

# Candidate genes investigation for severe nonalcoholic fatty liver disease based on bioinformatics analysis

Shan Qi, BM & Associate Chief Physician<sup>a</sup>, Changhong Wang, BM & Associate Chief Physician<sup>a</sup>, Chunfu Li, BM<sup>a</sup>, Pu Wang, BM<sup>b</sup>, Minghui Liu, MM & Attending Physician<sup>a,\*</sup>

## Abstract

**Background:** Nonalcoholic fatty liver disease (NAFLD) is the most common chronic liver condition worldwide. However, its etiology and fundamental pathophysiology for the disease process are poorly understood. In this study, we thus used bioinformatics to identify candidate genes potentially causative of severe NAFLD.

**Methods:** Gene expression profile data GSE49541 were downloaded from the Gene Expression Omnibus database. Tissues samples from 32 severe and 40 mild NAFLD patients were evaluated to identify differentially expressed genes (DEGs) between the 2 groups, followed by analyses of Gene Ontology (GO) functions and Kyoto Encyclopedia of Genes and Genomes pathways. Then, a weighted protein–protein interaction (PPI) network was constructed, and subnetworks and candidate genes were screened. Moreover, the GSE48452 data (14 normal liver tissue samples and 18 nonalcoholic steatohepatitis samples) were used to verify the results obtained from the above analyses.

**Results:** A total of 100 upregulated genes and 24 downregulated ones were identified in severe NAFLD. Functional enrichment and pathway analyses showed that these DEGs were mainly associated with cell adhesion, inflammatory response, and chemokine activity. The top 5 subnetworks were selected based on the PPI network. A total of 5 hub genes, including ubiquilin 4 (UBQLN4), amyloid-beta precursor protein (APP), sex hormone-binding globulin (SHBG), cadherin-associated protein beta 1 (CTNNB1) and collagen type I alpha 1 (COL1A1), were considered to be candidate genes for NAFLD. In addition, the verification data confirmed the status of *COL1A1*, *SHBG*, and *APP* as candidate genes.

**Conclusion:** *UBQLN4*, *APP*, *CTNNB1*, *SHBG*, and *COL1A1* might be involved in the development of NAFLD, and are proposed as the potential markers for predicting the development of this condition.

**Abbreviations:** APP = amyloid-beta precursor protein, COL1A1 = collagen type I alpha 1, CTNNB1 = cadherin-associated protein beta 1, DAVID = The Database for Annotation, Visualization, and Integrated Discovery, DEG = differentially expressed gene, ES = enrichment score, FC = fold change, GEO = Gene Expression Omnibus, GO = Gene Ontology, HPRD = Human Protein Reference Database, KEGG = Kyoto Encyclopedia of Genes and Genomes, NAFLD = Nonalcoholic fatty liver disease, NASH = nonalcoholic steatohepatitis, PPI = protein–protein interaction, RMA = Robust Multichip Averaging, SHBG = sex hormone-binding globulin, UBQLN4 = ubiquilin 4.

**Keywords:** differentially expressed genes, functional enrichment analysis, hub genes, nonalcoholic fatty liver disease, protein–protein interaction network

Editor: Ewa Janczewska.

The authors have no conflicts of interest to disclose.

Supplemental Digital Content is available for this article.

<sup>a</sup> Department of Traditional Chinese Medicine, China-Japan Union Hospital of Jilin University, <sup>b</sup> Clinical Medicine College, Jilin University, Changchun, Jilin Province, China.

\* Correspondence: Minghui Liu, Department of Traditional Chinese Medicine, China-Japan Union Hospital of Jilin University, No. 126, Xiantai Street, Economic-Technological Development Area, Changchun, Jilin Province 130033, China (e-mail: liumh@jlu.edu.cn).

Copyright © 2017 the Author(s). Published by Wolters Kluwer Health, Inc. This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Medicine (2017) 96:32(e7743)

Received: 20 January 2017 / Received in final form: 25 June 2017 / Accepted: 18 July 2017

<http://dx.doi.org/10.1097/MD.0000000000007743>

## 1. Introduction

Nonalcoholic fatty liver disease (NAFLD) is the most common chronic liver condition worldwide.<sup>[1]</sup> Its incidence is as high as 30% in developed countries and nearly 10% in developing countries.<sup>[2]</sup> NAFLD is a clinicopathologic syndrome that encompasses several clinical entities, ranging from simple steatosis to steatohepatitis, fibrosis, and end-stage liver disease.<sup>[3]</sup> Because the etiology and fundamental pathophysiology underlying NAFLD are poorly understood,<sup>[4]</sup> it is challenging to diagnose and treat NAFLD patients before symptomatic cirrhosis or arises.

Gene expression profiling is considered as a powerful tool for exploring diagnostic and predictive biomarkers, especially in the targeting therapy for diseases such as cancer.<sup>[5,6]</sup> In a previous study, the differences in expression of the transcriptome and proteome in the liver between NAFLD and normal were determined using a gene expression profile, and several key pathways involved xenobiotic and lipid metabolism, inflammatory response, and cell-cycle control were identified.<sup>[7]</sup> A study on the differential gene expression in nonalcoholic steatohepatitis

(NASH) also showed that the uptake transporter genes were coordinately targeted for downregulation at the global level during the pathological development of NASH.<sup>[8]</sup> In addition, using microarray analysis, Moylan et al<sup>[9]</sup> indicated that the expression of certain metabolism-related genes was induced in severe NAFLD. Although many genes have been reportedly related to the process of NAFLD development, candidate genes potentially causative of severe NAFLD have not yet been screened, and it has remained unclear whether such genes induce severe NAFLD by interacting with each other.

Against the above background, the present study involved an exploration of the candidate genes of NAFLD, the establishment of a weighted regulatory network and the mining of hub genes in severe NAFLD samples compared with mild NAFLD samples were explored by using bioinformatic methods. We aimed to provide molecular mechanisms underlying NAFLD, and investigate new therapeutic targets of NAFLD.

## 2. Materials and methods

### 2.1. Samples

Gene expression profile data GSE49541<sup>[9]</sup> were downloaded from Gene Expression Omnibus (GEO) database (<http://www.ncbi.nlm.nih.gov/geo/>) with platform GPL570 (HG-U133\_2) Affymetrix Human Genome U133 Plus 2.0 Array. Data on 72 tissue samples were downloaded, including those from 32 severe NAFLD patients (fibrosis stages 3–4) and 40 mild NAFLD patients (fibrosis stage 0–1). The groups were matched for sex, age ( $\pm 5$  years) and body mass index ( $\text{kg/m}^2$ ) ( $\pm 3$  points). The samples were collected from NAFLD cases in the Duke University Health System NAFLD Biorepository, with approval from the Institutional Review Board of Duke University. Biorepository liver samples are remnants from clinically indicated liver biopsies.

### 2.2. Data preprocessing and differential expression analysis

The normalization of gene expression profile data was performed using the Robust Multichip Averaging (RMA) method<sup>[10]</sup> of the affy package<sup>[11]</sup> in R (v.3.0.0) (<http://bioconductor.org/biocLite.R>), and the Linear Models for Microarray Data (limma, <http://www.bioconductor.org/packages/release/bioc/html/limma.html>) package<sup>[12]</sup> was applied to identify the differentially expressed genes (DEGs) by comparing the gene expression levels in samples between mild and severe NAFLD cases. Resampling-based empirical Bayes multiple testing procedures<sup>[13]</sup> were also conducted to correct the *P* value. Subsequently, an adjusted *P* value  $< .05$  and  $|\log\text{FC}$  (log fold change)  $> 0.58$  were selected as the thresholds for DEG screening.

### 2.3. Gene Ontology annotation and pathway analysis

The Database for Annotation, Visualization, and Integrated Discovery (DAVID)<sup>[14]</sup> (<http://david.abcc.ncifcrf.gov/>) is a tool for the functional classification of genes that provides a comprehensive set of functional annotation tools enabling investigators to understand the biological meaning behind large lists of genes. Here, DAVID was used for GO annotation analysis (including the 3 categories of biological process, cellular component, and molecular function 3 aspects) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis. A *P* value  $< .05$  was considered to indicate a significant difference.

### 2.4. Weighted regulatory network construction

The protein–protein interactions (PPIs) that were related to genes in GSE49541 were selected according to the Human Protein Reference Database (HPRD, (<http://www.hprd.org/>)).<sup>[15]</sup> The average value of rank correlation coefficient ( $|\overline{r_{E_{ij}}}|$ ) and the difference ( $|\Delta r_{E_{ij}}|$ ) for pairs of regulatory relationship in PPIs were calculated according to the Eq. (1). The average absolute value of the rank correlation coefficient in control samples was considered as the weight of PPI,<sup>[16]</sup> and then the weighted PPI network was constructed.

$$r_{E_{ij}} = \frac{\sum(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum(x_{ik} - \bar{x}_i) \sum(x_{jk} - \bar{x}_j)}} \quad (1)$$

Here,  $E_{ij}$  is the edge between gene  $V_i$  and gene  $V_j$ ,  $k$  is the  $k^{\text{th}}$  sample,  $V_i$  and  $V_j$  are ranked by their expression in the samples, respectively;  $X_{jk}$  is the rank of  $V_j$  of  $k^{\text{th}}$  sample,  $X_{ik}$  is the rank of  $V_i$  of  $k^{\text{th}}$  sample,  $\bar{x}_i$  and  $\bar{x}_j$  are the average ranks of  $V_i$  and  $V_j$  in the samples, respectively.

$$|\overline{r_{E_{ij}}}| = \frac{1}{2} |r_{E_{ij1}} + r_{E_{ij2}}| \quad (2)$$

$$|\Delta r_{E_{ij}}| = \frac{1}{2} |r_{E_{ij1}} - r_{E_{ij2}}| \quad (3)$$

Here,  $r_{E_{ij1}}$  and  $r_{E_{ij2}}$  represent the Spearman coefficients of  $r_{E_{ij}}$  in 2 samples, respectively.

Finally, based on the permutation test, the random  $|\Delta r_{E_{ij}}|$  of each PPI was calculated. Subsequently, the sample labels were permuted for 10,000 times and a random  $|\Delta r_{E_{ij}}|$  was generated. The PPIs with a  $|\Delta r_{E_{ij}}|$  value  $> 90\%$  random  $|\Delta r_{E_{ij}}|$  value were filtered out.<sup>[16]</sup>

### 2.5. Subnetwork investigation and protein–protein interaction score calculation

In the PPI network, the nodes with a degree  $> 15$  were defined as candidate genes potentially causative of disease, and the subnetworks consisted of candidate genes and the genes with which they interact. The score of PPIs in the subnetworks were also calculated. Briefly, all PPIs in the weighted network were ranked from large to small according to their weight coefficient and this was defined as the background set (*E*), whereas the subnetwork was defined as the objective set (*S*). Then, the enrichment score (ES) of the subnetwork was calculated by walking down background set using the Gene Set Enrichment Analysis method<sup>[17]</sup> (<http://www.broadinstitute.org/gsea/index.jsp>). The formula is listed as below:

$$P_{\text{hit}}(S, i) = \sum_{E_j \in S, j \leq i} \frac{|r_j|^P}{N_R},$$

where

$$N_R = \sum_{E_j \in S} |r_j|^P$$

$$P_{\text{miss}}(S, i) = \sum_{E_j \in S, j \leq i} \frac{1}{N - N_H} \quad (4)$$

where  $E_j$  is the  $j^{\text{th}}$  PPI in the ranked regulatory pairs;  $r_j$  is the weight of the  $j^{\text{th}}$  PPI pair in background set;  $P$  is a parameter and set as 1;  $N$  is the number of PPI in *E*;  $N_H$  is the number of PPI in the subnet *S*. The ES was equal to the maximum deviation between  $P_{\text{hit}}$  and  $P_{\text{miss}}$ .

The PPI without contribution to ES was removed from subnetwork.<sup>[16,17]</sup> To estimate the significance of ES of the subnetwork, pairs of background regulatory relationships were rearranged randomly for 1000 times, the random ES of subnetwork was calculated, while the ES was transformed into Z value based on the equation<sup>[16]</sup>:

$$Z_s = \frac{ES - \overline{ES}}{S'}, \quad (5)$$

where ES (bar) is the mean of the random ES set and  $S'$  is the standard deviation of the random ES set.

### 2.6. Calculating the enrichment score of differentially expressed genes in subnetwork

On the basis of the subnetwork obtained above, the ES of DEGs were calculated. Briefly, the genes in gene microarray and subnetwork were considered as background set and objective set, respectively; then, these genes were arranged from large to small; the ES of subnetwork walked by background set was calculated using the following equation<sup>[16]</sup>:

$$P_{\text{hit}}(S_{\text{trimmed}}, i) = \sum_{g_j \in S_{\text{trimmed}}, j \leq i} \frac{|r_j|^P}{M_R},$$

where

$$M_R = \sum_{g_j \in S_{\text{trimmed}}} |r_j|^P$$

$$P_{\text{miss}}(S_{\text{trimmed}}, i) = \sum_{g_j \in S_{\text{trimmed}}, j \leq i} \frac{1}{M - MH}, \quad (6)$$

Here,  $g_j$  is the  $j^{\text{th}}$  gene in the ranked genes,  $r_j$  is the magnitude of differential expression of the  $j^{\text{th}}$  gene,  $P$  is a parameter and set as 1,  $M$  is the number of genes in  $L$ , and  $MH$  is the number of genes in  $S_{\text{trimmed}}$ .

To estimate the significance of ES of the subnetwork, the background genes were rearranged randomly for 1000 times. The random ES of the subnetwork was calculated, followed by the ES being transformed into a Z value based on the Eq. (5).

### 2.7. Candidate genes screening

The 2 Z values obtained as described above were normalized and summed, followed by the acquisition of the combined these 2 parts. Here, using this combined score, the top 5 subnetworks were considered as candidate subnetworks of the disease, whereas the hub genes were defined as candidate genes of the disease.<sup>[16]</sup>

### 2.8. Data verification

Gene expression profile data GSE48452 were downloaded from the GEO database (<http://www.ncbi.nlm.nih.gov/geo/>) with the platform (HuGene-1\_1-st) Affymetrix Human Gene 1.1 ST Array [transcript (gene) version], including data on 14 normal liver tissue samples and 18 NASH samples. The liver samples had been obtained from NAFLD patients, who had provided written informed consent. The study protocol was also approved by the institutional review board ("Ethikkommission der Medizinischen Fakultät der University Kiel," D425/07, A111/99). Empirical Bayes analysis using  $t$  test procedures in the limma package (Version 3.10.3, <http://www.bioconductor.org/packages/2.9/bioc/html/limma.html>) was applied to identify the DEGs by comparing gene expression levels between normal and NASH samples.

Subsequently, an adjusted  $P$  value  $<.05$  and  $|\log_{2}FC| >0.58$  were selected as the significance thresholds for DEGs. Subsequently, the Pearson correlation analysis was performed to confirm the differences between the key candidate genes in severe NAFLD and the target DEGs with which they interacted.

## 3. Results

### 3.1. Identification of differentially expressed genes

The gene expression profile data were normalized by using the RMA method (Fig. 1). Upon applying the thresholds of an adjusted  $P$  value  $<.05$  and  $|\log_{2}FC| >0.58$ , a total of 124 DEGs were identified, including 100 upregulated genes and 24 downregulated ones (Supplementary Table 1, <http://links.lww.com/MD/B825>). The heatmap of the DEGs is shown in Figure 2. The top 10 upregulated genes and downregulated genes are listed in Table 1.

### 3.2. Gene Ontology annotation analysis and pathway analysis

GO annotation analysis of the DEGs was performed by using DAVID, the main results of which are listed in Table 2 (Supplementary Table 2, <http://links.lww.com/MD/B826>). The biological processes that were particularly commonly associated with NAFLD were cell adhesion, chemotaxis, collagen fibril organization, and inflammatory response; the cellular components were extracellular matrix and collagen; and the molecular function was chemokine activity. Furthermore, the identified DEGs were significantly clustered in cell adhesion and chemokine signaling pathways.

### 3.3. Weighted protein-protein interaction network analysis

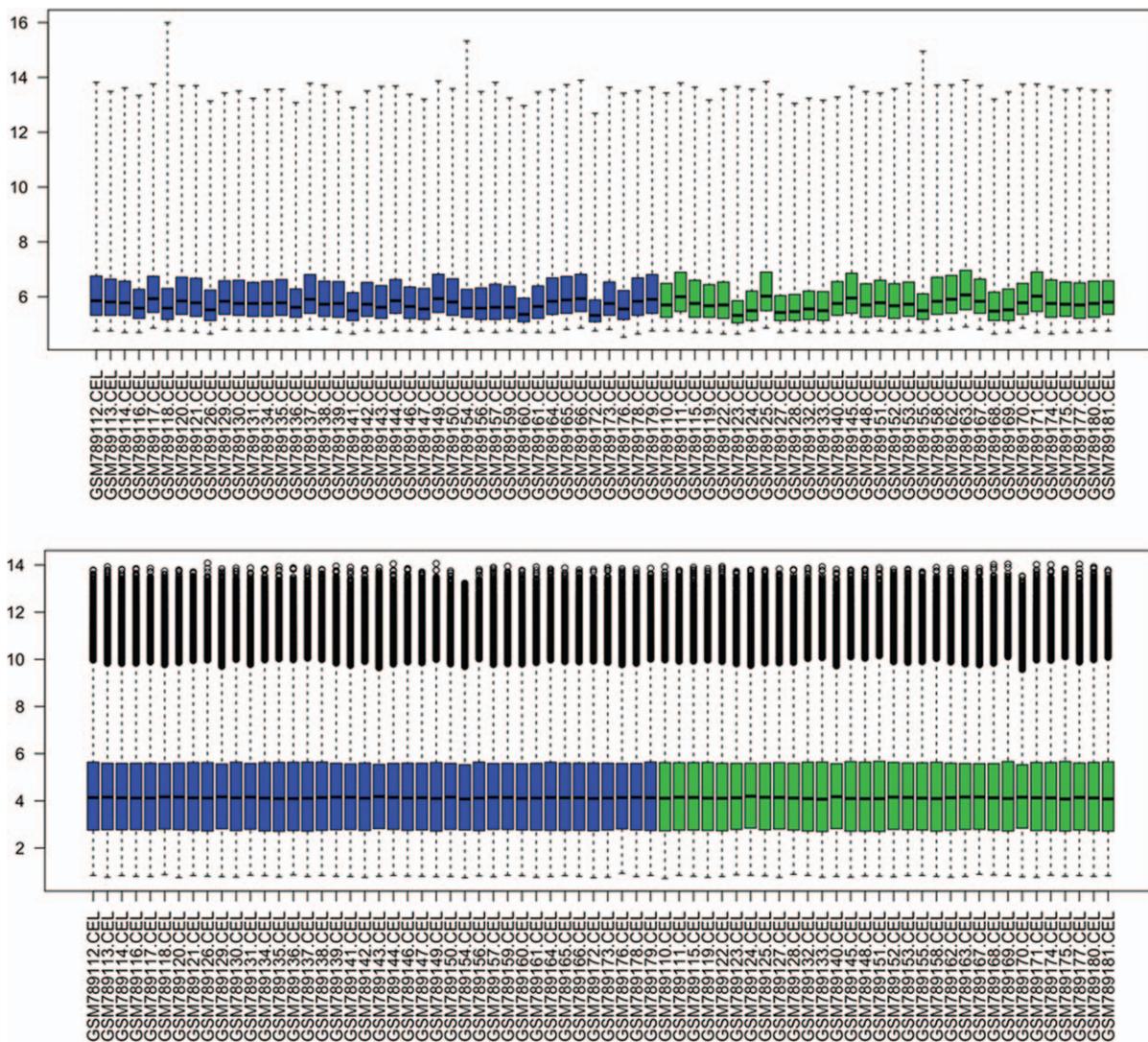
Based on the HPRD, the PPIs related to genes in the gene microarray were selected, and a PPI network was constructed. The entire network consisted of 8964 nodes and 34,915 edges (Fig. 3).

### 3.4. The enrichment score of subnetwork and candidate genes

By calculating the weighted PPIs of the subnetwork and the score of DEGs, the significance of ES was estimated with Z value. Then, the top 5 subnetworks with the highest Z values were selected (Fig. 4). The corresponding hub genes were considered as candidate genes, which included ubiquitin 4 (UBQLN4), also known as ataxin-1-interacting protein (A1UP), amyloid-beta precursor protein (APP), sex hormone-binding globulin (SHBG), cadherin-associated protein beta 1 (CTNNA1) and collagen type I alpha 1 (COL1A1) (Table 3).

### 3.5. The results of verification data

To confirm the above results, the DEGs between normal and NASH samples were identified through data verification. A total of 181 DEGs were identified, including 119 upregulated DEGs and 62 downregulated DEGs (Supplementary Table 3, <http://links.lww.com/MD/B827>). As shown in Table 4, *COL1A1* and *SHBG* exhibited significant differences in their expression level. *AKR1B10* and *CYP2C19*, which were included in the original top10 DEGs, were also verified. *UBQLN4*, *APP*, and *SHBG* interacted with the target DEGs in the PPI networks. Pearson correlation analysis revealed that there were significant positive



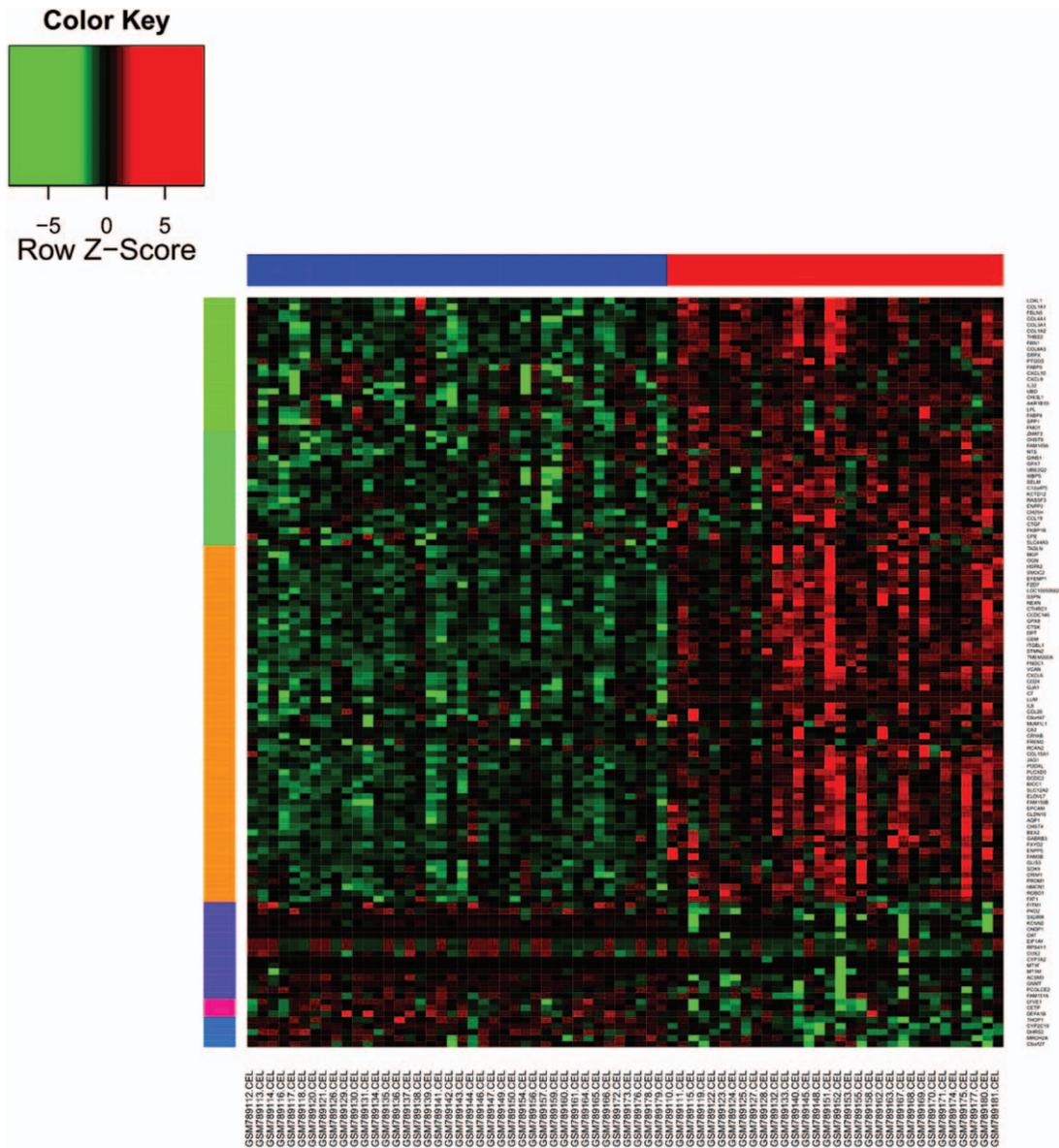
**Figure 1.** Box plot of gene expression profile data before and after normalization. The top box represents the distribution of data before normalization, and the bottom box represents the distribution of data after normalization. Horizontal axis represents 72 nonalcoholic fatty liver disease samples.

correlations between *APP* and its target-DEGs (*COL1A2*, *CRYAB*, *COL4A1*) ( $P < .05$ , Fig. 4A–C).

**4. Discussion**

NAFLD, is an emerging public health problem, that may be a highly chronic liver condition.<sup>[1]</sup> Although the histologic method has been approved, a useful therapy method for treating NAFLD is yet to be established.<sup>[18]</sup> Gene expression profiling is valuable for researching diagnostic and predictive biomarkers of disease, including NAFLD,<sup>[9]</sup> which may facilitate the development of new therapeutic drugs. In the present study, a total of 124 DEGs associated with severe NAFLD were identified by comparing the gene expression profile with that in mild NAFLD. These DEGs were mainly associated with cell adhesion, inflammatory response, and chemokine signaling pathways. Selection of top 5 subnetworks based on the PPI network indicated that *UBQLN4*, *APP*, *CTNNB1*, *SHBG*, and *COL1A1* may be involved in the development of NAFLD, and these were considered to be markers with potential utility for predicating NAFLD.

In a previous study, inflammation was proved to be associated with the process of NAFLD development.<sup>[19]</sup> Mikolasevic et al<sup>[20]</sup> demonstrated that in patients maintained on hemodialysis, there is probably some interaction between NAFLD and inflammation, malnutrition, and atherosclerosis. In support of this, a significant correlation between the intima-media thickness of the carotid artery and hepatic inflammation score was identified in NAFLD rats.<sup>[21]</sup> Moreover, Browning and Horton<sup>[22]</sup> indicated that the histological hallmarks of NASH, such as inflammation, cell death, and fibrosis promoted the progression of NAFLD. In fact, findings showed that inflammation resulting in a stress response of hepatocytes, might lead to lipid accumulation, and therefore could precede steatosis in NASH.<sup>[23]</sup> In another study based on gene expression in human cases of NAFLD, Greco et al<sup>[19]</sup> asserted that cell adhesion was significantly associated with liver fat content. In severe liver injury, neural cell adhesion molecules weaken the cell–cell and cell–matrix interactions, thereby allowing ductular reactions/hepatic progenitor cells to migrate for normal development and regeneration.<sup>[24]</sup> In fact, each cell adhesion molecule may play an important role during development in hepatic histogenesis, including hepatoblast/



**Figure 2.** The expression of differentially expressed genes in nonalcoholic fatty liver disease samples. The bottom horizontal axis represents nonalcoholic fatty liver disease samples, the first 40 samples were from mild nonalcoholic fatty liver disease samples and the other 32 samples were from severe nonalcoholic fatty liver disease. The right vertical axis represents genes. Red indicates upregulated genes and green indicates downregulated genes.

**Table 1**

**The top 10 upregulated and downregulated differentially expressed genes.**

Gene	logFC	Adjusted P value	Gene	logFC	Adjusted P value
CXCL6	1.901064	7.62E-11	CYP2C19	-1.80146	1.00E-05
EPCAM	1.87717	1.27E-09	DHRS2	-1.21959	2.38E-04
AKR1B10	1.690964	7.00E-03	MT1M	-1.13824	8.38E-03
CD24	1.571155	8.31E-11	RPS4Y1	-0.98888	4.60E-02
THBS2	1.386229	3.86E-12	GNMT	-0.96755	1.51E-03
LUM	1.34156	2.70E-13	C5orf27	-0.94409	2.82E-02
STMN2	1.314227	1.19E-12	OAT	-0.89612	3.22E-04
UBD	1.269138	1.37E-04	EIF1AY	-0.8862	4.15E-02
GEM	1.266165	1.22E-11	FAM151A	-0.86835	3.12E-02
CHI3L1	1.260242	1.41E-04	CNDP1	-0.85161	5.27E-04

FC = fold change.

**Table 2**  
Gene Ontology annotation analysis of differentially expressed genes in 3 aspects.

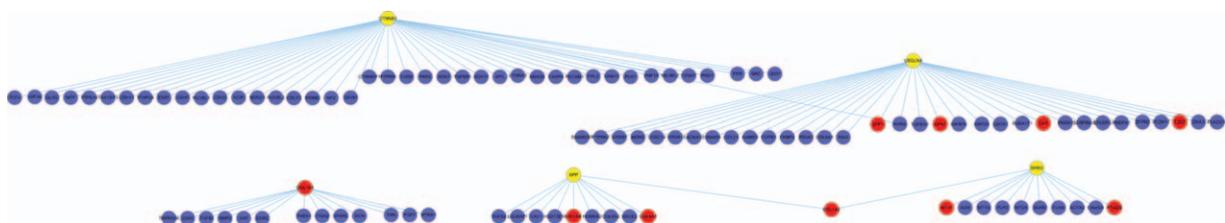
GO	Term	Count	P	
BP	GO:0007155~cell adhesion	23	2.60E-09	
	GO:0022610~biological adhesion	23	2.67E-09	
	GO:0009611~response to wounding	17	8.94E-07	
	GO:0042330~taxis	9	1.76E-05	
	GO:0006935~chemotaxis	9	1.76E-05	
	GO:0030199~collagen fibril organization	5	4.94E-05	
	GO:0006954~inflammatory response	11	1.01E-04	
	GO:0007626~locomotory behavior	10	1.40E-04	
	GO:0006952~defense response	14	3.91E-04	
	GO:0007160~cell-matrix adhesion	6	4.20E-04	
	GO:0005576~extracellular region	51	1.88E-16	
	GO:0044421~extracellular region part	35	5.59E-15	
	GO:0031012~extracellular matrix	22	9.60E-14	
CC	GO:0005578~proteinaceous extracellular matrix	21	2.46E-13	
	GO:0044420~extracellular matrix part	11	1.80E-08	
	GO:0005615~extracellular space	21	1.60E-07	
	GO:0005581~collagen	7	2.22E-07	
	GO:0005583~fibrillar collagen	4	8.87E-05	
	GO:0005604~basement membrane	6	3.09E-04	
	GO:0005584~collagen type I	2	0.015121	
	GO:0005201~extracellular matrix structural constituent	9	1.33E-07	
	GO:0008009~chemokine activity	6	1.57E-05	
	GO:0042379~chemokine receptor binding	6	2.15E-05	
MF	GO:0048407~platelet-derived growth factor binding	4	5.11E-05	
	GO:0005125~cytokine activity	9	6.07E-05	
	GO:0001871~pattern binding	8	9.45E-05	
	GO:0030247~polysaccharide binding	8	9.45E-05	
	GO:0005198~structural molecule activity	14	4.07E-04	
	GO:0005539~glycosaminoglycan binding	7	4.18E-04	
	GO:0030246~carbohydrate binding	10	7.61E-04	
	KEGG_pathway	hsa04512:ECM-receptor interaction	7	9.37E-05
		hsa04510:Focal adhesion	7	8.86E-03
		hsa04062:Chemokine signaling pathway	6	2.58E-02
	hsa00591:Linoleic acid metabolism	3	2.61E-02	

BP = biological process, CC = cellular component, MF = molecular function.

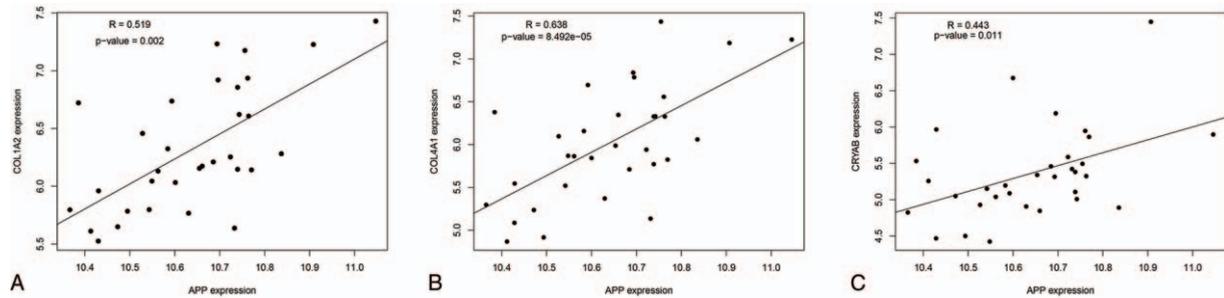
hepatocyte-stellate cell interactions.<sup>[25]</sup> In the present study, the selected DEGs were particularly associated with processes such as cell adhesion and inflammatory response. Thus, we speculated that these biological processes such as cell adhesion, fibrosis, and inflammatory response might play important roles in the development of NAFLD.

*COL1A1* has been demonstrated to be upregulated in liver fibrosis by the activation of stellate cells and the progression of liver fibrosis.<sup>[26,27]</sup> Zhao et al<sup>[28]</sup> showed that *COL1A1* gene polymorphism was associated with liver fibrogenesis, since the T allele at 1997 of *COL1A1* was crucial to the increased transcriptional activity. *COL1A2* is an independent predictor

of survival in diseases.<sup>[29]</sup> Many diseases that are inherited in an autosomal dominant fashion are caused by mutations in the *COL1A1/COL1A2* genes.<sup>[30,31]</sup> Because *COL1A1* and *COL1A2* were revealed to be DEGs associated with NAFLD in this study, we speculated that not only *COL1A2*, but also the dysregulation of *COL1A1/COL1A2* participated in the progression of NAFLD. Furthermore, the present study based on GO and KEGG pathway analyses showed that *COL1A1* might be related to cell adhesion and inflammation; thus, *COL1A1/COL1A2* might play important roles in diseases by regulating genes associated with cell adhesion and inflammation. In addition, *COL1A1* was confirmed to be upregulated in NASH samples



**Figure 3.** The top 5 subnetworks with the highest Z scores. Red spots represent differentially expressed genes, and blue or yellow spots represent nondifferentially expressed genes.



**Figure 4.** Pearson correlation analysis between *APP* and the target-DEGs which they interact. A, the correlation between *APP* and *COL1A2* ( $R=0.519$ ,  $P=.002$ ); B, the correlation between *APP* and *COL4A1* ( $R=0.638$ ,  $P<.001$ ); C, the correlation between *APP* and *CRYAB* ( $R=0.443$ ,  $P=.011$ ).  $R$  represents Pearson correlation coefficient,  $R>0$  indicates a positive correlation;  $R<0$  indicates a negative correlation; if  $|R|$  is further away from 0, the correlation is stronger. The  $P$  value indicates the significant difference. *APP* = amyloid-beta precursor protein, *DEG* = differentially expressed gene.

**Table 3**  
The candidate genes and their interacted genes.

Candidate genes	Z score	Target-DEGs
UBQLN4	1.68	SPP1, GPX7, OAT . . .
APP	1.56	COL1A2, CRYAB, COL4A1 . . .
SHBG	1.43	COL1A2, MT1F, PTGDS . . .
CTNNB1	1.39	RXRA, TCF4, H1F1A . . .
COL1A1	1.38	TXN, FGF7, HTRA1 . . .

compared with the level in normal controls. *SHBG*, a glycoprotein expressed predominantly in the hepatocytes, regulates the transport of sex steroid hormones in the bloodstream to their target tissues.<sup>[32]</sup> As one of the circulation factors released from fatty liver, *SHBG* was reported to be directly involved in the pathogenesis of local and systemic inflammation, and peripheral as well as hepatic insulin resistance.<sup>[33]</sup> In the present study, *APP* and *SHBG* were shown to be connected by *COL1A2*, which further indicates that the *APP* and *SHBG* genes have a close relationship in the process of NAFLD development.

Ubiquilins (*UBQLN*), a family of ubiquitin-binding proteins, are involved in several protein degradation pathways and have

been implicated in various diseases.<sup>[34]</sup> Previous studies indicated that the members of this family-mediated degradation of misfolded proteins and they were implicated in a number of pathological and physiological conditions.<sup>[34-36]</sup> For example, Matsuda et al<sup>[37]</sup> indicated that *UBQLN4* was highly expressed in organs such as liver. Unfortunately, to date, few details of the relationship between *UBQLN4* and NAFLD have been clarified. In this study, *UBQLN4* was revealed to be a hub gene in the PPI network. Thus, we speculated that *UBQLN4* might participate in the progression of via the degradation of misfolded proteins. *CTNNB1* is located on the short arm of chromosome 3, as determined by in situ fluorescence analysis,<sup>[38]</sup> and has been reported to be commonly involved in benign liver tumorigenesis.<sup>[39]</sup> Mutation of the *CTNNB1* gene mutation is also an important indicator of prognosis in primary sporadic aggressive fibromatosis.<sup>[40]</sup> In addition, *CTNNB1* was also found to often display point mutations resulting in loss of function in a range of cancers, with the notable exception in hepatocellular carcinoma,<sup>[41]</sup> revealing that the expression level of *CTNNB1* might have great significance for liver function. Kubota et al<sup>[42]</sup> indicated that mutational analysis of *CTNNB1* for predicting disease such as solid-pseudopapillary neoplasm is feasible. In the present study, *CTNNB1* investigated given its

**Table 4**  
The results of verification data including the top 10 differentially expressed genes and the 5 candidate genes.

Genes	LogFC	P	Original	
Top 10 DEGs	<i>CXCL6</i>	0.153	.323	Up
	<i>EPCAM</i>	0.532	.204	Up
	<b><i>AKR1B10</i></b>	<b>1.693</b>	<b>.020</b>	<b>Up</b>
	<i>CD24</i>	0.495	.189	Up
	<i>THBS2</i>	0.453	.141	Up
	<b><i>CYP2C19</i></b>	<b>-0.674</b>	<b>.030</b>	<b>Down</b>
	<i>DHRS2</i>	-0.271	.405	Down
	<i>MT1M</i>	-0.369	.321	Down
	<i>RPS4Y1</i>	-0.340	.656	Down
	<i>GNMT</i>	-0.671	.055	Down
The 5 candidate genes	<b><i>COL1A1</i></b>	<b>0.713</b>	<b>.005</b>	<b>Up</b>
	<i>UBQLN4</i>	0.113	.241	/
	<i>APP</i>	0.056	.375	/
	<b><i>SHBG</i></b>	<b>-1.008</b>	<b>.001</b>	<b>/</b>
	<i>CTNNB1</i>	0.014	.833	/

Original indicates the upregulation and downregulation of DEGs in the manuscript. DEG = differentially expressed gene, FC = fold change. Bold values signify the  $P$  value in the table.

status here as a hub protein. The findings suggest that *CTNNB1* is useful as a target gene for further investigation of NAFLD. The analyses demonstrated the detailed action of these 2 candidate genes in NAFLD, and also showed that the interaction between them is rare. Therefore, further experiments are needed to explain the functions of *CTNNB1* and *UBQLN4* in NAFLD.

Despite all these results, there are some limitations in the present study. First, verification experiments were not performed because of the lack of sufficient liver tissue samples. Second, we did not obtain the gene expression data, including for normal liver tissue, severe NAFLD tissue, and mild NAFLD tissue, categorized according to fibrosis stage in the NCBI database. Therefore, more experiments and data are required to verify the above results.

In conclusion, cell adhesion, fibrosis, and inflammatory response might play crucial roles in severe NAFLD. Similarly, 5 candidate genes, namely *UBQLN4*, *APP*, *CTNNB1*, *SHBG*, and *COL1A1* might be involved in the development of NAFLD, and might be considered as potential markers for predicting the development of NAFLD. The findings of this study might explain the development of NAFLD and reveal new target genes for treating NAFLD.

## References

- Erickson SK. Nonalcoholic fatty liver disease. *J Lipid Res* 2009;50 (suppl):S412–6.
- Smith BW, Adams LA. Non-alcoholic fatty liver disease. *Crit Rev Clin Lab Sci* 2011;48:97–113.
- Grant LM, Lisker-Melman M. Nonalcoholic fatty liver disease. *Ann Hepatol* 2004;3:93–9.
- Northup PG, Argo CK, Shah N, Caldwell SH. Hypercoagulation and thrombophilia in nonalcoholic fatty liver disease: mechanisms, human evidence, therapeutic implications, and preventive implications. Paper presented at: Thieme Medical Publishers, 2012, 32:39–48.
- Van't Veer LJ, Dai H, Van De Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530–6.
- Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease. *Nature* 2008;452:423–8.
- Kirpich IA, Gobejishvili LN, Homme MB, et al. Integrated hepatic transcriptome and proteome analysis of mice with high-fat diet-induced nonalcoholic fatty liver disease. *J Nutr Biochem* 2011;22:38–45.
- Lake AD, Novak P, Fisher CD, et al. Analysis of global and absorption, distribution, metabolism, and elimination gene expression in the progressive stages of human nonalcoholic fatty liver disease. *Drug Metab Dispos* 2011;39:1954–60.
- Moylan CA, Pang H, Dellinger A, et al. Hepatic gene expression profiles differentiate pre-symptomatic patients with mild versus severe nonalcoholic fatty liver disease. *Hepatology* 2014;59:471–82.
- Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003;4:249–64.
- Gautier L, Cope L, Bolstad BM, et al. Affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 2004;20:307–15.
- Diboun I, Wernisch L, Orengo CA, et al. Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. *BMC Genomics* 2006;7:252.
- Dudoit S, Gilbert HN, Van Der Laan MJ. Resampling-based empirical bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: focus on the false discovery rate and simulation study. *Biomet J* 2008;50:716–44.
- Alvord G, Roayaei J, Stephens R, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* 2007;8:R183.
- Prasad TK, Goel R, Kandasamy K, et al. Human protein reference database—2009 update. *Nucleic Acids Res* 2009;37(suppl 1):D767–72.
- Wu C, Zhu J, Zhang X. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. *BMC Bioinformatics* 2012;13:182.
- Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;102:15545–50.
- Mazzella N, Ricciardi LM, Mazzotti A, et al. The role of medications for the management of patients with NAFLD. *Clin Liver Dis* 2014;18:73–89.
- Greco D, Kotronen A, Westerbacka J, et al. Gene expression in human NAFLD. *Am J Physiol Gastrointest Liver Physiol* 2008;294:G1281–7.
- Mikolasevic I, Lukenda V, Racki S, et al. Nonalcoholic fatty liver disease (NAFLD)—a new factor that interplays between inflammation, malnutrition, and atherosclerosis in elderly hemodialysis patients. *Clin Interv Aging* 2014;9:1295–303.
- Wu J, Zhang H, Zheng H, et al. Hepatic inflammation scores correlate with common carotid intima-media thickness in rats with NAFLD induced by a high-fat diet. *BMC Vet Res* 2014;10:162.
- Browning JD, Horton JD. Molecular mediators of hepatic steatosis and liver injury. *J Clin Invest* 2004;114:147–52.
- Tilg H, Moschen AR. Evolution of inflammation in nonalcoholic fatty liver disease: the multiple parallel hits hypothesis. *Hepatology* 2010;52:1836–46.
- Tsuchiya A, Lu WY, Weinhold B, et al. Polysialic acid/neural cell adhesion molecule modulates the formation of ductular reactions in liver injury. *Hepatology* 2014;60:1727–40.
- Sugiyama Y, Koike T, Shiojiri N. Developmental changes of cell adhesion molecule expression in the fetal mouse liver. *Anat Rec (Hoboken)* 2010;293:1698–710.
- Asselah T, Bièche I, Laurendeau I, et al. Liver gene expression signature of mild fibrosis in patients with chronic hepatitis C. *Gastroenterology* 2005;129:2064–75.
- Iizuka M, Ogawa T, Enomoto M, et al. Induction of microRNA-214-5p in human and rodent liver fibrosis. *Fibrogenesis Tissue Repair* 2012; 5:12.
- Zhao YP, Wang H, Fang M, et al. Study of the association between polymorphisms of the COL1A1 gene and HBV-related liver cirrhosis in Chinese patients. *Dig Dis Sci* 2009;54:369–76.
- Misawa K, Kanazawa T, Misawa Y, et al. Hypermethylation of collagen alpha2 (I) gene (COL1A2) is an independent predictor of survival in head and neck cancer. *Cancer Biomark* 2011;10:135–44.
- Wang W, Wu Q, Cao L, et al. Mutation analysis of COL1A1 and COL1A2 in fetuses with osteogenesis imperfecta type II/III. *Gynecol Obstet Invest* 2015;79: sgmppl =-109.
- Stephen J, Shukla A, Dalal A, et al. Mutation spectrum of COL1A1 and COL1A2 genes in Indian patients with osteogenesis imperfecta. *Am J Med Genet A* 2014;164A:1482–9.
- Flechtner-Mors M, Schick A, Oeztuerk S, et al. Associations of fatty liver disease and other factors affecting serum SHBG concentrations: a population based study on 1657 subjects. *Horm Metab Res* 2013;46: 287–93.
- Stefan N, Häring H-U. The metabolically benign and malignant fatty liver. *Diabetes* 2011;60:2011–7.
- Tsukamoto S, Shimada K, Honoki K, et al. Ubiquitin 2 enhances osteosarcoma progression through resistance to hypoxic stress. *Oncol Rep* 2015;33:1799–806.
- El Ayadi A, Stieren ES, Barral JM, et al. Ubiquitin-1 and protein quality control in Alzheimer disease. *Prion* 2013;7:164–9.
- Daoud H, Rouleau GA. A role for ubiquitin 2 mutations in neurodegeneration. *Nature Rev Neurol* 2011;7:599–600.
- Matsuda M, Koide T, Yorihuzi T, et al. Molecular cloning of a novel ubiquitin-like protein, UBIN, that binds to ER targeting signal sequences. *Biochem Biophys Res Commun* 2001;280:535–40.
- Kraus C, Liehr T, Hülsken J, et al. Localization of the human (-catenin gene (CTNNB1) to 3p21: a region implicated in tumor development. *Genomics* 1994;23:272–4.
- Nault JC, Fabre M, Couchy G, et al. GNAS-activating mutations define a rare subgroup of inflammatory liver tumors characterized by STAT3 activation. *J Hepatol* 2012;56:184–91.
- Van Broekhoven DL, Verhoef C, Grunhagen DJ, et al. Prognostic value of CTNNB1 gene mutation in primary sporadic aggressive fibromatosis. *Ann Surg Oncol* 2015;22:1464–70.
- Ozen C, Yildiz G, Dagcan AT, et al. Genetics and epigenetics of liver cancer. *New Biotechnol* 2013;30:381–4.
- Kubota Y, Kawakami H, Natsuizaka M, et al. CTNNB1 mutational analysis of solid-pseudopapillary neoplasms of the pancreas using endoscopic ultrasound-guided fine-needle aspiration and next-generation deep sequencing. *J Gastroenterol* 2015;50:203–10.