DR. WEICHEN ZHOU (Orcid ID : 0000-0003-4755-1072)

DR. YIDA PAN (Orcid ID : 0000-0002-1173-1074)

Received Date : 29-Nov-2016

Revised Date : 10-Mar-2017

Accepted Date : 14-Mar-2017

Article type : Original Articles

Handling Editor: Mario Mondelli

Predictive model for inflammation grades of chronic hepatitis B: large-scale analysis of clinical parameters and gene expressions

Weichen Zhou[2,3]*, Yanyun Ma[2]*, Jun Zhang[1], Jingyi Hu[1,2], Menghan Zhang[2], Yi Wang[2], Yi Li[2], Lijun Wu[1], Yida Pan[1], Yitong Zhang[1,2], Xiaonan Zhang[5], Xinxin Zhang[9], Zhanqing Zhang[5], Jiming Zhang[6], Hai Li[7], Lungen Lu[8], Li Jin[2], Jiucun Wang[2#], Zhenghong Yuan[4,5#], Jie Liu[1, 4#]

[1] Department of Digestive Diseases of Huashan Hospital, Collaborative Innovation Center for Genetics and Development, Fudan University, 12 Middle Wulumuqi Road, Shanghai 200040, China.

[2] State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, 2005 Songhu Road, Shanghai 200438, China.

[3] Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA.

[4] Key Laboratory of Medical Molecular Virology of MOE/MOH, Institutes of

Biomedical Sciences and Department of Immunology, Shanghai Medical School, Fudan University, 138 Yixueyuan Road, Shanghai 200032, China.

[5] Shanghai Public Health Clinical Center, Fudan University, 2901 Caolang Road, Shanghai 201508, China.

[6] Department of Infectious Diseases, Huashan Hospital, Fudan University, 12 Middle Wulumuqi Road, Shanghai 200040, China.

[7] Department of Gastroenterology, Renji Hospital, Shanghai Jiaotong University School of Medicine, 145 Middle Shandong Road, Shanghai 200001, China.

[8] Department of Gastroenterology, Shanghai General Hospital, Shanghai Jiaotong University School of Medicine, 100 Haining Road, Shanghai 20080, China.

[9] Department of Infectious Diseases, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, 197 Ruijin Er Road, Shanghai 200025, China.

* These authors have contributed equally as joint first authors.

[#] These authors have contributed equally as senior authors.

Correspondence

Jiucun Wang, Ph.D.,
School of Life Sciences, Fudan University, 2005 Songhu Road, Shanghai 200438, China, E-mail: jcwang@fudan.edu.cn

Jie Liu, M.D. Ph.D.,
Department of Digestive Diseases, Huashan Hospital, Fudan University, 12 Middle Wulumuqi Road, Shanghai 200040, China, E-mail: jieliu@fudan.edu.cn

Zhenghong Yuan, M.D. Ph.D.,
Shanghai Public Health Clinical Center, Fudan University, 2901 Caolang Road, Shanghai 201508, China, E-mail: zhyuan@shaphc.org

List of abbreviations

HBV, hepatitis B virus; HCC, hepatocellular carcinoma; CHB, chronic hepatitis B; ALT, alanine amino transaminase; AST, aspartate amino transaminase; GO, gene ontology; LARS, least angle regression; PCA, principal component analysis; RF, random forest; SVM, support vector machine; KNN, K-nearest neighbor; AUC, area under the ROC curve; CI, confidence interval, MDA, mean decrease accuracy.

Conflict of interest

None.

Financial support

This work was supported by grants from the National Natural Science Foundation of China (31521003, 91129702 and 81125001), the Ministry of Science and Technology of China (2006AA02A411), the Major National Science and Technology Program of China (2008ZX10002-002) and the 111 Project (B13016) from Ministry of Education (MOE). Computational support was provided by the High-End Computing Center located at Fudan University.

Word count: 4934

Number of figures and tables: 7

Abstract

Background

Liver biopsy is the gold standard to assess pathological features (e.g. inflammation grades) for hepatitis B virus infected patients, although it's invasive and traumatic; meanwhile, several gene profiles of chronic hepatitis B (CHB) have been separately described in relatively small HBV-infected samples. We aimed to analyze correlations among inflammation grades, gene expressions and clinical parameters (serum alanine amino transaminase, aspartate amino transaminase, and HBV-DNA) in large-scale CHB samples, and to predict inflammation grades by using clinical parameters and/or gene expressions.

Methods

We analyzed gene expressions with three clinical parameters in 122 CHB samples by an improved regression model. Principal component analysis and machine learning methods including Random Forest, K-Nearest Neighbor, and Support Vector Machine, were used for analysis and further diagnosis models. Six normal samples were conducted to validate the predictive model.

Results

Significant genes related to clinical parameters were found enriching in the immune system, interferon-stimulated, regulation of cytokine production, anti-apoptosis and etc. A panel of these genes with clinical parameters can effectively predict binary classifications of inflammation grade (AUC: 0.88, 95% CI: 0.77-0.93), validated by normal samples. A panel with only clinical parameters was also valuable (AUC: 0.78, 95% CI: 0.65-0.86), indicating that liquid biopsy method for detecting the pathology of CHB is possible.

Conclusions

This is the first study to systematically elucidate the relationships among gene expressions, clinical parameters and pathological inflammation grades in CHB, and to build models predicting inflammation grades by gene expressions and/or clinical parameters as well.

Abstract word count

244

Keyword

Clinical predictive model; Inflammation grades; Gene expressions; HBV infection.

Key points

1. Correlations among inflammation grades, clinical parameters and gene expressions in CHB patients are only partially enclosed; meanwhile, liquid biopsy prediction of inflammation grades is still unexplored.
2. A list of significant genes correlated with clinical parameters was revealed in several functions and pathways from large-scale samples.

3. A panel of genes and clinical parameters can effectively predict binary classifications of inflammation grade (AUC: 0.88, 95% CI: 0.77-0.93).
4. A panel with only clinical parameters also has a power (AUC: 0.78, 95% CI: 0.65-0.86) to predict inflammation, which can be further used in the liquid biopsy method for detecting the pathology of CHB.

Introduction

In clinic, liver biopsy is a gold standard to directly assess pathological features (e.g. the inflammation level G) and determine prognosis for HBV-infected patients [1]. But it is invasive and traumatic. Serum parameters (e.g. Alanine amino transaminase (ALT) and aspartate amino transaminase (AST)) are utilized to access the damage of liver and HBV viral infection [1, 2]. In certain cases, these three clinical parameters are necessities for the decision of following appropriate therapy [1, 3].

Microarray is a well-established and widely used technology, which can effectively provide an image of gene expressions [4]. Researchers have only identified several gene profiles in relative small number of HBV-infected patients [5, 6], some of which have investigated gene expressions with single clinical parameter, e.g. ALT or HBV expression [7]. There are few studies systematically combining clinical parameters, gene expressions and pathological inflammation levels to acquire a comprehensive view of CHB, not to mention in a large-scale sample size. Other researchers began to explore a liquid biopsy method to assess liver function based on single clinical parameter or in other liver disease (e.g. chronic hepatitis C) [8-10]. There is barely any effective predictive model for inflammation grades of CHB right now, and liquid biopsy method is even more elusive.

In this paper, we carried out the first study combining three clinical parameters (serum ALT, AST and HBV-DNA), gene microarray data and inflammation grades of CHB. We determined a batch of gene expressions significantly correlated with these clinical continuous parameters, and uncovered pathways and networks related to CHB by comprehensive bioinformatics analyses. More importantly, it is the first time to construct an effective model to diagnose and predict inflammation grades in

HBV-infected patients by using these significant gene expressions and/or three clinical parameters, which can help to develop liquid biopsy method for detecting the pathology of CHB.

Materials and methods

Collection of samples and clinical data

This study was approved by the ethics committees of Fudan University (Shanghai, China). All subjects provided written informed consents according to institutional guidelines. A standardized procedure was established for preservation of liver biopsy sample and RNA extract method. Briefly, after the biopsy was taken, it was quickly submerged in RNAlater, which is an effective stabilizer of tissue RNA, and stored at 4 degree. The sample was later shipped to a biobank and stored at -80 degree for long-term storage. The workflow of microarray analysis requires rigorous quality control in RNA integrity. Only RNA samples extracted with RIN>=7.0 and 28S/18S>0.7 were processed further. Four sampling sites must completely follow this standardized procedure, and patients must have same chronic hepatitis B diagnostic criteria, including HBV persistent infection, HBsAg positive, liver biopsy showed varying degrees of inflammatory necrosis. There is no distribution bias of liver disease grades among these sites. Normal samples were obtained and validated by liver biopsy with non-HBV-infected. Hepatitis samples were obtained by liver biopsy and conducted blood sampling. The samples with HCV infection or metabolic liver injury (e.g. fatty liver, chronic alcoholic hepatitis, etc.) were excluded. After extraction of cRNA, liver tissues were processed by GeneChip Human Genome U133 Plus 2.0 Arrays.

Three clinical parameters were measured in blood. The samples with inexact values (>5*10^{7} or <500) of HBV-DNA were excluded. The activity of inflammation was measured and confirmed by pathological examination of liver biopsies from two experienced pathologists separately. They were characterized into five grades (G0-4) following the pathological analysis of the biopsies [1, 11, 12].

Data processing and bioinformatics analysis

CEL files were performed by Affymetrix Expression Console. Probe set signals were normalized and summarized by the robust multi-array average algorithm [13] to adjust different batch effects. All samples passed quality control. The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus and are accessible through accession number GSE83148.

We normalized the values of parameters, by $\log_{10}$ transformation of HBV-DNA values and min-max normalization of values of ALT and AST. Subsequently, the Least Angle Regression (LARS) algorithm (package ***Lars*** [14]) was performed to obtain significant probes that correlated with ALT, AST and HBV-DNA, respectively. We only used the expression data of the samples with valid information. Later, significant probe-level sets were converted to gene-level by using annotation file.

Pathway and gene ontology (GO) enrichment were performed by using the Database for Annotation, Visualization and Integrated Discovery (DAVID, (http://david.abcc.ncifcrf.gov/). Cytoscape was applied to build gene networks with geneMANIA plugin [15].

Principal component analysis (PCA) and linear regression

PAST (http://folk.uio.no/ohammer/past) was used to carry out PCA and linear regression to investigate expressions of significant genes correlated with ALT and AST. The loading coefficients of significant genes were obtained according to different PCs. The scatter values of HBV-infected samples in each PC were transformed from the expression values of each sample by loading coefficients. In the linear regression (Fig. 4D) for PC3, inflammation grades were considered as numerical variables 0-4 and PC3 scatter values were considered as dependent variables, with box-plots and fitted lines plotted.

Binary classifications of inflammation grades and predictive models by machine-learning methods

G0 and G1 were considered as mild inflammation, and G 2-4 as moderate or severe inflammation [1, 12]. Based on these, binary classifications (mild or exacerbated) of G were introduced. The expressions of significant genes, the above three clinical parameters and information of sex and age were then utilized to predict these binary classifications of G.

Based on the G classifications, feature selections were conducted by Random Forest (RF) among significant genes that correlate with either ALT and HBV-DNA, ALT and AST, or AST and HBV-DNA. A gene panel was obtained. Here we used K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and RF to build predictive models for three modules. In general, these are all machine-learning methods for classification and regression [16, 17]. KNN is a non-parametric algorithm, assigning weights to the contributions of neighbors on the basis of the basic principle of majority voting; SVM is a non-probabilistic binary linear classifier, assigning new examples to one category or the other based on a set of training examples; And RF constructs decision trees by training sets, and outputs the class either is the mode of classification or regression of the individual trees. KNN was implemented in Matlab. SVM (Package ***e1071***) and RF (Package ***randomForest***) were run by R. Three modules for predictive models were separately built: Module 1 (with information of three clinical parameters and adjustment of sex and age), Module 2 (with genes panel obtained by feature selections), and Module 3 (with all information of selected genes panel, clinical parameters, and adjustment of sex and age). All modules with three predictive methods were performed by five-fold cross-validation to avoid over-fitting. ROC curves were plotted (package ***ROCR***) and the area under the ROC curve (AUC) was calculated (package ***pROC***) with 95% confidence interval (CI). All normal samples were conducted as validations by Module 2 with RF. All packages from R project can be downloaded from Bioconductor (http://www.bioconductor.org).

Results

Distribution of HBV-infected patients by clinical parameters

One hundred and twenty-two liver hepatitis tissues infected with HBV were obtained. Of these (Fig. 1 & Table S1), 90 had exact quantitative HBV-DNA values (ranging from 603 to 1*E9), and 105 samples had valid quantitative ALT (normal values: 7-40, and abnormal values: 41-1554.3) and AST values (normal values: 10-35, and abnormal values: 36-706.1). One hundred nineteen samples were portrayed by G (from G0 to G4), with 34 G0 samples, 33 G1 samples, 31 G2 samples, 15 G3 samples and 6 G4 samples. Six normal samples were all identified as G0.

Analysis workflow

We devised a framework for analyzing three clinical parameters, gene microarray data and inflammation grades of CHB (Fig. 2A). After normalization, we manipulated LARS into 90 samples with exact values of HBV-DNA to analyze the significances correlated with HBV-DNA and 105 samples with exact values of ALT or AST to analyze the probes significantly correlated with ALT or AST. After annotation, we finally identified 80 significant genes correlated with serum HBV-DNA, including 48 positive and 32 negative, 96 significant genes (53 positive and 43 negative) correlated with serum ALT, and 92 significant ones (45 positive and 47 negative) correlated with serum AST, respectively (Fig. 2B). Two genes, *IGHA1* and *ZNF75A*, significantly correlated with both values of serum HBV-DNA and ALT. Sixteen others significantly correlated with both values of serum ALT and AST (Fig. 2B & Table S2).

Significant gene pathway, GO and gene networks

A gene pathway consists of a group of interacting components, acting in concert to perform specific biological tasks [18]. Utilizing the DAVID, we identified 7 significant pathways. For HBV-DNA, hTert transcriptional regulation (p-value = 2.42*E-02), lectin-induced complement pathway (p-value = 4.11*E-02), and classical complement pathway (p-value = 4.11*E-02) were identified. For ALT, B cell activation (p-value = 2.51*E-02) was highlighted. For AST, B cell activation pathway was also significantly (p-value = 4.97*E-02) enriched, so was pathways in cancer (p-value = 7.61*E-04) and pathway of melanoma (p-value = 4.69*E-02) (Table 1). Significant GO terms are listed in Table S2, which are mostly enriched in immune response (GO: 0006955), apoptosis (GO: 0042981, GO: 0043066, GO: 0060548), positive regulation of cytokine production (GO: 0001819) and etc.

For gene networks, geneMANIA can search large, publicly available biological datasets to illuminate interactions [15]. Genes were linked by different color lines, referring to different interactions (Fig. 3): co-expression, co-localization, physical interaction, and shared protein domains. The number of lines represents the importance of the gene in the network. In the ALT-correlated network, 11 significant genes had more than 3 lines, e.g. *FLI1*, *STK17B*, *ANK2* and etc. In the AST-correlated network, there are 3 sub-networks, and *DGUOK* is an important one which is not in the significant gene list but interacts closely with others. In the HBV-DNA-correlated network, *Sept10* and *SLC9A3R2* are at the core of two sub-networks.

PCA reveals gene expressions correlated with three biological categories: clinical parameters, gene functions and inflammation grades

PCA can be used to determine key variables in gene expression data by using an orthogonal transformation [19]. By applying PCA to 16 significant gene expressions that correlated with ALT and AST, we obtained three highlighted PCs , each of which can explain more than 10% of variance: the first (PC1) explained 19.1% of variance (eigenvalue = 3.052), the second (PC2) explained 13.8% of variance (eigenvalue = 2.202), and the third (PC3) explained 10.7% of variance (eigenvalue = 1.705). The more portion of variance it can explain, the more important one component is. We thereby were figuring out biological meanings behind these corresponding components.

In Fig. 4A, genes with positive loading coefficients in PC1 are the same ones that positively correlate with serum ALT and AST, and the others with negative loading coefficients have negative correlations. Therefore, PC1 mainly represents the correlative effects of serum ALT and AST.

According to loading coefficients in PC2 (Fig. 4B), *DLX3*, *PRDX2* and *YBX1* are enriched in the GO term regarding regulation of transcription with positive coefficients (Table. S2). *TTLL4*, *TTLL7* and *DCTN4* are enriched in microtubules, and *IGF1R* and *NRXN1* are related to axon-genesis, all of which represent significant genes with negative coefficients correlated with the function of cell cytoskeleton. Therefore, PC2 mainly represents the functional differentiation of genes as serum ALT and AST levels are changing in the HBV-infected patients.

For PC3 (Fig. 4C), we carried out a linear regression analysis between inflammation grades of CHB and scatter values of each sample generated by loading coefficients in PC3, and found a significant linear correlation between them (p-value = 6.69*E-03) (Fig. 4D). Therefore, PC3 mainly explains a linear correlation between inflammation grades and gene expressions.

Random Forest model efficiently diagnoses the inflammation grades in CHB

According to PCA, there is a correlation between inflammation grades of CHB and gene expression in PC3. Inspired from this, we further constructed diagnosis models to predict inflammation grades based on the 18 sharing significant genes (two correlated with ALT and HBV-DNA and 16 correlated with ALT and AST).

A binary classification of G was introduced based on the categories of all inflammation grades [1, 3]. By utilizing RF, a gene panel with nine genes (*DLX3*, *ALPK1*, *YBX1*, *ZNF75A*, *SPP2*, *TTLL4*, *TTLL7*, *AGAP3*, and *DCTN4*) among 18 significant ones for binary classification of G were selected. RF, SVM, and KNN were further used to construct predictive models, with the involvement of three clinical phenotypic parameters and adjustment of sex and age. To remove the impact of missing data on results, we only utilized 81 samples with valid information of HBV-DNA, ALT, AST, sex and age (Table S1). Sensitivity, specificity and classification accuracy of each method are shown in Table 2, and ROC curves are plotted in Fig. 5. All values are averaged in fivefold validations and values of AUC are shown with 95% CI, according to different predictive modules and models.

Using genes panel, Module 2 generally performed better than Module 1 by using clinical parameters, based on the results of SVM (0.749 vs 0.734), KNN (0.723 vs 0.729) and RF (0.801 vs 0.784) (Table 2). Notably, when combining all information (Module 3), the predictive power of KNN (0.806, 95% CI: 0.711-0.898) and RF (0.880, 95% CI: 0.771-0.933) increased dramatically. More importantly, even though the powers of Module 1 are relatively low (RF: 0.784, SVM: 0.734 and KNN: 0.729), it is still an improvable model by only using clinical parameters to predict inflammation grades, indicating that liquid biopsy method for detecting the pathology of CHB is possible. Lastly, we carried out validations by conducting Module 2 of RF method on six normal samples. All six samples were predicted as mild inflammation (G0 or G1), with predicting probability up to 0.827 ± 0.037. In conclusion, RF is the most powerful model for the diagnosis of inflammation grades of CHB when combining expressions of nine genes, three clinical parameters, sex and age.

Discussion

In this study, we considered clinical parameters as continuous variables, and analyzed gene expressions by an advanced regression algorithm (LARS), which is more efficient to obtain significant genes than regular linear regression method [14]. Besides, in CHB samples, part of them have a normal level of ALT, AST, or inflammation grade (G0), which can be considered as baseline values in LARS analysis, PCA, and predictive models, as healthy controls in the regular case-control study. Moreover, to maximize the utilization of all information and eliminate the impact of missing data, we discarded samples with missing data in separated steps.

Several genes and pathways correlated with HBV infection and immune response were discovered. *TRD*, *CD84*, *HLA-DRB4* and B cell activation pathway (Table 1) with genes *IGHG3*, *POU2F2*, and *IGHM* positively correlated with ALT values, suggesting that a proliferation of immune cells and regeneration of liver cells occurs as an increase of serum ALT after HBV infection. Intriguingly, though the gene expression profiles came from a mix of different types of cells, these inflammation-related genes and pathway indicate the inflammatory cells mixing with hepatic cells may have contributed to the overall gene expression patterns as HBV infection getting worse. In the core of ALT-correlated network (Fig. 3A), *STK17B* with positive correlation was reported to form a novel signaling module which controls calcium homeostasis following T cell activation [20]. Another core gene *PRF1* was reported as an important role in liver cell injury after HBV infection [21] and

HBV-DNA cleanup [22]. Additionally, in the HBV-DNA network (Fig. 3C), a core significant gene *SLC9A3R2*, co-expressing with *ACACB* and *SP1*, is a membrane transporter of HBV and HDV entry [23]. The AST positive specific-related gene *CD58* (Fig. 3B) was also found related to the microtubule and immune response system and significantly increased with the severity of HBV infection [24]. Intriguingly, by utilizing interferon-stimulated genes datasets (Interferome: http://interferome.its.monash.edu.au), we found seven interferon-stimulated genes that are significantly correlated with serum HBV-DNA. *MKX*, *TSNARE1* and *EFR3A* are positively, and *ACSF3*, *H2AFJ*, *XRN1* and *ZNF677* are negatively correlated with the increasing value of serum HBV-DNA in infected hepatocytes.

In the AST-related pathway, pathways related to cancer were found (Table 1), supported by the fact that an increasing serum AST often indicates a severe progression of liver cell damage. *SP1* and *WT1*, clustered in the pathway of *hTert* transcriptional regulation (Table 1), are reported to significantly correlate with *HBx* expression [25] and HCC development [26-28]. Interestingly, among 80 significant genes correlated with HBV-DNA, 8 (10%) of them have been reported to correlate with HCC [26-30] or other cancers [22, 31, 32], indicating that they may also play important roles in the progression from HBV-induced inflammation to HCC. Gene *IGHA1*, which shares significant positive correlation with HBV-DNA and ALT, is also reported involving in gastric tumorigenesis [33].

Notably, in PCA of expression data of 16 significant genes, five principal components (PCs) had an eigenvalue more than 1 and explained 58.7% of variance in total. The top three PCs can reveal specific biological insights and explain 43.5% of variance. In the feature selection of predictive model, nine genes were selected based on their Mean Decrease Accuracy (MDA): *YBX1* (MDA=47.8), *ALPK1* (MDA=28.0), *ZNF75A* (MDA=18.2), *SPP2* (MDA=13.2), *DCTN4* (MDA=12.3), *AGAP3* (MDA=7.95), *DLX3* (MDA=6.86), *TTLL4* (MDA=4.68) and *TTLL7* (MDA=2.57). All genes above were mainly related to protein phosphorylation, transcription functions and the major histocompatibility complex. Five of them are related to transcription, indicating the importance between transcription and inflammation grades.

For predictive models, three modules were conducted separately to find the most appropriate model. We suggest RF as a machine-learning black box to aid in prediction and diagnosis for binary classification of inflammation grade of CHB, which has an effective power (0.880) with the help of three indispensable clinical parameters and nine genes. In previous studies, there established a predictive model by Xu et al. [9] using red blood cell distribution width value, ALT and other blood parameters (albumin and platelet) from 446 patients to predict CHB inflammation with highest AUC of 0.765. While, we have had a relative higher power of AUC of 0.784 (Random Forest, AUC: 0.784, 95% CI: 0.65-0.86) by using the three clinical parameters from 81 samples to predict inflammation grades in the present study. More samples and studies are highly required based on our models, which may substitute liver biopsy by liquid biopsy method into a practical clinic protocol to characterize the pathological inflammation.

In conclusion, we carried out the first analysis of large-scale HBV-infected samples by combining gene expressions data and three clinical parameters (ALT, AST and HBV-DNA). We considered the parameters as continuous variables and found differentially expressed genes related to these parameters. Most of these significant genes are enriched in immune response, interferon-stimulated, anti-apoptosis, and cell proliferation. Some important ones are also reported to correlate with HCC or other cancers.

We found genes correlated with clinical parameters provide insights for inflammation grades of CHB. We thereby constructed models with novel panels and validated by six normal samples, which can effectively predict binary classifications of inflammation and aid in the diagnosis of CHB. Notably, the novel panel with only clinical parameters was quite valuable, indicating that liquid biopsy method for detecting the pathology of CHB is possible.

Acknowledgements

We would like to thank Weilin Pu, Kelin Xu, Hua Dong, Chao Chen, Qianqian Peng, Feng Qian and Catherine Ketcham for their critical suggestions.

Authors' contributions

W.Z., Y.M., J.W., Z.H. and J.L. designed the project. J.L., J.Z., Y.M., X.Z., X.Z., Z.Z., J.Z., H.L., L.L., and Z.Y. provided HBV-infected samples and conducted experiments. W.Z., Y.M., J.H, L.W., Y.P., Y.Z ., M.Z., Y.W., Y.L. and J.Z contributed to the analyses. W.Z., Y.M., J.Z., J.L. and J.W. wrote the manuscript. Z.Y., L.J., J.L. and J.W. contributed to the final revision. All authors read and approved the final manuscript.

References

1. Chinese Society of Hepatology CMA, Chinese Society of Infectious Diseases CMA, Hou JL, et al. The guideline of prevention and treatment for chronic hepatitis B: a 2015 update. Zhonghua Gan Zang Bing Za Zhi 2015;23:888-905.

2. Li W, Zhao J, Zou Z, et al. Analysis of hepatitis B virus intrahepatic covalently closed circular DNA and serum viral markers in treatment-naive patients with acute and chronic HBV infection. PLoS One 2014;9:e89046.

3. ter Borg MJ, van Zonneveld M, Zeuzem S, et al. Patterns of viral decline during PEG-interferon alpha-2b therapy in HBeAg-positive chronic hepatitis B: relation to treatment response. Hepatology 2006;44:721-7.

4. Barrett T, Edgar R. Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. Methods Mol Biol 2006;338:175-90.

5. He D, Liu ZP, Honda M, et al. Coexpression network analysis in chronic hepatitis B and C hepatic lesions reveals distinct patterns of disease progression to hepatocellular carcinoma. J Mol Cell Biol 2012;4:140-52.

6. Ura S, Honda M, Yamashita T, et al. Differential microRNA expression between hepatitis B and hepatitis C leading disease progression to hepatocellular carcinoma. Hepatology 2009;49:1098-112.

7. He D, Li M, Guo S, et al. Expression pattern of serum cytokines in hepatitis B virus infected patients with persistently normal alanine aminotransferase levels. J Clin Immunol 2013;33:1240-9.

8. Castera L. Noninvasive methods to assess liver disease in patients with hepatitis B or C.

Gastroenterology 2012;142:1293-1302 e4.

9. Xu WS, Qiu XM, Ou QS, et al. Red blood cell distribution width levels correlate with liver fibrosis and inflammation: a noninvasive serum marker panel to predict the severity of fibrosis and inflammation in patients with hepatitis B. Medicine (Baltimore) 2015;94:e612.

10. Praneenararat S, Chamroonkul N, Sripongpun P, et al. HBV DNA level could predict significant liver fibrosis in HBeAg negative chronic hepatitis B patients with biopsy indication. BMC Gastroenterology 2014;14:218.

11. Desmet VJ, Gerber M, Hoofnagle JH, et al. Classification of chronic hepatitis: diagnosis, grading and staging. Hepatology 1994;19:1513-20.

12. Bedossa P, Poynard T. An algorithm for the grading of activity in chronic hepatitis C. The METAVIR Cooperative Study Group. Hepatology 1996;24:289-93.

13. 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.

14. Efron B, Hastie T, Johnstone I, et al. Least angle regression. The Annals of statistics 2004;32:407-499.

15. Montojo J, Zuberi K, Rodriguez H, et al. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics 2010;26:2927-8.

16. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7:3.

17. Wang Y, Li Y, Pu W, et al. Random Bits Forest: a Strong Classifier/Regressor for Big Data. Sci Rep 2016;6:30086.

18. Peng G, Luo L, Siu H, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet 2010;18:111-7.

19. Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 2000:455-66.

20. Newton RH, Leverrier S, Srikanth S, et al. Protein kinase D orchestrates the activation of DRAK2 in response to TCR-induced Ca2+ influx and mitochondrial reactive oxygen generation. J Immunol 2011;186:940-50.

21. Lee JY, Chae DW, Kim SM, et al. Expression of FasL and perforin/granzyme B mRNA in chronic

hepatitis B virus infection. J Viral Hepat 2004;11:130-5.

22. Hofmann I, Schlechter T, Kuhn C, et al. Protein p0071 - an armadillo plaque protein that characterizes a specific subtype of adherens junctions. J Cell Sci 2009;122:21-4.

23. Ni Y, Lempp FA, Mehrle S, et al. Hepatitis B and D viruses exploit sodium taurocholate co-transporting polypeptide for species-specific entry into hepatocytes. Gastroenterology 2014;146:1070-83.

24. Li J, Qi B, Chen P, et al. The expression of CD2 in chronic HBV infection. Cell Mol Immunol 2008;5:69-73.

25. Park IY, Sohn BH, Yu E, et al. Aberrant epigenetic modifications in hepatocarcinogenesis induced by hepatitis B virus X protein. Gastroenterology 2007;132:1476-94.

26. Kou XX, Hao T, Meng Z, et al. Acetylated Sp1 inhibits PTEN expression through binding to PTEN core promoter and recruitment of HDAC1 and promotes cancer cell migration and invasion. Carcinogenesis 2013;34:58-67.

27. Horikawa I, Barrett JC. Transcriptional regulation of the telomerase hTERT gene as a target for cellular and viral oncogenic mechanisms. Carcinogenesis 2003;24:1167-76.

28. Uesugi K, Hiasa Y, Tokumoto Y, et al. Wilms' tumor 1 gene modulates Fas-related death signals and anti-apoptotic functions in hepatocellular carcinoma. J Gastroenterol 2013;48:1069-80.

29. Li HG, Xie DR, Shen XM, et al. Clinicopathological significance of expression of paxillin, syndecan-1 and EMMPRIN in hepatocellular carcinoma. World J Gastroenterol 2005;11:1445-51.

30. Song Z, Li R, You N, et al. Loss of heterozygosity of the tumor suppressor gene Tg737 in the side population cells of hepatocellular carcinomas is associated with poor prognosis. Mol Biol Rep 2010;37:4091-101.

31. Raimondi C, Chikh A, Wheeler AP, et al. A novel regulatory mechanism links PLCgamma1 to PDK1. J Cell Sci 2012;125:3153-63.

32. Abuli A, Fernandez-Rozadilla C, Giraldez MD, et al. A two-phase case-control study for colorectal cancer genetic susceptibility: candidate genes from chromosomal regions 9q22 and 3q22. Br J Cancer 2011;105:870-5.

33. Rajkumar T, Vijayalakshmi N, Gopal G, et al. Identification and validation of genes involved in gastric tumorigenesis. Cancer Cell Int 2010;10:45.

Figure legends

Fig. 1. Value distribution of three clinical parameters.

The Y-axis is the number of samples, and the X-axis is the value of corresponding serum parameters. (A) and (B), the distributions of serum ALT and AST in 105 hepatitis samples, respectively; (C), the distribution of serum HBV-DNA in 90 hepatitis samples with transformed by $\log_{10}$.

Fig. 2. Workflow for this study and Venn diagram of significant genes for three clinical parameters.

(A), workflow for analyzing three clinical parameters, gene microarray data and inflammation grades of CHB; (B), the Venn diagram of significant genes, including left circle representing significant genes (positive versus negative) correlated with HBV-DNA, the middle representing significant genes correlated with ALT, and the right representing significant genes correlated with AST. The intersection sets are significant genes shared in the results of HBV-DNA and ALT and ALT and AST, respectively.

Fig. 3. Networks generated by significant genes correlated with three clinical parameters.

(A), networks correlated with ALT; (B) and (C), networks correlated with AST and HBV-DNA, respectively. The red circles represent positively correlated genes and the green represents negative ones. The grey circles represent important genes which are not in the significant gene lists but interact closely with significances. The lines interlinking two genes represent the type of interaction between two genes: orange lines represent co-expression, dark blue ones represent co-localization, red ones represent physical interaction, and purple ones represent protein domain sharing.

Fig. 4. PCA of 16 significant genes correlated both with ALT and AST.

(A), the plot of genes for PC1 with 10 positive and 6 negative loading coefficients;

(B), the genes for PC2 with 8 positive and 8 negative coefficients; (C), the genes for PC3 with 7 positive and 9 negative coefficients; (D), boxplot between G and PC3 scatter values of 102 HBV-infected samples. Fitted linear regression lines, R value and p-values are shown.

Fig. 5. ROC curves based on three predictive models and three modules with five-fold cross-validations.

Mean boxplot curves for each model are shown with the values of AUC and 95% CI, according to different predictive modules (using three clinical parameters only, nine genes only, or clinical parameters and genes). Dotted curves represent five-fold validations of each experiment.

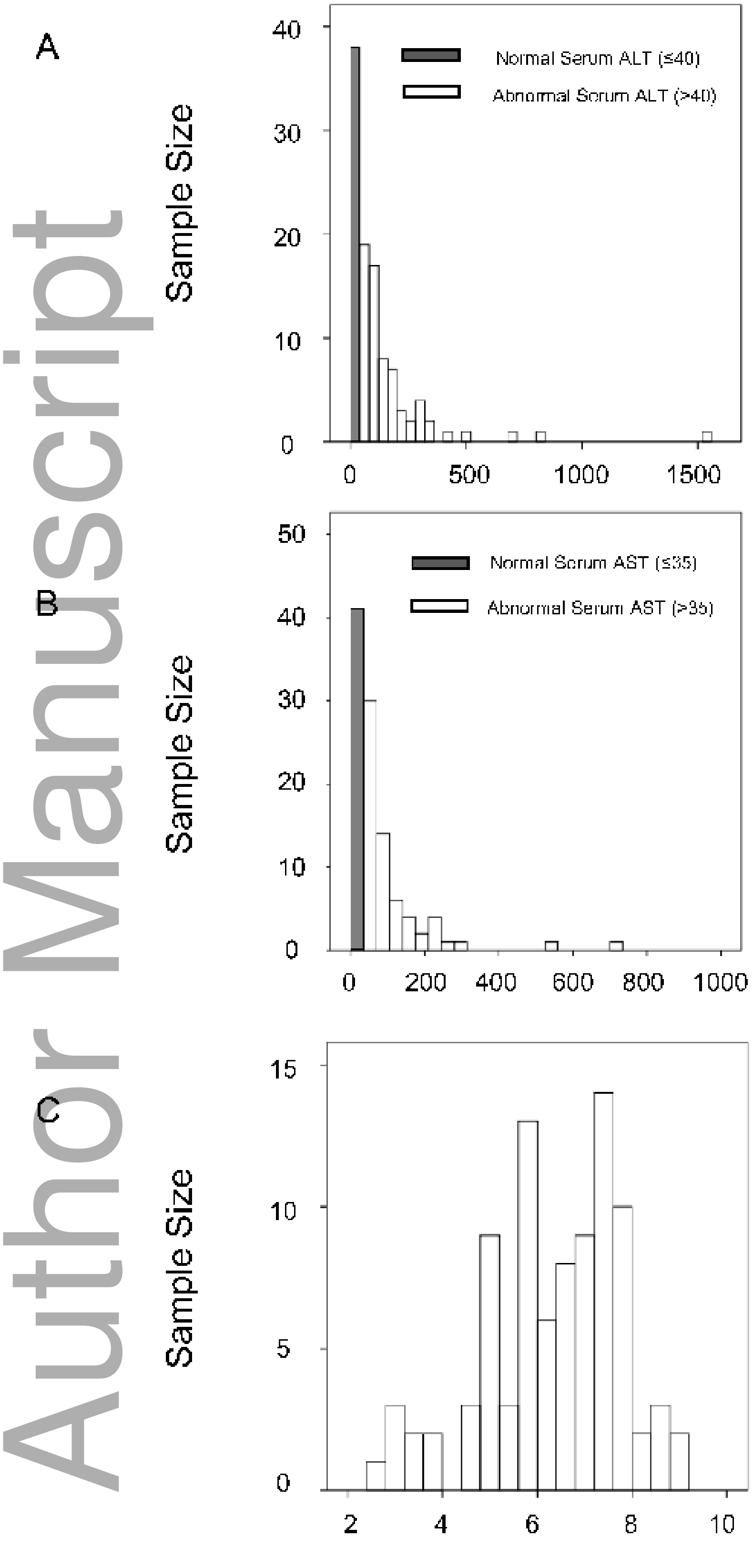
Tables

Table 1. Significant pathways correlated with three clinical parameters (P-values < 0.05). Three databases (BBID, KEGG and BIOCARTA) were subjected to pathway enrichment analysis.

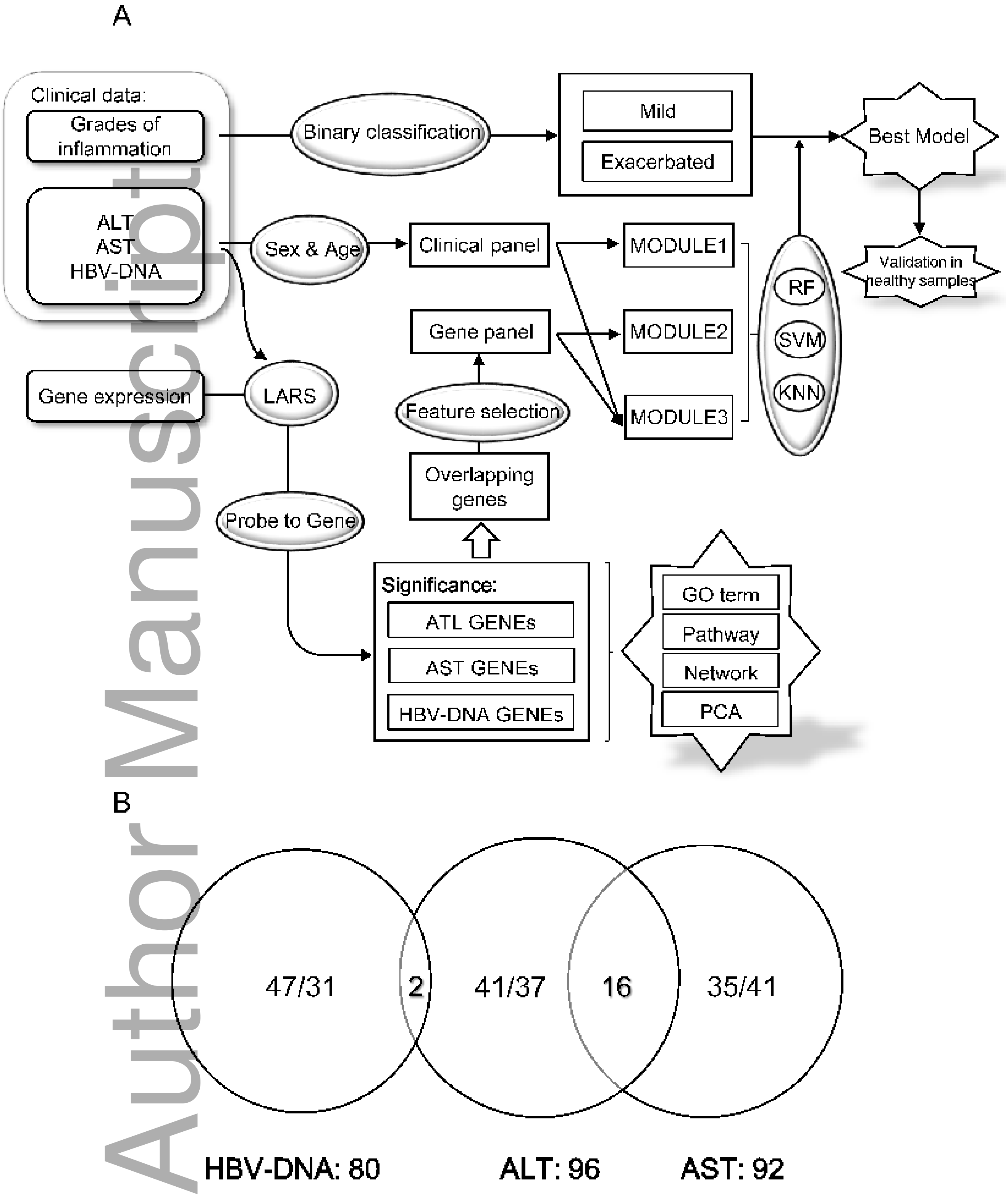
Type	Database	P-value	Genes	Term
ALT	BBID	0.0251	IGHG3, POU2F2, IGHM	B cell Activation
AST	KEGG	0.0008	IGF1R, FGF16, SMAD3, MDM2, BRCA2, BIRC5, ITGB1, TRAF4	Pathways in cancer (hsa05200)
	KEGG	0.0469	IGF1R, FGF16, MDM2	Melanoma (hsa05218)
	BBID	0.05	IGHG1, POU2F2	B cell Activation
HBV-DNA	BIOCARTA	0.0241	SP1, WT1	Overview of telomerase protein component gene hTert Transcriptional Regulation
	BIOCARTA	0.0411	C4A, C4B	Lectin Induced Complement Pathway
	BIOCARTA	0.0411	C4A, C4B	Classical Complement Pathway

Table 2. Specificity, sensitivity, accuracy of classification and AUC of predictive modules based on three methods with five-fold cross-validation.

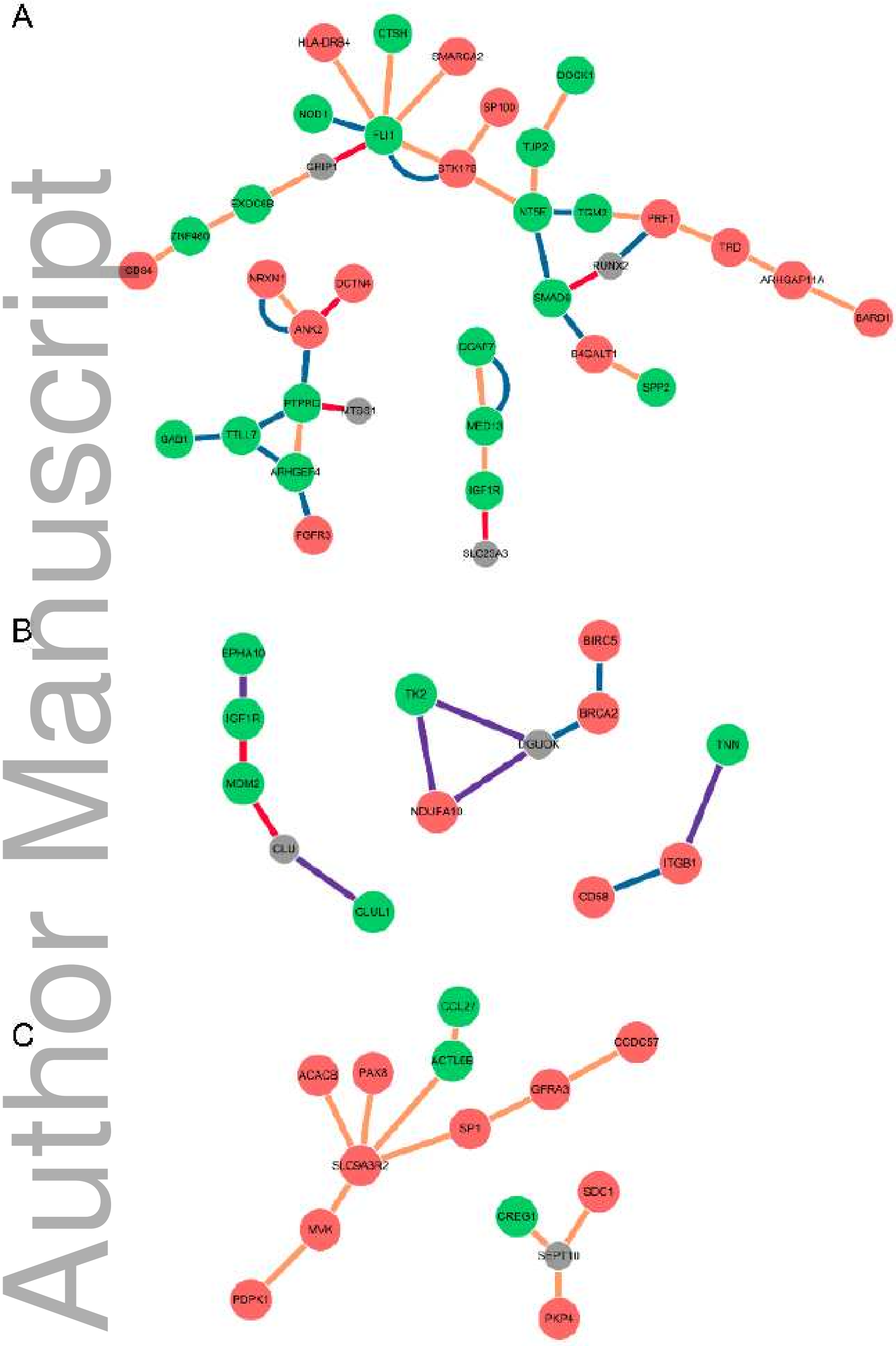
	Specificity	Sensitivity	Accuracy	AUC (95% CI)	Module
SVM	0.6053	0.7209	0.6667	0.7339 (0.5832-0.8146)	Module 1
	0.6579	0.6744	0.6667	0.7489 (0.5832-0.8165)	Module 2
	0.6579	0.6512	0.6543	0.7093 (0.5849-0.8129)	Module 3
KNN	0.5802	0.7971	0.6931	0.7286 (0.6244-0.8407)	Module 1
	0.6938	0.8150	0.7187	0.7226 (0.6543-0.8604)	Module 2
	0.6178	0.8857	0.7666	0.8057 (0.7108-0.8982)	Module 3
RF	0.6053	0.7674	0.6914	0.7841 (0.6450-0.8562)	Module 1
	0.6842	0.7209	0.7037	0.8015 (0.6903-0.8874)	Module 2



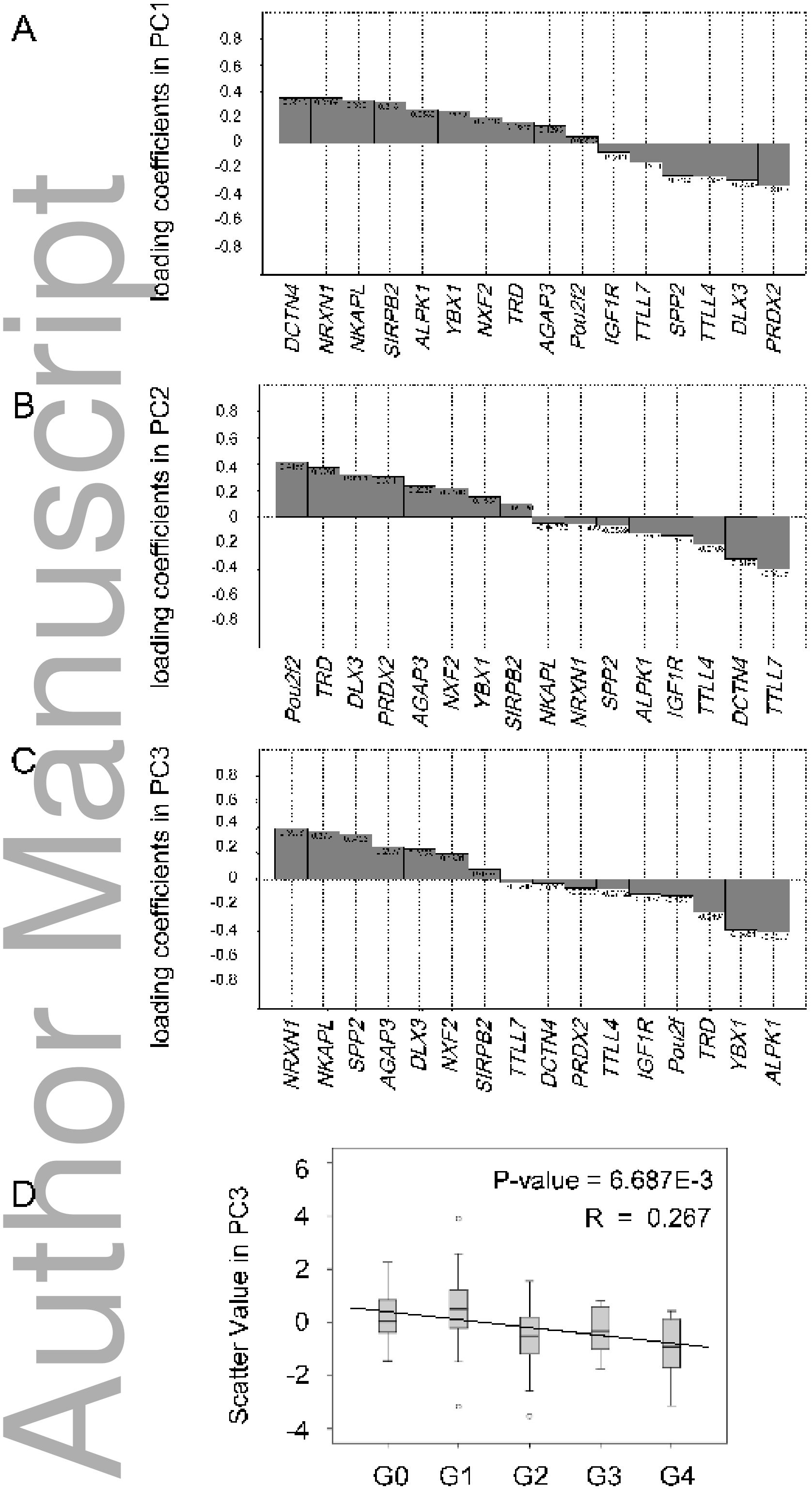
liv_13427_f1.tiff



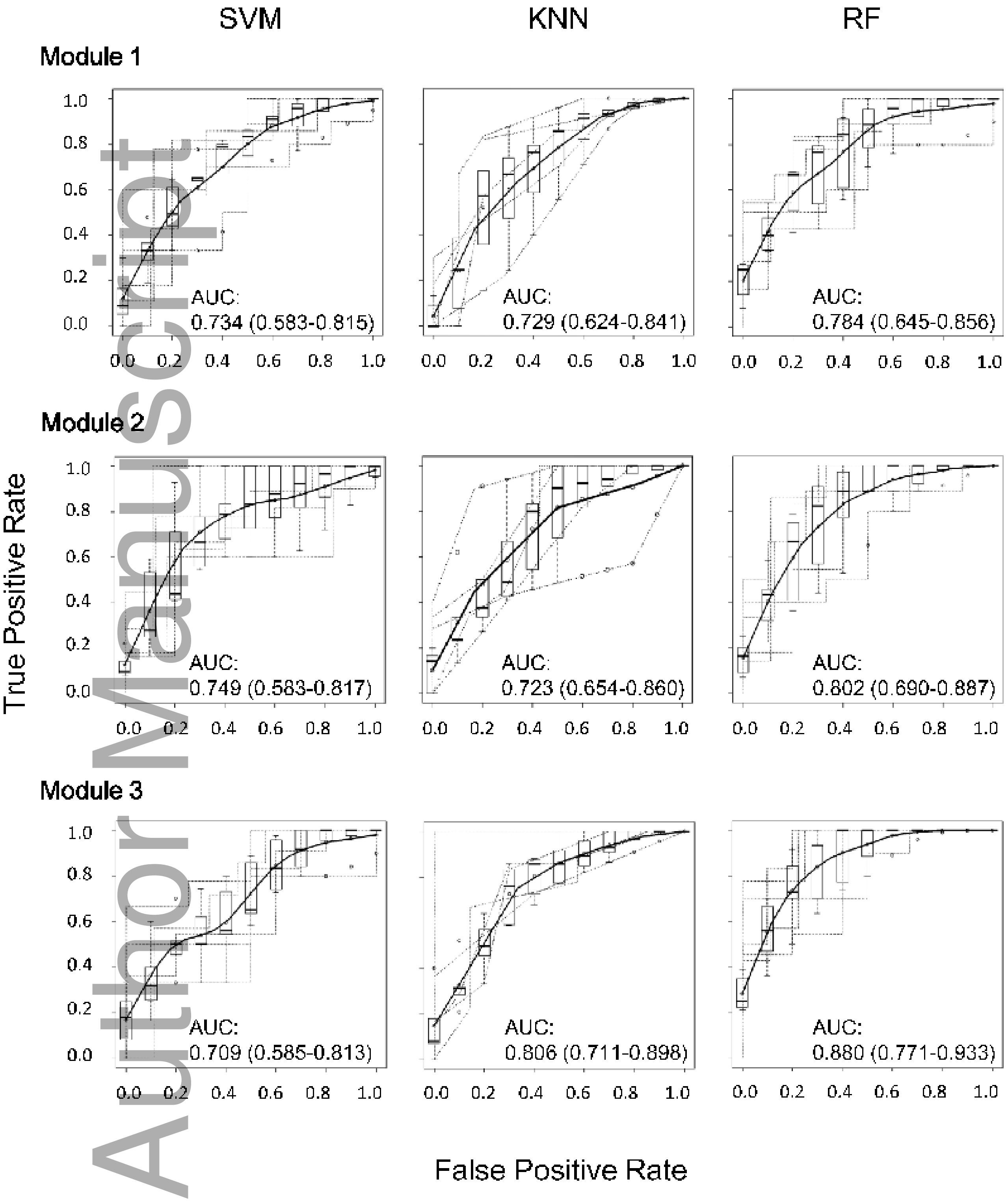
liv_13427_f2.tiff



liv_13427_f3.tiff



liv_13427_f4.tiff



liv_13427_f5.tiff