, we modelled the expression of each RBP as a dependent variable with age and gender as independent variables. Using the analysis of covariance -a statistical test used to explain the variance between independent and dependent variables, available from the R statistical framework we computed the significance of association. For each of the two microarray datasets, normalized expression of RBPs was simultaneously associated with the metadata - age and gender (Figure 6). Associations predicted at $p<0.001$ (FDR=5%, Benjamini-Hochberg [83] procedure) were considered significant. Thus, RBPs that were identified to significantly associate with age from either of the datasets were merged to form a non-redundant set. Similarly, a non-redundant set of RBPs associated with gender was obtained from the two datasets.

Further, correlation analysis was performed between RBPs expression and age for the two datasets separately. Spearman correlation was calculated and p-value for each correlation association (expression vs age) was calculated. Associations identified at $p <0.001$ (FDR = 5%, Benjamini-Hochberg procedure) were considered significant. A similar approach was adopted to identify the associations of 1348 transcription factors extracted from DBD: Transcription factor prediction database[84] and ~18,000 Non-RBPs in the human genome.

3.2.3 Network dynamics of RBPs associated with age and gender

We categorized the 1344 RBPs into two groups – those that are significantly associated with age ($p<0.001$ & FDR = 5%) in either of the two liver cohorts and the remaining genes that are not associated with age. A protein-protein interaction (PPI) network was constructed for each of the two groups of RBPs, using the publicly available experimentally verified interaction data from BioGRID database[85]. Age-associated RBP interaction network comprised of 47 nodes and 86 edges while the non-age associated RBP interaction network comprised of 1112 nodes and 9325 edges. Since the non-age associated network was significantly larger in terms of the number of nodes and edges, we employed a custom randomization model to generate random networks of the same size as the observed age-associated RBP network. In other words, the interaction network of proteins that are not associated with age was used to create 100 random networks each comprising of the same number of edges to serve as a control set. Briefly, let there be x number of edges in the network where RBPs are associated with age and y edges in the network not associated with age (assuming $y > x$). We obtained 100 random networks to represent control networks not associated with age and with each iteration we randomly obtained x edges and thus constructed 100 random networks with RBPs not associated with age interacting among themselves. The same protocol was applied to generate interaction networks for genes/proteins associated and non-associated with gender. Non-gender RBP interactions were then used to construct 100 random networks of RBPs not associated with gender.

To study the properties of the PPI networks associated with age and gender, we used igraph, a publicly available R package for analyzing graphs [see http://cneurocvs.rmki.kfki.hu/igraph/ and http://www.r-project.org]. In particular, since the network of PPIs analyzed in this study is undirected, we used the corresponding versions of the functions: degree, clustering coefficient

(transitivity), betweenness and closeness for calculating the degree, clustering coefficient, betweenness and closeness centralities of a node. Betweenness centrality, which is the number of shortest paths going through a node was calculated using the brandes algorithm [86] implemented in R. Similarly, closeness, measured as the inverse of the average length of the shortest paths to all the other nodes in the graph, was obtained using the implementation in R. Since the centrality measures, betweenness and closeness use the shortest path lengths between all pairs of nodes in a graph, for cases where no path exists between a particular pair of nodes, shortest path length was taken as one less than the maximum number of nodes in the graph. Note that this is also the default assumption for calculating centrality measures in igraph. All the network centrality measures were compared between age-associated and non-age-associated random RBP networks to study for differences in the distributions by performing non-parametric tests. Similarly, gender-associated RBPs were compared for their corresponding network properties in non-associated random networks.

3.2.4 Validation of the RBP associations in liver tissues of mouse and rat

Mouse:

Liver tissues were collected from mice at the indicated time points for mRNA expression analysis. For further details regarding mice strain and age refer to Supplementary Table 4. Total RNA was isolated from mouse liver tissue using TRIzol. Total RNA (~5 µg) was reverse transcribed using Random Hexamer Primer (Thermo) and the Thermo Scientific Maxima Reverse Transcriptase kit. Quantitative real-time PCR were performed in duplicates using 130 ng of cDNA per reaction on an Eco Real-Time PCR system using Quanta PerfeCTa SYBR Green FastMix. An initial activation step for 10 min at 95 °C was followed by 40 cycles of 95 °C for 10 s and 60 °C for 30 s. Primer sequences used to measure relative steady-state expression of mRNA are provided in Supplementary Table 4. Fold change of the mRNA was

calculated as previously described [87]. Data is presented as mean ± SEM.

We followed the NIH Guide for the Use and Care of Laboratory Animals. The Institutional Animal Care and Use Committees of Universities of Illinois approved all experiments.

Rat:

The orthologs of 88 human RBPs associated with age in rat were extracted from the ENSEMBL compara, resulting in 65 corresponding genes. Expression profiles of these orthologs in 8 rat liver tissues RNA-sequenced at four developmental stages – 2 weeks, 6 weeks, 21 weeks and 104 weeks were extracted from the Rat Body Map Database (http://pgx.fudan.edu.cn/ratbodymap/). This data was then used to the compare the expression patterns of RBPs between human and rat livers to identify the genes which followed the same trend as in humans. Heat maps were generated for RBPs which increased or decreased in both humans and rats (Supplementary Figure 2).

3.3 Results

3.3.1 RBPs exhibit significant alterations in their expression with aging human liver

Association and correlation analysis revealed the expression of 88 RBPs to be significantly varying with age in human liver tissues ($p<0.001$, FDR=5%) (Figure 6, Materials and Methods, Supplementary Tables 1 and 2). These included 54 RBPs whose expression was observed to be increasing with aging liver (Figure 7A). Among these, LRPPRC – a leucine rich RNA Binding protein associated with post splicing nuclear RNP complexes and predominantly located within mitochondria [88] was found to have the highest correlation. Following this is SUMO1- a qualified RBP identified from the interactome capture of the human HeLa cell lines [89]. In addition to these, we also capture one of the genes in the p53 developmental pathway –

ZMAT3, a zinc finger protein involved in the mRNA stability whose overexpression suppresses tumour growth [90, 91]. On the other hand, we found 34 genes to be significantly associated but negatively correlated with age (Figure 7B). For instance, we find PCDH20 – a novel member of the protocadherin gene family to be significantly negatively correlated with age. Moreover, an independent study has recently identified that the mRNA levels of this proteins to be down-regulated in the hepatocellular carcinomas when compared to the disease free livers and thus functioning as a tumour suppressor [92]. Another notable example is PPIA – member of the peptidyl-prolyl cis-trans isomerase family known to accelerate the folding of proteins which in our analysis, was seen to negatively correlate with age. Additionally, a recent study proposed that stable expression of PPIA (or CypA) in liver cells confers resistance to anticancer drugs like doxorubicin and vincistrine [93] . In summary, we found 54 RBPs (61.4 %) to be positively correlated with age suggesting that the majority of the age-associated post-transcriptional control in liver is due to increased expression of RBPs indicating a potential for increased number of targets with age[14].

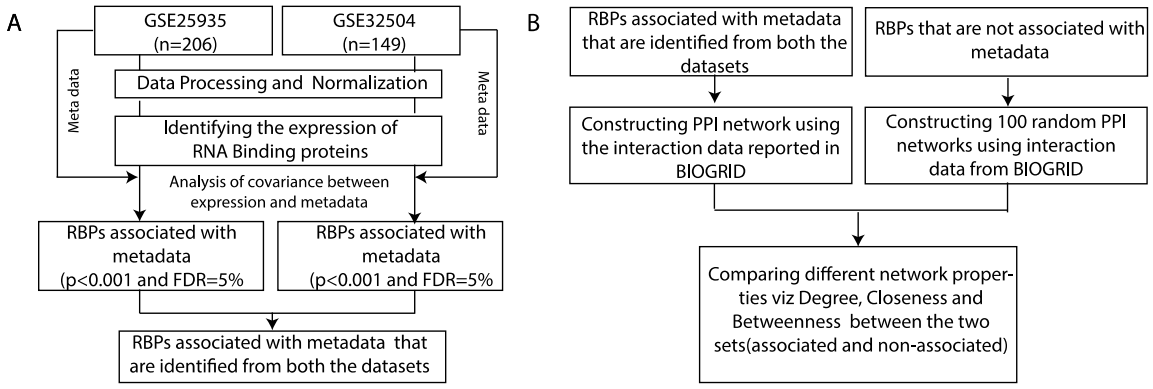


Figure 1 : The flowcharts in Figure A and B are idependently carried out for metadata associated with age and gender

Figure 6: Flowcharts summarizing the major steps involved in association and network analysis. A) Flowchart showing the various steps adopted for pre-processing and normalization of raw microarray data from both the liver cohorts followed by analysis of covariance between expression and metadata (age and gender) using functions in R statistical framework to identify the most significant associations. B) Flowchart showing the steps for construction of protein-protein interaction (PPI) networks between RBPs associated with metadata (either age or gender) and a control set of 100 randomized PPI networks from RBPs not associated with metadata followed by the comparison of network properties between them.

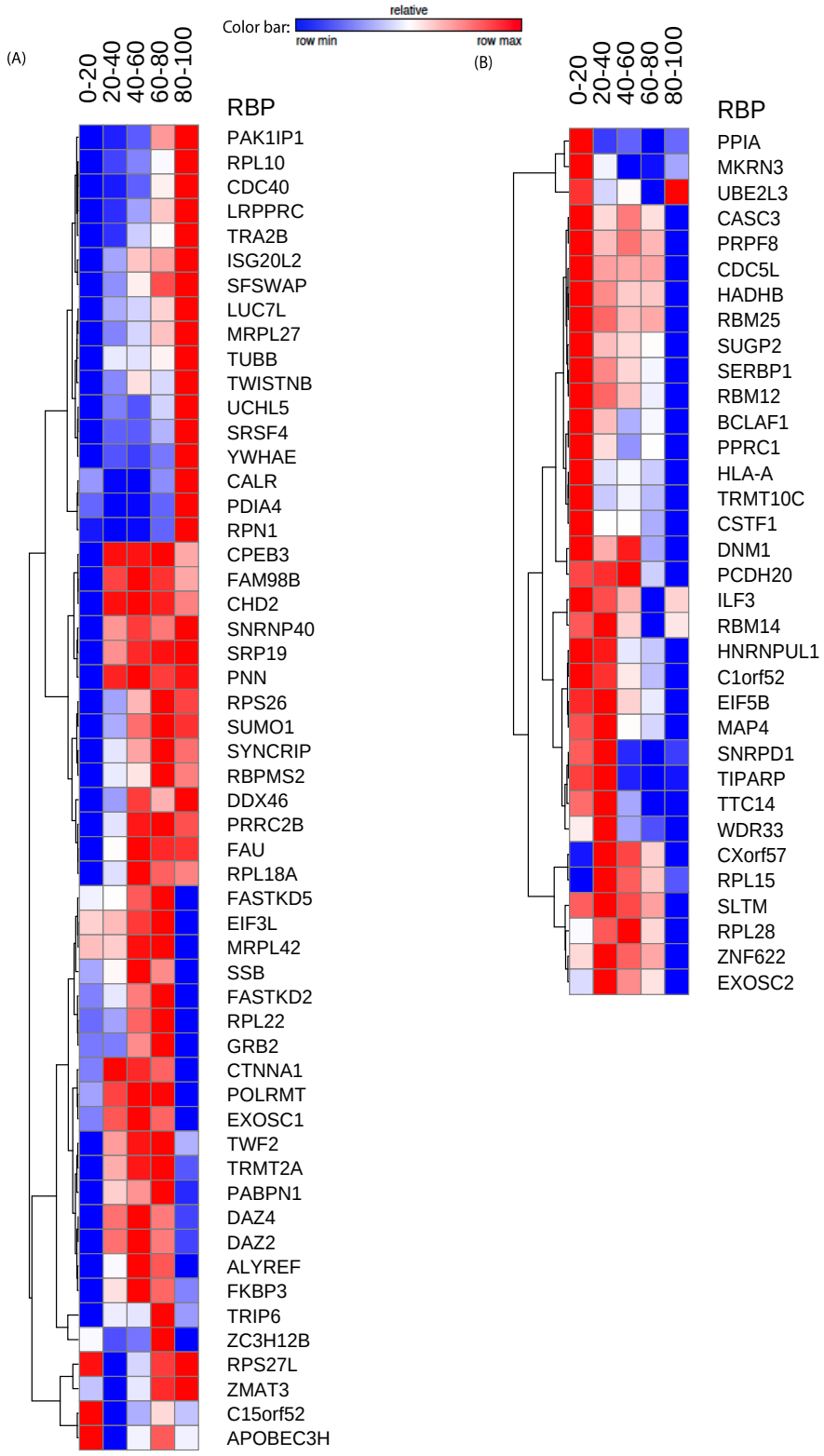


Figure 7: Expression patterns of RBPs associated with age. A) Heat map showing RBPs that are positively correlated with age (p<0.001, FDR<=5%) from both the cohorts in the study B) Heat map showing the list of RBPs that are negatively correlated with age (p<0.001, FDR

<=5%). Age ranges are shown from 0 to 100 years and expression levels are binned into 20 year intervals.

To know if the repertoire of proteins important for liver development is enriched for RBPs, we performed a similar association analysis for TFs and Non-RBPs (See Materials and Methods). This indicated that a higher proportion of RBPs are seen to be significantly associated (~6.5%) when compared to the TFs (~5.6%) and Non-RBPs (~5.4%) ($p = 0.09$, Fisher's exact test). This strengthens the notion that RBPs are at least as important as other regulatory molecules implicated during liver development.

3.3.2 Several RBPs showed similar changes in expression within aging mouse and rat liver tissue

To validate whether our observations in the human liver samples are reproducible, we measured the relative fold change of the RBPs which showed strong positive correlation with age, using real-time-PCR in mouse liver for four different RBPs (see Materials and Methods) (Figure 8A). In general, RBPs which showed increasing expression with age in human liver samples also showed a very similar trend in mouse liver. As shown in Figure 8A, this trend is observed for Sumo1, Srsf4 and Mrpl42 which show a late and immediate increase in expression (> two-fold) at 10 months (Supplementary Table 3).

Similar analysis was performed on four RBPs, which exhibited a strong negative correlation between expression and age, in the mouse liver (Figure 8B). Of the four RBPs validated, Cstf1 and Map4 exhibited strong decrease in expression levels with aging mice. Ppia and Mkrn3 followed decreasing trends in expression levels albeit weakly (Supplementary Table 3).

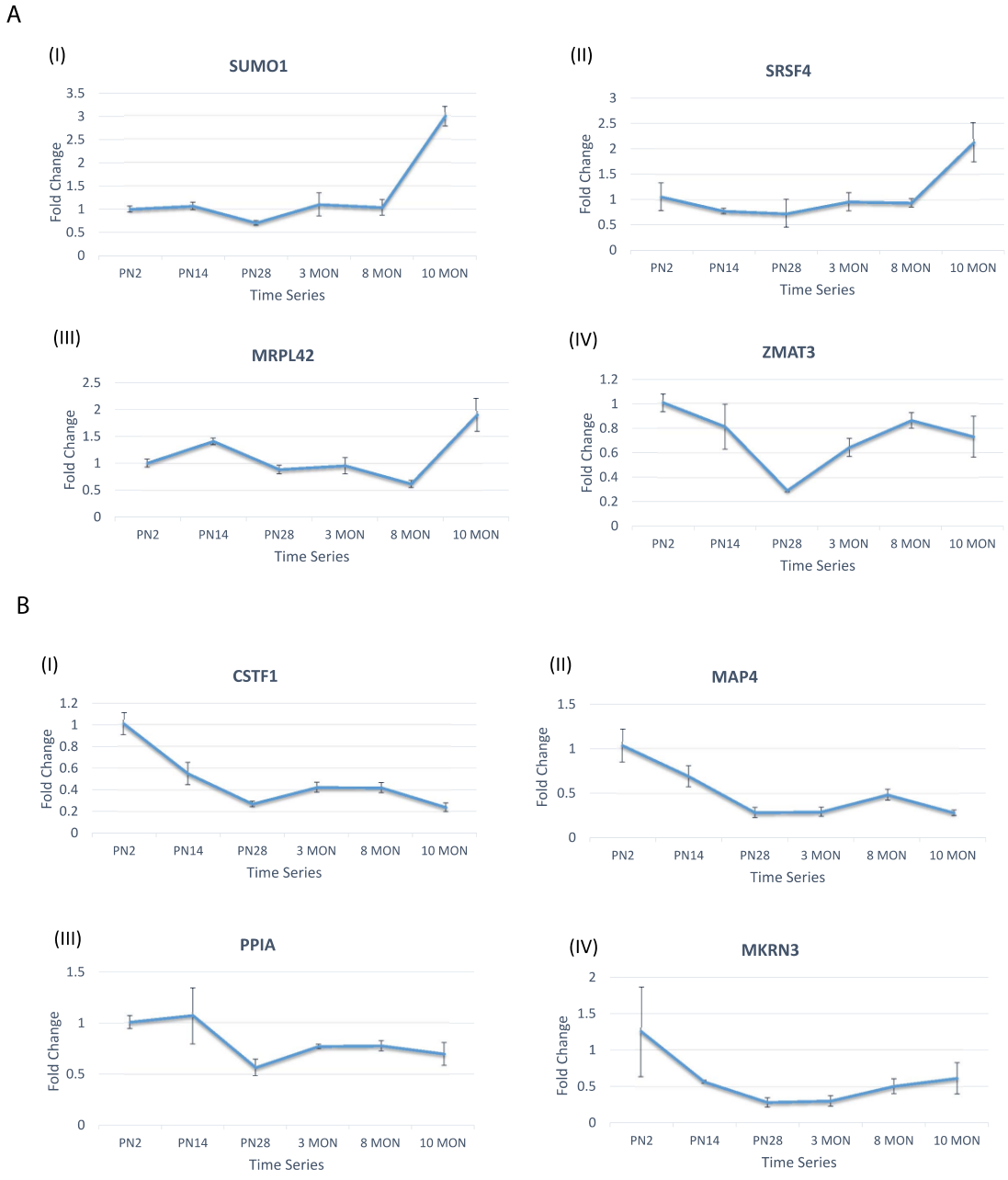


Figure 8: Expression levels measured using RT-PCR for eight RBPs in mouse for the time series ranging from postnatal 3 days to 10 months. X-axis shows the age of the mouse and y-axis represents the average fold change in expression with standard error of mean shown for each time point. A) shows four RBPs tested in mouse model to see if the expression is positively correlated with age B) shows four RBPs tested in the mouse model which were found to be negatively correlated with age in the human samples. All the tested RBPs show strong correlation with age in the respective directions in human through our computational pipeline.

To further validate our findings, we extracted the expression profiles of these RBPs in rat livers (See Materials and Methods). We observed a strong reproducibility in the expression patterns of 33 human RBPs in the aging rat tissues (See Supplementary Figure 2). For example, we observe SUMO1, SYNCRIP and PDIA4 to be positively correlated while HADHB and PPRC were found to be negatively correlated with age in the rat livers. More generally, to study the enrichment of RBPs changing with age in human liver to be reproducible in the rat liver, we compared the proportions of the respective trends of the RBPs. RBPs increasing with age in human were found to be significantly enriched to be also detected as positively-correlated with age in rat ($p=0$, Hypergeometric probability). We also found the RBPs decreasing with age in humans to be markedly over-represented to be negatively correlated with age in rat samples ($p < 3.3e\text{-}13$, Hypergeometric probability) suggesting that the observations are reproducible in mouse and rat genomes despite the variations in the samples and the absence of orthologs for several RBPs due to evolutionary divergence between the species.

3.3.3 A small fraction of the RBPs are sexually dimorphic in humans

Association analysis also enabled us to identify sex-specific RBP expression and to uncover variations between male and female samples (see Materials and Methods). Briefly, this analysis revealed 45 RBPs whose expression was found to have a significant difference between genders ($p < 0.001$, FDR = 5%) (Supplementary Tables 1 and 2). We also found that a small set of these gender-associated RBPs (~40%) are up-regulated in males when compared to females. Several RBPs including WDR6, RBM4, GSPT1 and EIF1AX were found to be significantly differentially expressed between male and female samples (Figure 9). Of these, WDR6, member of the WD repeat protein family, implicated in the cell growth arrest was observed to be expressed at relatively higher levels in males. WD repeats are minimally conserved regions of around 40 amino acids generally bracketed by gly-his and trp-asp, which

may further facilitate multi-protein complex formation. It is known to enhance STK11/LKB1-induced cell growth suppression activity, a negative regulator of amino acid starvation–induced autophagy[94, 95]. Similarly, EIF1AX-eukaryotic translation initiation factor was observed to be expressed in higher levels in males than in females. In contrast, RBPs like GSTP1, LGAS1 and RBM4 were found to exhibit higher expression in females compared to males (Figure 4). Overall, our results suggest that a relatively smaller fraction of the RBPs are associated with gender compared to age in human liver and majority of these are up-regulated in females compared to males.

3.3.4 Age-associated RBPs form a dense modular network of protein interactions

RBPs attain their precise spatio-temporal control of gene expression most typically by forming protein complexes in the cell [96]. Therefore, to understand how the 88 RBPs, whose expression was found to be strongly associated with age in human samples, are interacting, we constructed a network of experimentally known protein-protein interactions (PPIs) between them. This was achieved using documented interactions between RBPs from the BioGRID database [85] (Materials and Methods). This resulted in a network of 86 protein-protein interactions between the age-associated RBPs (Figure 10). A closer inspection of the distribution of clustering coefficients of the nodes, which is a proxy for the modularity of the network, indicated that the network is significantly modular compared to random networks of the same size (Materials and Methods) (Supplementary Figure 1). Indeed, we found that the age-associated RBP network exhibited twice the clustering coefficient than random networks (0.27 vs 0.14 median clustering coefficient, p=2.2e-146, Wilcox test). Additional analysis to cluster the network into likely protein complexes indicated the presence of two high confidence protein complexes using ClusterONE [97] (Supplementary Table 5). One of these complexes comprised of ILF3, RPS26, FAU, SYNCRIP- a member of the cellular heterogeneous nuclear ribonucleoprotein (hnRNP) family and several members of the 60S ribosomal subunit.

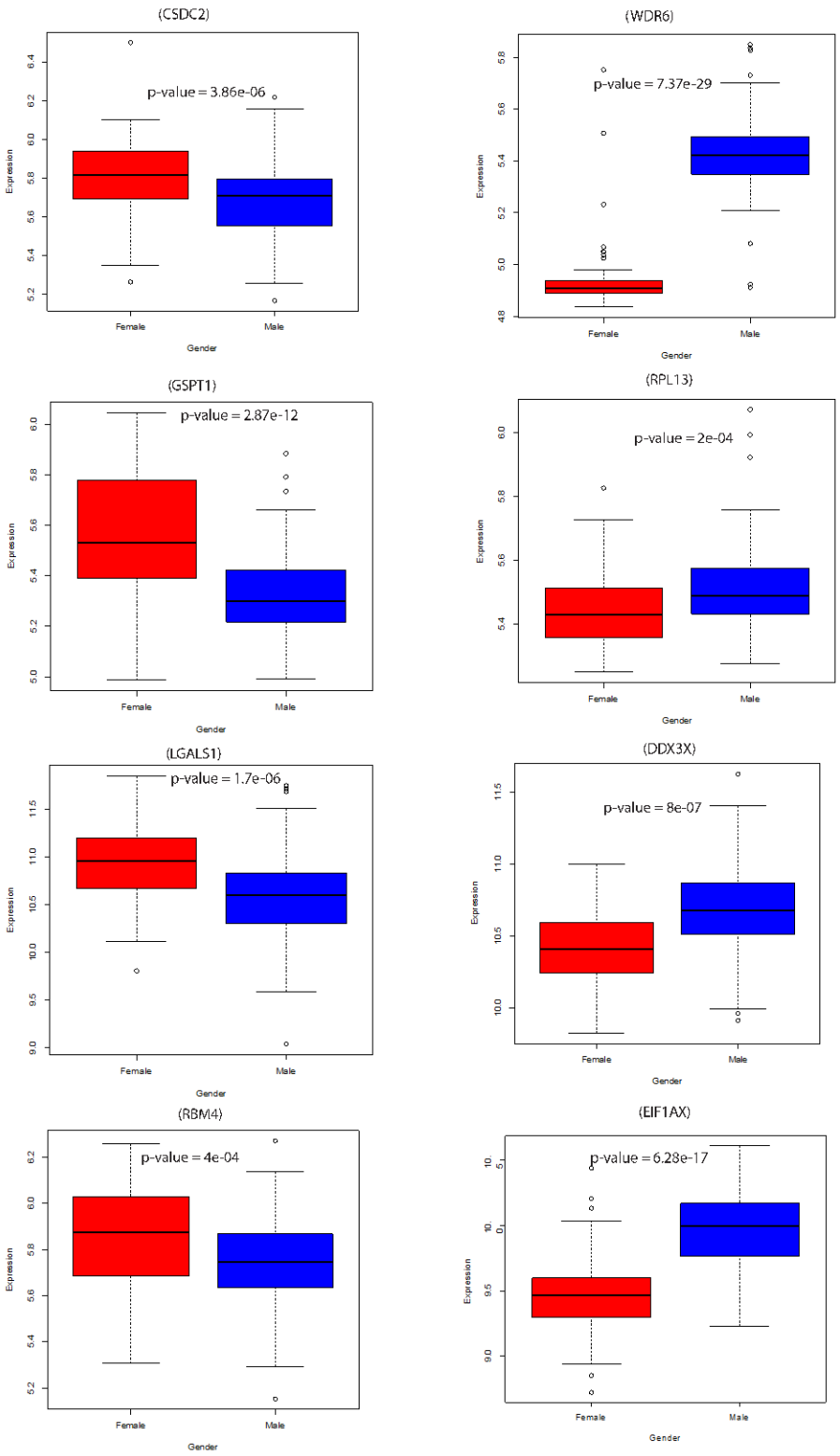


Figure 9: Boxplots showing expression levels of selected set of RBPs which were found to be gender specific and thus are differentially expressed in male vs female samples in humans. All the comparisons are significant at p-value <= 0.05 (Wilcox test).

This dense network also comprised of SUMO1 and ILF3, forming hubs with highest number of interactions (Figure 10). Figure 10 also shows the edge betweenness (normalized) defined as the number of shortest paths going through an edge of interest and is analogous to the node betweenness, a centrality measure for nodes in a network (see Materials and Methods). We found a high edge betweenness score for the edge connecting SUMO1-protein whose expression increases with age and ILF3-protein whose expression decreases with age suggesting that there might be an inverse relationship in their stoichiometry with age to separate the age-related RNP complexes/network into different partitions.

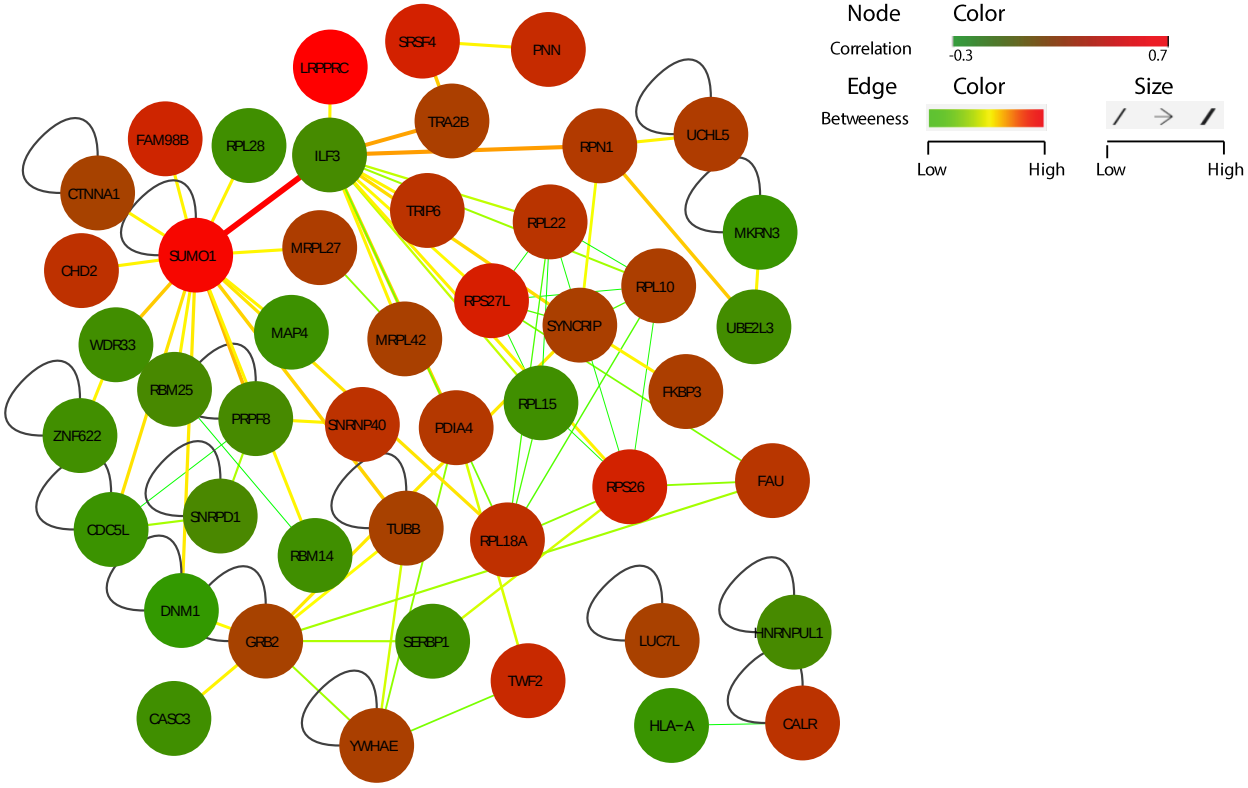


Figure 10: Protein interaction network of RBPs associated with age. Network showing the experimentally verified protein-protein interactions from the BioGRID database [85] between age-associated RBPs. Node colours vary between green and red signifying the extent of correlation of expression with age, with positively correlated RBPs coloured in red. Similarly, edge thickness and colour are highlighted with edge betweenness scores. Edges with higher edge betweenness centrality scores are shown in a thick red colour. Network also highlights how RBPs positively and negatively correlated with age, interact among themselves to form dense modular RNP complexes.

3.3.5 Age and gender-specific associations in liver show significantly higher network centrality measures

The importance of a protein can be assessed by measuring the centrality of a node in the PPI network. Centrality of a node can be measured by a number of different metrics including the most commonly used measures - degree, closeness and betweenness [98]. Therefore, to assess the significance of the centrality of the nodes in the age-associated RBP network, we compared the various centrality measures with that observed in a control set of random networks that contains a network of interactions among proteins that are not associated with age (see Materials and Methods). Normalized degree of the nodes in the network of proteins associated with age was found to be significantly higher than that seen in the control set ($p=2.46e-52$, Wilcox test) suggesting a higher connectivity among the age-associated RBPs (Figure 11A). Similarly, the distribution of closeness values was observed to be significantly different ($p=1.97e-31$, Wilcox test) and higher than the control set suggesting the existence of a denser network of interactions among proteins associated with age. The distribution of betweenness centrality scores were also found to be significantly different ($p=1.33e-13$, Wilcox test) and higher than the control set suggesting that these proteins are likely to play essential roles. Our results also suggest that these genes encoding for RBPs and strongly associated with age might form a dense and intertwined network of protein complexes contributing to the regulation of several age-related post-transcriptional processes in the liver tissue such as development and regeneration. Similar results were observed when we compared network of RBPs which expressed differentially in male vs female samples to the set of random networks of RBPs which are not gender specific (Figure 11B). In particular, we found that each of the centrality measures- degree, betweenness and closeness were significantly higher for gender-associated RBPs compared to that seen in the random networks.

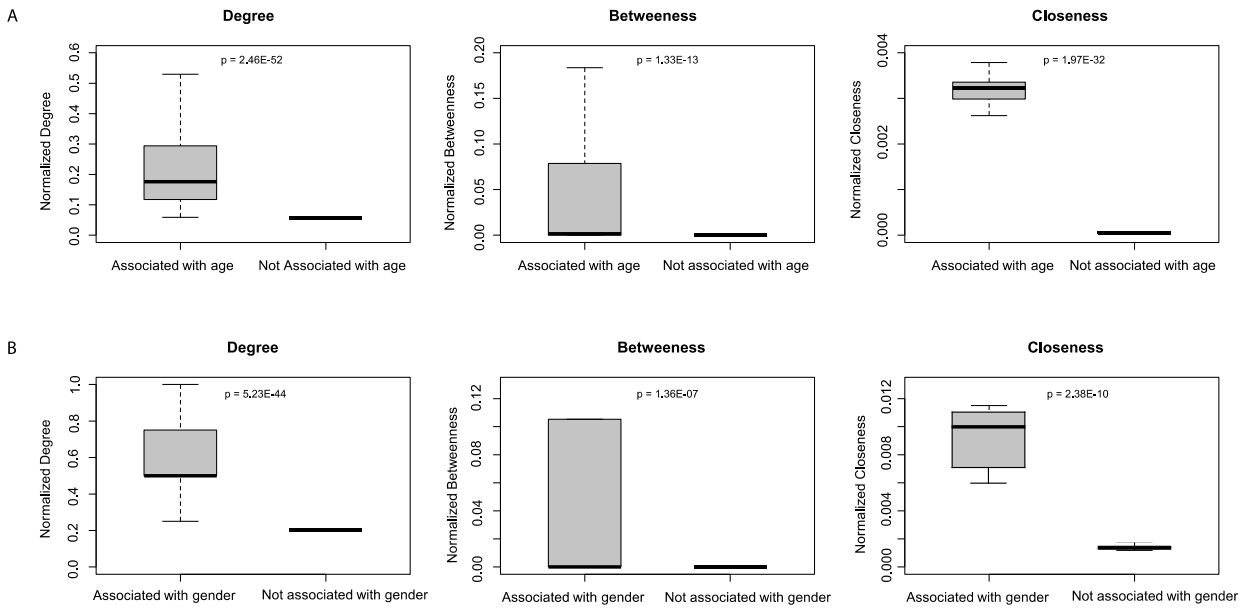


Figure 11: Boxplots for comparison of various network centrality measures A) between the networks of RBPs associated with age and not associated with age. The network of RBPs not associated with age comprises of an ensemble of 100 random networks constructed based on the randomization approach discussed in Materials and Methods section. Boxplots correspond to the comparison of normalized degree, betweenness and closeness values in the respective networks. All normalizations are based on the maximum value observed for the particular centrality measure in the corresponding network/s. B) between the networks of RBPs associated with gender and not associated with gender. The network of RBPs not associated with gender comprises of an ensemble of 100 random networks constructed based on the randomization approach discussed in Materials and Methods section. Boxplots correspond to the comparison of normalized degree, betweenness and closeness values in the respective networks. All normalizations are based on the maximum value observed for the particular centrality measure in the corresponding network/s.

Chapter 4 Discussion

Over the past decade, although significant progress has been made in understanding the pathogenesis of NAFLD, the detailed molecular pathways underlying the disease process are not well understood. Our study for the first time carried out a genome-wide integrated miR-mRNA network analysis focusing on hepatic fat accumulation as an objective measure of NAFLD. The study revealed major dysregulated miRs controlling mRNAs significantly associated with hepatic fat accumulation. We also found that the miR-mRNA regulatory network was significantly rewired in NAFLD liver samples compared to normal livers. These findings provided important targets that foster strong rationale for further mechanistic validations and to investigate potential treatment targets.

Our strategy aiming at the hepatic fat content as a quantitative objective measure of NAFLD allows us to identify genetic factors associated with NAFLD in a robust manner. Indeed, the list of functionally important miRs, as shown in **Table 1**, demonstrates the high confidence of our analysis. Many miRs identified in our analysis were linked to NAFLD and NASH in previous studies, e.g. miR-122, the most studied miR in liver, is considered as biomarker involved in various liver diseases such as drug- and alcohol-induced liver disease, non-alcoholic fatty liver disease, fibrosis, viral infections (hepatitis), cirrhosis and hepatocellular carcinoma [99]. Similarly, miR-34a, which was found significantly correlated with HFC ($p \leq 0.003$ with $q \leq 0.01$) in our study, was previously documented as being dysregulated in multiple liver disorders (Table 1). Apart from these two, some other miRs like miR-33b, miR-378 and miR-125b were also extensively studied for their role in lipid metabolism [9, 100] further supporting the significance of our observations. Confirmation of these earlier observations in our study underscores the critical role that these “very important microRNAs” (VIMs) may play in the pathogenesis of NAFLD. In addition, several miRs identified in this

study to be associated with hepatic fat accumulation (e.g., miR-19b and miR-518b), which were also predicted to regulate multiple mRNAs associated with HFC, have not been reported in the context of liver diseases previously. How these miR-mRNA regulatory connections confer risk for NAFLD warrants further investigation.

Our analysis also confidently identified a list of VIM-targeted genes that are associated with HFC and NAFLD. For instance, *STMN2* was found in our analysis as a significant ($p \leq 3.36 \times 10^{-5}$ with $q < 0.001$) and most positively associated gene with HFC. Our GO term analysis suggested its involvement in cell communication and developmental processes. This gene encodes for a phosphoprotein which helps in stabilizing cytoskeleton microtubule organization [101]. Our analysis showed a list of high confident post transcriptional regulators (i.e., miR-374a, miR-30b*, miR-887-5p, miR-129-5p, miR-142-3p, miR-885-5p, miR-204 and miR-122) that could modulate this gene in NAFLD. This gene was reported to be significantly up-regulated in individuals with liver fibrosis [102]. Thus, it could be conjectured that dysregulation of this gene may confer susceptibility to the distortion of cytoskeleton organization in hepatocytes with fat accumulation and there by promoting ballooning and fibrosis. Similarly, *LAMA1*, encoding Laminin alpha 1, was the second most positively correlated gene with HFC, and was previously found to be involved in multiple biological processes including metabolism, signaling, communication, immunity, cell adhesion etc. from their GO term annotations. Laminin can interact with extracellular matrix components and mediate the attachment, migration and organization of cells within tissue during developmental processes. Its expression was reported to be 2.1 fold increased in NASH patients compared to normal steatosis patients [103], further supporting our observations. We found, miR-30a, b, c, d, e, and miR-26a and b, as major post transcriptional regulators for this gene.

Lipid metabolism and immune response were identified as most affected processes recurring in functional theme with high significance and gene count and this confirms a key role for dysregulated lipid metabolism and inflammation-related pathways in the pathogenesis of NAFLD. HFC-associated 23 genes **(Figure 4)** belong to specific immunity and inflammation pathways like natural killer cell mediated cytotoxicity, Fc gamma R-mediated phagocytosis, steroid biosynthesis and IL-2 Receptor Beta Chain in T cell Activation provide the mechanistic information prone to molecular alteration as an effect of hepatic fat accumulation. For instance, one previous study suggested that dysregulation of Gas6 is implicated in the progression of steatosis to steatohepatitis and fibrosis in mice liver [104]. Our results identified two miRs significantly associated with HFC, miR-491-3p and miR-151-5p, as major regulators of the Gas6 gene.

Most interestingly, we found that the miR-mRNA network was significantly rewired between normal and NAFLD liver samples. Our results suggest that all the HFC-associated miRs identified in this study were likely to contribute to the rewiring of the post-transcriptional network in NAFLD to differing extents. In particular, most of the miRNA-mediated-dysfunction in NAFLD might be because of the decrease in the functionality of the miRNAs which would otherwise target higher number of transcripts in a healthy state. In other words, it can be hypothesized that the effective changes in miR-mRNA interaction in NAFLD for miRs like miR-19a, miR-26a is because, either they skip most of their targets due to influence of some unplugged thorn in NAFLD or they degraded an important post-transcriptional target as an immediate response during fat accumulation which is in turn responsible for triggering the decrease in the number of targets.

Some aspects of our study deserve further comments. Our study is based on human liver samples obtained from transplant donors, which were subsequently characterized in terms of

their HFC and NAFLD. It would be ideal to conduct this study on samples obtained from living humans but such a study design is very difficult because it is challenging (a) to obtain liver samples from individuals with normal livers and (b) to obtain liver tissue of sufficient quantity to measure hepatic fat, protein, mRNA and miRs. One could argue that individuals undergoing bariatric surgery may serve as an avenue for obtaining liver samples of sufficient quantity, but morbidly obese undergoing bariatric surgery represent an extreme phenotype and thus may not reflect NAFLD as seen in the general population. Another potential limitation of our study from a genomic view point is that our microarray platform for miRNAs does not include the recently identified novel miRNAs and hence the repertoire of miRNAs being reported here might be incomplete as there might be additional miRNAs which might be correlated with HFC belonging to the new families of miRNAs. A possible solution to address this problem is to perform RNA-seq of the samples in the future, which can provide a comprehensive overview of the complete set of miRNAs in the liver samples which are associated with HFC. Yet another potential limitation of our current dataset is that it does not shed light into the pathogenesis of nonalcoholic steatohepatitis (NASH) and advanced fibrosis. Our cohort lacked sufficient number of samples with NASH and advanced fibrosis. Future studies should focus of cohorts which are enriched with these advanced forms of NAFLD.

In summary, our study for the first time provided detailed insights into the HFC-associated miR-mRNA regulatory network and these observations provide clues to better understand the pathogenesis and treatment targets of NAFLD. Our findings further highlighted the miRs and mRNAs previously identified, and more importantly generated a number of new hypotheses to be investigated in the future. Similar systems biology studies should be undertaken to better understand the pathogenesis of NASH and advanced fibrosis in humans.

Many studies to date have focused on genes that change with age and gender but the regulatory molecules mediating such mechanisms remain unclear. Previous studies have shown that eQTLs (expression Quantitative Trait Loci) interact with age and gender and play a major role in disease susceptibility[71, 105]. However, the contributing regulatory factors remain unclear. For instance, Masuda et. al, show that mRNA turnover and translation regulatory (TTR) RBPs show similar pattern in many tissues like gastrointestinal, urinary and immune systems in an age dependent manner [106]. This article emphasises that change in expression of a specific group of RBPs - important players in mRNA turnover and/or translation, as a function of age in human liver samples is an important but poorly studied level of association. However, post-transcriptional control is mediated by several hundreds of RBPs in the human genome and our knowledge about their role in controlling age-associated processes is rather limited. Therefore, in this study, to address this problem, we have compiled two different liver expression cohorts and performed a global association analysis between hundreds of experimentally characterized RBPs in the human genome for the first time to uncover the compendium of RBPs varying with age and gender. We observe several RBPs to be changing in expression with age and gender suggesting an important role for post transcriptional regulation during liver development. Majority (~60%) of the age-associated RBPs were found to be increasing in their expression levels with age while a small set of gender-associated RBPs (~40%) were found to be up-regulated in males. Several of these predicted associations were confirmed in the rat tissues and experimentally validated in a mouse model. This resulted in a significant number of human RBPs showing similar expression patterns in aging liver of rat and mouse, thus corroborating our findings. Further analysis of the protein interaction network of RBPs associated with age and gender based on the centrality measures like degree, betweenness and closeness revealed that several of these RBPs might be prominent players in liver development and impart gender specific alterations in gene expression via the formation of protein complexes. Furthermore, SUMO1 – gene which was positively correlated with age was observed to be sharing a highest

edge betweenness with ILF3 that was negatively correlated with age signifying a possible existence of a co-regulatory mechanism linking sumoylation and translational regulation – processes controlled by these respective proteins during liver development. This also suggests that the interaction between SUMO1 and ILF3 might facilitate a link between post transcriptional control, protein stability, nuclear cytosolic transport of proteins and translational control. It is worth mentioning that in this study we are only looking at a static network of physical interactions between RBPs associated with age (Figure 10). However, it is easy to note that this dynamic network might evolve/vary with age suggesting that varying sub-networks might be active depending on the developmental time frame and interplay with other cellular processes. So it is possible to speculate that there may be a time point in dynamics when the stoichiometry of the physical interactions between RBPs increasing and decreasing with age may disrupt leading to causative effects or disease phenotypes. With increasing high-resolution data from next generation sequencing and proteomics pipelines from hundreds of individuals it should be possible in the near future to understand such complex and rich set of associations between genes and proteins across cell types. Hence, this study will not only help us gain insight into the role of post-transcriptional regulatory molecules in aging and gender specific expression of genes but also provide a foundation for identifying the causative players contributing to the splicing eQTLs.

1. Kusenda, B., et al., *MicroRNA biogenesis, functionality and cancer relevance.* Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub, 2006. **150**(2): p. 205-15.
2. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.
3. Visone, R., F. Petrocca, and C.M. Croce, *Micro-RNAs in gastrointestinal and liver disease.* Gastroenterology, 2008. **135**(6): p. 1866-9.
4. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.
5. Janga, S.C. and S. Vallabhaneni, *MicroRNAs as post-transcriptional machines and their interplay with cellular networks.* Adv Exp Med Biol, 2011. **722**: p. 59-74.
6. Vickers, K.C., et al., *MicroRNA-27b is a regulatory hub in lipid metabolism and is altered in dyslipidemia.* Hepatology, 2013. **57**(2): p. 533-42.
7. Lynn, F.C., et al., *MicroRNA expression is required for pancreatic islet cell genesis in the mouse.* Diabetes, 2007. **56**(12): p. 2938-45.
8. Grueter, C.E., et al., *A cardiac microRNA governs systemic energy homeostasis by regulation of MED13.* Cell, 2012. **149**(3): p. 671-83.
9. Sacco, J. and K. Adeli, *MicroRNAs: emerging roles in lipid and lipoprotein metabolism.* Curr Opin Lipidol, 2012. **23**(3): p. 220-5.
10. Tian, Q., et al., *MicroRNA-550a acts as a pro-metastatic gene and directly targets cytoplasmic polyadenylation element-binding protein 4 in hepatocellular carcinoma.* PLoS One, 2012. **7**(11): p. e48958.
11. Bae, H.J., et al., *MicroRNA-29c functions as a tumor suppressor by direct targeting oncogenic SIRT1 in hepatocellular carcinoma.* Oncogene, 2013.
12. Ladeiro, Y. and J. Zucman-Rossi, *[miRNAs in cancer: the case of liver tumors]*. Med Sci (Paris), 2009. **25**(5): p. 467-72.
13. Glisovic, T., et al., *RNA-binding proteins and post-transcriptional gene regulation.* FEBS Lett, 2008. **582**(14): p. 1977-86.
14. Mittal, N., et al., *Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks.* Proc Natl Acad Sci U S A, 2009. **106**(48): p. 20300-5.
15. Colegrove-Otero, L.J., N. Minshall, and N. Standart, *RNA-binding proteins in early development.* Crit Rev Biochem Mol Biol, 2005. **40**(1): p. 21-73.
16. Blech-Hermoni, Y. and A.N. Ladd, *RNA binding proteins in the regulation of heart development.* Int J Biochem Cell Biol, 2013. **45**(11): p. 2467-78.
17. Vizzutti, F., et al., *Non-invasive assessment of fibrosis in non-alcoholic fatty liver disease.* Ann Hepatol, 2009. **8**(2): p. 89-94.
18. McCullough, A.J., *The clinical features, diagnosis and natural history of nonalcoholic fatty liver disease.* Clin Liver Dis, 2004. **8**(3): p. 521-33, viii.
19. Masuoka, H.C. and N. Chalasani, *Nonalcoholic fatty liver disease: an emerging threat to obese and diabetic individuals.* Ann N Y Acad Sci, 2013. **1281**: p. 106-22.
20. Yilmaz, Y., et al., *Linking nonalcoholic fatty liver disease to hepatocellular carcinoma: from bedside to bench and back.* Tumori, 2013. **99**(1): p. 10-6.
21. Rottiers, V. and A.M. Naar, *MicroRNAs in metabolism and metabolic disorders.* Nat Rev

Mol Cell Biol, 2012. **13**(4): p. 239-50.
22. Blade, C., et al., *miRNAs, polyphenols, and chronic disease.* Mol Nutr Food Res, 2013. **57**(1): p. 58-70.
23. Ahn, J., et al., *High fat diet induced downregulation of microRNA-467b increased lipoprotein lipase in hepatic steatosis.* Biochem Biophys Res Commun, 2011. **414**(4): p. 664-9.
24. Li, S., et al., *Differential expression of microRNAs in mouse liver under aberrant energy metabolic status.* J Lipid Res, 2009. **50**(9): p. 1756-65.
25. Shibayama, M., et al., *Polypyrimidine tract-binding protein is essential for early mouse development and embryonic stem cell proliferation.* FEBS J, 2009. **276**(22): p. 6658-68.
26. Suckale, J., et al., *PTBP1 is required for embryonic development before gastrulation.* PLoS One, 2011. **6**(2): p. e16992.
27. Katsanou, V., et al., *The RNA-binding protein Elavl1/HuR is essential for placental branching morphogenesis and embryonic development.* Mol Cell Biol, 2009. **29**(10): p. 2762-76.
28. Wakamatsu, Y. and J.A. Weston, *Sequential expression and role of Hu RNA-binding proteins during neurogenesis.* Development, 1997. **124**(17): p. 3449-60.
29. Leeds, P., et al., *Developmental regulation of CRD-BP, an RNA-binding protein that stabilizes c-myc mRNA in vitro.* Oncogene, 1997. **14**(11): p. 1279-86.
30. Torres-Padilla, M.E., C. Fougere-Deschatrette, and M.C. Weiss, *Expression of HNF4alpha isoforms in mouse liver development is regulated by sequential promoter usage and constitutive 3' end splicing.* Mech Dev, 2001. **109**(2): p. 183-93.
31. Thoren, L.A., et al., *UPF2 is a critical regulator of liver development, function and regeneration.* PLoS One, 2010. **5**(7): p. e11650.
32. Innocenti, F., et al., *Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue.* PLoS Genet, 2011. **7**(5): p. e1002078.
33. Gamazon, E.R., et al., *A genome-wide integrative study of microRNAs in human liver.* BMC Genomics, 2013. **14**: p. 395.
34. Li, M., et al., *Comparing morphometric, biochemical, and visual measurements of macrovesicular steatosis of liver.* Hum Pathol, 2011. **42**(3): p. 356-60.
35. Kleiner, D.E., et al., *Design and validation of a histological scoring system for nonalcoholic fatty liver disease.* Hepatology, 2005. **41**(6): p. 1313-21.
36. Strimmer, K., *fdrtool: a versatile R package for estimating local and tail area-based false discovery rates.* Bioinformatics, 2008. **24**(12): p. 1461-2.
37. Mi, H., A. Muruganujan, and P.D. Thomas, *PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees.* Nucleic Acids Res, 2013. **41**(Database issue): p. D377-86.
38. Mi, H. and P. Thomas, *PANTHER pathway: an ontology-based pathway database coupled with data analysis tools.* Methods Mol Biol, 2009. **563**: p. 123-40.
39. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
40. Reimand, J., et al., *g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W193-200.
41. John, B., et al., *Human MicroRNA targets.* PLoS Biol, 2004. **2**(11): p. e363.

42. Betel, D., et al., *Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites.* Genome Biol, 2010. **11**(8): p. R90.
43. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs.* Genome Res, 2009. **19**(1): p. 92-105.
44. Kinsella, R.J., et al., *Ensembl BioMarts: a hub for data retrieval across taxonomic space.* Database (Oxford), 2011. **2011**: p. bar030.
45. Flicek, P., et al., *Ensembl 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D48-55.
46. Griffiths-Jones, S., et al., *miRBase: tools for microRNA genomics.* Nucleic Acids Res, 2008. **36**(Database issue): p. D154-8.
47. Bauer-Mehren, A., *Integration of genomic information with biological networks using Cytoscape.* Methods Mol Biol, 2013. **1021**: p. 37-61.
48. Saldanha, A.J., *Java Treeview--extensible visualization of microarray data.* Bioinformatics, 2004. **20**(17): p. 3246-8.
49. Zhang, J., et al., *microRNA-150 inhibits human CD133-positive liver cancer stem cells through negative regulation of the transcription factor c-Myb.* Int J Oncol, 2012. **40**(3): p. 747-56.
50. Estep, M., et al., *Differential expression of miRNAs in the visceral adipose tissue of patients with non-alcoholic fatty liver disease.* Aliment Pharmacol Ther, 2010. **32**(3): p. 487-97.
51. Venugopal, S.K., et al., *Liver fibrosis causes downregulation of miRNA-150 and miRNA-194 in hepatic stellate cells, and their overexpression causes decreased stellate cell activation.* Am J Physiol Gastrointest Liver Physiol, 2010. **298**(1): p. G101-6.
52. Magrelli, A., et al., *Altered microRNA Expression Patterns in Hepatoblastoma Patients.* Transl Oncol, 2009. **2**(3): p. 157-63.
53. Schmitt, F., et al., *Lentiviral vectors that express UGT1A1 in liver and contain miR-142 target sequences normalize hyperbilirubinemia in Gunn rats.* Gastroenterology, 2010. **139**(3): p. 999-1007, 1007 e1-2.
54. Meng, F., et al., *Epigenetic regulation of miR-34a expression in alcoholic liver injury.* Am J Pathol, 2012. **181**(3): p. 804-17.
55. Castro, R.E., et al., *miR-34a/SIRT1/p53 is suppressed by ursodeoxycholic acid in the rat liver and activated by disease severity in human non-alcoholic fatty liver disease.* J Hepatol, 2013. **58**(1): p. 119-25.
56. Mizuguchi, Y., et al., *Sequencing and bioinformatics-based analyses of the microRNA transcriptome in hepatitis B-related hepatocellular carcinoma.* PLoS One, 2011. **6**(1): p. e15304.
57. Tryndyak, V.P., et al., *Down-regulation of the microRNAs miR-34a, miR-127, and miR-200b in rat liver during hepatocarcinogenesis induced by a methyl-deficient diet.* Mol Carcinog, 2009. **48**(6): p. 479-87.
58. Scagnolari, C., et al., *Differential expression of interferon-induced microRNAs in patients with chronic hepatitis C virus infection treated with pegylated interferon alpha.* Virol J, 2010. **7**: p. 311.
59. Ji, J., et al., *MicroRNA expression, survival, and response to interferon in liver cancer.* N Engl J Med, 2009. **361**(15): p. 1437-47.
60. Wei, X., et al., *miR-101 is down-regulated by the hepatitis B virus x protein and induces aberrant DNA methylation by targeting DNA methyltransferase 3A.* Cell Signal, 2013. **25**(2): p. 439-46.

61. Bae, J.S., et al., *Association study of genetic variations in microRNAs with the risk of hepatitis B-related liver diseases.* Dig Liver Dis, 2012. **44**(10): p. 849-54.
62. Au, S.L., et al., *Enhancer of zeste homolog 2 epigenetically silences multiple tumor suppressor microRNAs to promote liver cancer metastasis.* Hepatology, 2012. **56**(2): p. 622-31.
63. Ji, F., et al., *Circulating microRNAs in hepatitis B virus-infected patients.* J Viral Hepat, 2011. **18**(7): p. e242-51.
64. Gui, J., et al., *Serum microRNA characterization identifies miR-885-5p as a potential marker for detecting liver pathologies.* Clin Sci (Lond), 2011. **120**(5): p. 183-93.
65. Zhou, J., et al., *Plasma microRNA panel to diagnose hepatitis B virus-related hepatocellular carcinoma.* J Clin Oncol, 2011. **29**(36): p. 4781-8.
66. Villanueva, A., et al., *New strategies in hepatocellular carcinoma: genomic prognostic markers.* Clin Cancer Res, 2010. **16**(19): p. 4688-94.
67. Padgett, K.A., et al., *Primary biliary cirrhosis is associated with altered hepatic microRNA expression.* J Autoimmun, 2009. **32**(3-4): p. 246-53.
68. Lu, T., et al., *Gene regulation and DNA damage in the ageing human brain.* Nature, 2004. **429**(6994): p. 883-91.
69. Rodwell, G.E., et al., *A transcriptional profile of aging in the human kidney.* PLoS Biol, 2004. **2**(12): p. e427.
70. Thomas, R.P., et al., *Age-associated changes in gene expression patterns in the liver.* J Gastrointest Surg, 2002. **6**(3): p. 445-53; discussion 454.
71. Glass, D., et al., *Gene expression changes with age in skin, adipose tissue, blood and brain.* Genome Biol, 2013. **14**(7): p. R75.
72. Buckberry, S., et al., *Integrative transcriptome meta-analysis reveals widespread sex-biased gene expression at the human fetal-maternal interface.* Mol Hum Reprod, 2014.
73. Jansen, R., et al., *Sex differences in the human peripheral blood transcriptome.* BMC Genomics, 2014. **15**: p. 33.
74. Jessen, H.M. and A.P. Auger, *Sex differences in epigenetic mechanisms may underlie risk and resilience for mental health disorders.* Epigenetics, 2011. **6**(7): p. 857-61.
75. Castello, A., et al., *System-wide identification of RNA-binding proteins by interactome capture.* Nat Protoc, 2013. **8**(3): p. 491-500.
76. Baltz, A.G., et al., *The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts.* Mol Cell, 2012. **46**(5): p. 674-90.
77. Ray, D., et al., *A compendium of RNA-binding motifs for decoding gene regulation.* Nature, 2013. **499**(7457): p. 172-7.
78. Kwon, S.C., et al., *The RNA-binding protein repertoire of embryonic stem cells.* Nat Struct Mol Biol, 2013. **20**(9): p. 1122-30.
79. Cook, K.B., et al., *RBPDB: a database of RNA-binding specificities.* Nucleic Acids Res, 2011. **39**(Database issue): p. D301-8.
80. Schroder, A., et al., *Genomics of ADME gene expression: mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver.* Pharmacogenomics J, 2013. **13**(1): p. 12-20.
81. Dunning, M.J., et al., *beadarray: R classes and methods for Illumina bead-based data.* Bioinformatics, 2007. **23**(16): p. 2183-4.
82. Du, P., W.A. Kibbe, and S.M. Lin, *lumi: a pipeline for processing Illumina microarray.* Bioinformatics, 2008. **24**(13): p. 1547-8.

83. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
84. Charoensawan, V., D. Wilson, and S.A. Teichmann, *Genomic repertoires of DNA-binding transcription factors across the tree of life.* Nucleic Acids Res, 2010. **38**(21): p. 7364-77.
85. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.
86. Brandes, U., *A Faster Algorithm for Betweenness Centrality.* Journal of Mathematical Sociology 2001. **25**(2): p. 163-177.
87. Kalsotra, A., et al., *The Mef2 transcription network is disrupted in myotonic dystrophy heart tissue, dramatically altering miRNA and mRNA expression.* Cell Rep, 2014. **6**(2): p. 336-45.
88. Mili, S. and S. Pinol-Roma, *LRP130, a pentatricopeptide motif protein with a noncanonical RNA-binding domain, is bound in vivo to mitochondrial and nuclear RNAs.* Mol Cell Biol, 2003. **23**(14): p. 4972-82.
89. Castello, A., et al., *Insights into RNA biology from an atlas of mammalian mRNA-binding proteins.* Cell, 2012. **149**(6): p. 1393-406.
90. Vilborg, A., et al., *The p53 target Wig-1 regulates p53 mRNA stability through an AU-rich element.* Proc Natl Acad Sci U S A, 2009. **106**(37): p. 15756-61.
91. Vilborg, A., et al., *The p53 target Wig-1: a regulator of mRNA stability and stem cell fate?* Cell Death Differ, 2011. **18**(9): p. 1434-40.
92. Lv, J., et al., *PCDH20 functions as a tumour-suppressor gene through antagonizing the Wnt/beta-catenin signalling pathway in hepatocellular carcinoma.* J Viral Hepat, 2014.
93. Nigro, P., G. Pompilio, and M.C. Capogrossi, *Cyclophilin A: a key player for human disease.* Cell Death Dis, 2013. **4**: p. e888.
94. Xie, X., Z. Wang, and Y. Chen, *Association of LKB1 with a WD-repeat protein WDR6 is implicated in cell growth arrest and p27(Kip1) induction.* Mol Cell Biochem, 2007. **301**(1-2): p. 115-22.
95. McKnight, N.C., et al., *Genome-wide siRNA screen reveals amino acid starvation-induced autophagy requires SCOC and WAC.* EMBO J, 2012. **31**(8): p. 1931-46.
96. Collins, L.J., et al., *The modern RNP world of eukaryotes.* J Hered, 2009. **100**(5): p. 597-604.
97. Nepusz, T., H. Yu, and A. Paccanaro, *Detecting overlapping protein complexes in protein-protein interaction networks.* Nat Methods, 2012. **9**(5): p. 471-2.
98. Koschutzki, D. and F. Schreiber, *Centrality analysis methods for biological networks and their application to gene regulatory networks.* Gene Regul Syst Bio, 2008. **2**: p. 193-201.
99. Lendvai, G., et al., *[Alterations in microRNA expression patterns in liver diseases]*. Orv Hetil, 2010. **151**(45): p. 1843-53.
100. Flowers, E., E.S. Froelicher, and B.E. Aouizerat, *MicroRNA regulation of lipid metabolism.* Metabolism, 2013. **62**(1): p. 12-20.
101. Bieche, I., et al., *Molecular profiling of early stage liver fibrosis in patients with chronic hepatitis C virus infection.* Virology, 2005. **332**(1): p. 130-44.
102. Asselah, T., et al., *Liver gene expression signature of mild fibrosis in patients with chronic hepatitis C.* Gastroenterology, 2005. **129**(6): p. 2064-75.
103. Sookoian, S., et al., *Liver transcriptional profile of atherosclerosis-related genes in*

human nonalcoholic fatty liver disease. Atherosclerosis, 2011. **218**(2): p. 378-85.
104. Fourcot, A., et al., *Gas6 deficiency prevents liver inflammation, steatohepatitis, and fibrosis in mice.* Am J Physiol Gastrointest Liver Physiol, 2011. **300**(6): p. G1043-53.
105. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.
106. Masuda, K., et al., *General RBP expression in human tissues as a function of age.* Ageing Res Rev, 2012. **11**(4): p. 423-31.

PRANEET CHATURVEDI

322 Canal Walk, Apt #275, Indianapolis, IN 46202
Ph: (317)-909-2567, Email: pranchat@iupui.edu

QUALIFICATION SUMMARY

1. Well versed with Next Generation Sequencing data analysis pipelines using NGS tools & Micro Array data analysis using Bioconductor packages in R"
2. Experience of 3 years at computational and analytical skills developed over the years in Bioinformatics
3. Proficient in perl, C & R for data analysis, handling and statistical analysis
4. Developed command line application using perl for conversion of large experimental datasets to desired database format for ADME dept. in Lilly(part of internship project)
5. Competent with UCSC, j-Browse. IGV, ENSEMBL for the visualization of different genomes and adept with Cytoscape and its plugins for visualization of biological networks
6. Analyzed datasets from 1000 genomes project, TCGA, SRA, GEO, Drugbank and Ensembl for different projects
7. Designed a pipeline to perform variate statistics like PCA, ANOVA, MANOVA & LDA using the combination of perl and R
8. Adept with Matrix eQTL R package for performing ultra-fast eQTL analysis
9. Thorough knowledge of Windows, Linux/Unix and MAC OS X
10. A highly enthusiastic & motivated team player with flexible thinking & adept at working in diverse roles

EDUCATIONAL QUALIFICATIONS

Master of Science, Bioinformatics **December, 2014**
Indiana University - Purdue University Indianapolis, Indianapolis, Indiana
GPA: 3.713/4.00

Bachelor of Technology, Bioinformatics **May, 2012**
VIT University, Vellore, India
GPA: 3.52/4.00

PROFESSIONAL EXPERIENCE

1. **Academic Intern, Eli Lilly & Company, Indianapolis, Indiana, USA**
   June 2014 – present
   - Help in direct curation, standardizing, automating and facilitating data extraction from tables and reports.
   - Design models and then simulate these models with ADME properties.
   - Developed command line application for converting raw experimental data into desired format for storage and analysis
   - Member of Research Agreement between Eli Lilly and IUPUI to conduct downstream analysis of study of pathway dysregulation due to CYPs dysregulation by inducement of drugs
2. **Intern, Bioinformatics Institute of India, Noida, India**
   June 2009

- Genomic Analysis of Genes Responsible for Cystic Fibrosis in Homo Sapiens (under the guidance of Dr. Kumud Sarin)
- During this training/internship I have developed a pipeline to analyze the genes involved in Cystic Fibrosis and did a pathway analysis to find out affected pathways due to mutation in genes involved in Cystic Fibrosis.

RESEARCH EXPERIENCE

Janga Lab of Genomics and Systems Biology, Indiana University-Purdue University Indianapolis, Indianapolis, Indiana

Research Assistant
Spring 2013 - Present

Thesis: Comprises of three projects with generalized title: "Transcriptional & post-transcriptional regulation aspects in liver and liver disease"

(I) Integrated miR-miRNA Network Underlying Hepatic Fat Accumulation in Humans

(II) Study of Transcriptional Regulation of CYP's Dysreguled in Non-Alcoholic Fatty Liver Diseases

(III) Understanding the importance of RNA binding proteins as a function of age and gender during liver development (collaboration project between Janga lab and Department of Biochemistry and Medical Biochemistry, UIUC)

Lab Projects:

1. Differential miRNA expression in VSMC (vascular smooth muscle cells) and MV (matrix vesicles) may be involved in vascular calcification (collaboration project between Janga lab and Department of nephrology, IU school of Medicine)
2. Correlating eQTLs in non-alcoholic fatty liver disease with respect to age and gender using Matrix eQTL(R Package)
3. Construction of Drug - Target networks across 16 tissues in human body based on side effect propensity

ACADEMIC PROJECTS

1. Mining of Anti-Bacterial & Anti-Tumor Metabolites in Human Microbiome Project data & Prediction of Body Sites with Significant Abundance of these Metabolites
2. Development of Pancreatic Cancer Research Database (Team Project)
3. Pathway analysis of differentially expressed Genes in PD controlled by miR's associated to PD using RNA-seq in human brain cortex
4. Uncovering the mechanism of alternative splicing in RBPs in LIHC (Liver Hepatocellular Carcinoma)(Team Project)
5. Identification of factors causing dysregulation of genes and discovering biomarkers using gene terrain in Parkinson Disease
6. Analysis of AaeL-AAEL015099 as a drug target for the disease Yellow Fever (Undergraduate Thesis Project)

BIOINFORMATICS QUALIFICATIONS

Languages	Perl, R, Bash, C, HTML & MySQL
Visualization Tools	Cytoscape, IGV, Java Tree View, j-Browse, UCSC, ENSEMBL, Gene Terrain & Tibco Spotfire
NGS Analysis	SAM Tools, BED Tools ,VCF Tools, Bowtie, Tophat, STAR aligner, SailFish, Cufflinks pipeline & GATK toolkit
Microarray Analysis	MeV, GSEA, Cluster & R Bio-conductor Packages
Functional Annotation & Pathway Analysis	IPA, David, Panther, Gprofiler & MetaNet
Motif Detection tools	MEME, TOMTOM & FIMO

POSTER PRESENTATION

1. Poster Presentation at RRM Meeting ’14, Pittsburg**,** OCT 2014 on “Understanding the importance of RNA binding proteins as a function of age and gender during liver development”

2. Poster Presentation at CTSI Indiana, Indianapolis, Sept 2013 on “Integrated miR-miRNA Network Underlying Hepatic Fat Accumulation in Humans”