VOLUME 30 NUMBER 2 April 2024

pISSN 2287-2728
eISSN 2387-285X

CLINICAL and MOLECULAR HEPATOLOGY

The forum for latest knowledge of hepatobiliary diseases

cfDNA ULP-WGS for prognosis in HCC

Linvencorvir phase 2 trial for HBV
Signature gene set for discrimination of MASLD progression
Incidence of adverse events associated with NAFLD
JCAD in cholestatic fibrosis
Prognosis of MASLD

KASL

pISSN 2287-2728
eISSN 2287-285X

CLINICAL and MOLECULAR HEPATOLOGY

https://doi.org/10.3350/cmh.2023.0449
Clinical and Molecular Hepatology 2024;30:247-262

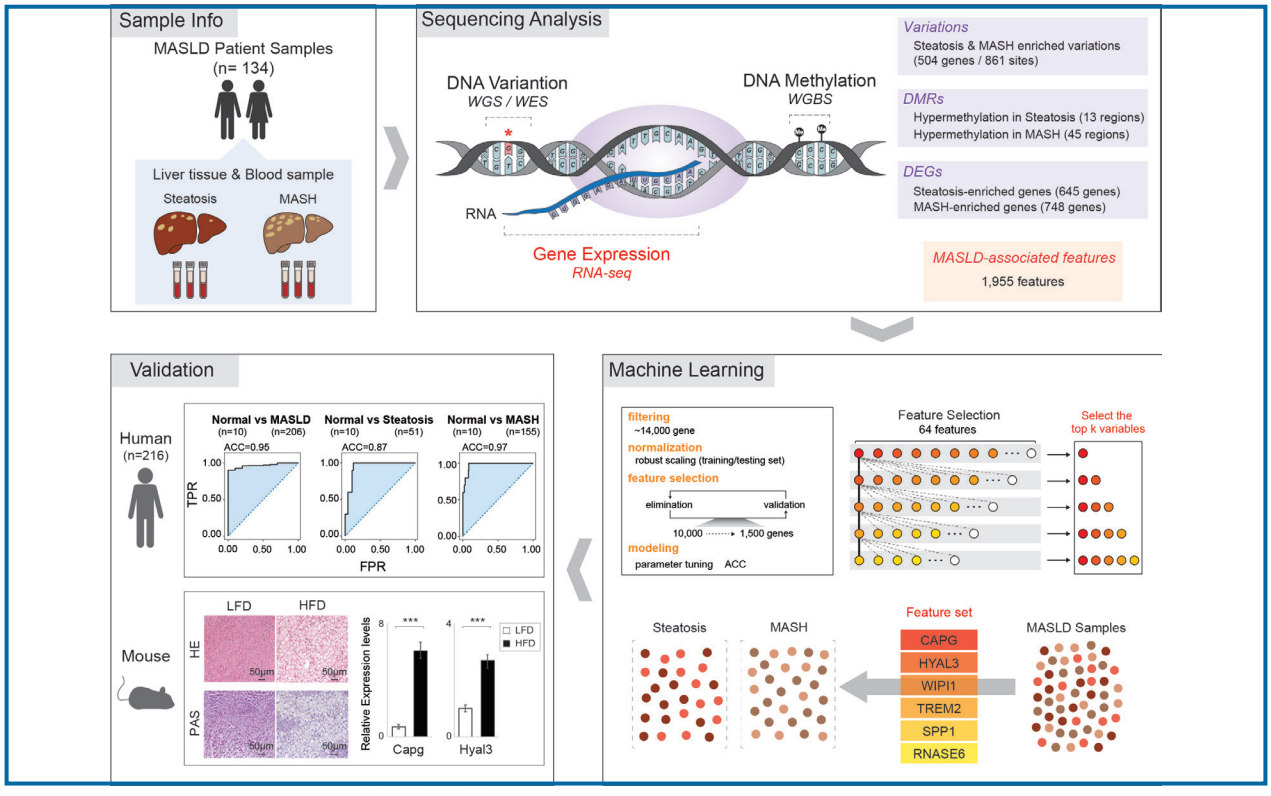
Original Article

Identification of signature gene set as highly accurate determination of metabolic dysfunction-associated steatotic liver disease progression

Sumin Oh[1,2,*], Yang-Hyun Baek[3,*], Sungju Jung[1,*], Sumin Yoon[1,*], Byeonggeun Kang[4,5,*], Su-hyang Han[1], Gaeul Park[6], Je Yeong Ko[7], Sang-Young Han[8], Jin-Sook Jeong[9], Jin-Han Cho[10], Young-Hoon Roh[11], Sung-Wook Lee[12], Gi-Bok Choi[13], Yong Sun Lee[6,14], Won Kim[15], Rho Hyun Seong[4], Jong Hoon Park[7], Yeon-Su Lee[6], and Kyung Hyun Yoo[1,2]

[1]Laboratory of Biomedical Genomics, Department of Biological Sciences, Sookmyung Women's University, Seoul, Korea; [2]Research Institute of Women's Health, Sookmyung Women's University, Seoul, Korea; [3]Liver Center, Department of Internal Medicine, Dong-A University College of Medicine, Busan, Korea; [4]Department of Biological Sciences and Institute of Molecular Biology and Genetics, Seoul National University, Seoul, Korea; [5]Bio-MAX Institute, Seoul National University, Seoul, Korea; [6]Division of Rare Cancer, Research Institute, National Cancer Center, Goyang, Korea; [7]Department of Biological Sciences, Sookmyung Women's University, Seoul, Korea; [8]Liver Center, On Hospital, Busan, Korea; [9]Department of Pathology, Dong-A University Medical Center, Busan, Korea; [10]Department of Diagnostic Radiology, Dong-A University Medical Center, Busan, Korea; [11]Department of Surgery, Dong-A University Medical Center, Busan, Korea; [12]Liver Center, Department of Internal Medicine, Dong-A University Medical Center, Busan, Korea; [13]Department of Radiology, On Hospital, Busan, Korea; [14]Department of Cancer Biomedical Science, Graduate School of Cancer Science and Policy, National Cancer Center, Goyang, Korea; [15]Department of Internal Medicine, Seoul National University College of Medicine, Seoul Metropolitan Government Boramae Medical Center, Seoul, Korea

Graphical Abstract



Study Highlights

- We used a multi-omics approach to investigate the genomes, epigenomes, and transcriptomes of 134 MASLD patients and identified 1,955 MASLD-associated features. Then, we used machine learning to select the features that most accuracy track MASLD progression. From this analysis, CAPG, HYAL3, WIPI1, TREM2, SPP1, and RNASE6 stood out as a signature gene set useful for discriminating the stages of MASLD progression. This signature gene set was verified using independent cohorts of MASLD, MASLD-associated cirrhosis, and liver cancer patients, suggesting it represents a group of biomarkers that apply to the full spectrum of MASLD-associated disease.

Corresponding author : **Kyung Hyun Yoo**
Laboratory of Biomedical Genomics, Department of Biological Sciences, Research Institute of Women's Health, Sookmyung Women's University, 100 Cheongpa-ro 47-gil, Yongsan-gu, Seoul 04310, Korea
Tel: +82-2-2077-7836, Fax: +82-2-2077-7258, E-mail: khryu@sookmyung.ac.kr
https://orcid.org/0000-0003-2172-5564

Yeon-Su Lee
Division of Rare Cancer, Research Institute, National Cancer Center, 323 Ilsan-ro, Ilsandong-gu, Goyang 10408, Korea
Tel: +82-31-920-2550, Fax: +82-31-920-2542, E-mail: yslee2@ncc.re.kr
https://orcid.org/0009-0009-4128-3557

Jong Hoon Park
Department of Biological Sciences, Sookmyung Women's University, 100 Cheongpa-ro 47-gil, Yongsan-gu, Seoul 04310, Korea
Tel: +82-2-710-9414, Fax: +82-2-2077-7322, E-mail: parkjh@sookmyung.ac.kr
https://orcid.org/0000-0002-8082-0214

*These authors equally contribute to this work.

Editor: Silvia Sookoian, CONICET, Argentina

Received : Nov. 1, 2023 / **Revised :** Jan. 9, 2024 / **Accepted :** Jan. 26, 2024

Abbreviations:

ACC, accuracy; AI, artificial intelligence; ATAC-seq, assay for transposase-accessible chromatin sequencing; C, cytosine; DEGs, differentially expressed genes; DMRs, differentially methylated regions; FC, fold change; FFA, free fatty acid; FPR, false positive rate; GATK, Genome Analysis Tool Kit; GLM, generalized linear regression model; GO, Gene ontology; H&E, hematoxylin and eosin; HCC, hepatocellular carcinoma; HFD, high-fat diet; HSI, hepatic steatosis index; KRGDB, Korean reference genome database; LFD, low-fat diet; MASLD, metabolic dysfunction-associated steatotic liver disease; MASH, metabolic dysfunction-associated steatohepatitis; NGS, next generation sequencing; PAS, periodic acid schiff; PCA, principal component analysis; PoN, panel of normal; PPI, protein-protein interaction; qRT-PCR, quantitative real time polymerase chain reaction; RBF, radical basis function; ROC, receiver operating characteristic; SNP, single nucleotide polymorphism; SNV, single nucleotide variant; SVM, support vector machine; T, thymine; TPR, true positive rate; TSS, transcription start site; WES, whole exome sequencing; WGBS, whole genome bisulfite sequencing; WGS, whole genome sequencing

Background/Aims: Metabolic dysfunction-associated steatotic liver disease (MASLD) is characterized by fat accumulation in the liver. MASLD encompasses both steatosis and MASH. Since MASH can lead to cirrhosis and liver cancer, steatosis and MASH must be distinguished during patient treatment. Here, we investigate the genomes, epigenomes, and transcriptomes of MASLD patients to identify signature gene set for more accurate tracking of MASLD progression.

Methods: Biopsy-tissue and blood samples from patients with 134 MASLD, comprising 60 steatosis and 74 MASH patients were performed omics analysis. SVM learning algorithm were used to calculate most predictive features. Linear regression was applied to find signature gene set that distinguish the stage of MASLD and to validate their application into independent cohort of MASLD.

Results: After performing WGS, WES, WGBS, and total RNA-seq on 134 biopsy samples from confirmed MASLD patients, we provided 1,955 MASLD-associated features, out of 3,176 somatic variant callings, 58 DMRs, and 1,393 DEGs that track MASLD progression. Then, we used a SVM learning algorithm to analyze the data and select the most predictive features. Using linear regression, we identified a signature gene set capable of differentiating the various stages of MASLD and verified it in different independent cohorts of MASLD and a liver cancer cohort.

Conclusions: We identified a signature gene set (i.e., *CAPG*, *HYAL3*, *WIPI1*, *TREM2*, *SPP1*, and *RNASE6*) with strong potential as a panel of diagnostic genes of MASLD-associated disease. **(Clin Mol Hepatol 2024;30:247-262)**

Keywords: MASLD; Multi-omics; Machine learning; Signature gene set; Biomarker

INTRODUCTION

Metabolic dysfunction-associated steatotic liver disease (MASLD) is a metabolic disease characterized by fat accumulation in the liver.[1-3] MASLD includes simple steatosis, which is relatively early-stage and low risk, and metabolic dysfunction-associated steatohepatitis (MASH), which is late-stage disease characterized by serious liver inflammation and fibrosis.[4-6] Since MASH is often a precursor of cirrhosis, liver cancer, and liver failure, it is critical to discriminate between steatosis and MASH to guide patient treatment.[7-9] MASLD can be diagnosed using various non-invasive assessment methods, including imaging techniques, blood tests, and fibrosis assessment. However, a combination of these methods is required for a more accurate diagnosis and to assess the severity of MASLD.[10-12] This underscores the need to identify novel molecular markers that would facilitate a faster and more precise MASLD staging.

Genome, transcriptome, and epigenome sequencing have already suggested potential biomarkers of MASLD in previous studies.[13] Genetic variants, specifically single nucleotide polymorphism (SNPs) in PNPLA3, GCKR, TM6SF2, and AGXT2, have been associated with MASLD progression.[14,15] Comprehensive RNA-seq analyses have identified differentially expressed genes (DEGs) related to MASLD, providing insights into its severity involving processes such as the ablation of extracellular molecules, cytokine responses, and immune system functions.[16,17] In addition, epigenetic markers, particularly DNA methylation, have been explored. DNA methylation signatures related to age acceleration were correlated with MASLD severity, and hepatic fat-associated CpGs in peripheral blood samples of patients with type 2 diabetes revealed differentially methylated regions (DMRs), including ABCG1, CPT1A, and TMEM50B.[18,19] Despite these efforts uncovering significant markers associated with various stages of MASLD progression, securing the optimal gene set for accurately diagnosing a patient's specific stage of MASLD progression remains an ongoing challenge.

Therefore, we decided to collect and analyze genomic, epigenomic, and transcriptomic data from a single cohort of patients progressing from steatosis to MASH, aiming to identify features that would enable an accurate diagnosis of MASLD stages. By feeding the MASLD-associated into a series of machine learning models that used linear regression methods, we were able to identify a set of 6 MASLD signature genes accurate enough to discriminate MASLD stage. We verified the utility of this gene set by using them to distinguish an independent cohort of MASLD and liver cancer patients from controls. Thus, this gene set will likely prove useful for the early diagnosis of MASLD and in guiding MASLD

patient treatment.

MATERIALS AND METHODS

Sample and sequencing library preparation

Pathologically proven biopsy-tissue and blood samples were obtained from a cohort of 134 MASLD patients, comprising 60 steatosis and 74 MASH patients in the study cohort who were recruited from the Dong-A University Hospital (Informed consent was obtained from all subjects, DAUHIRB-17-197) and Onhospital (Informed consent was obtained from all subjects, ONHIBR-19-001), Busan, Rep. of Korea. All fresh samples were frozen immediately after biopsy and stored at –70℃ according to the protocols approved by the institutional review board for the human subject guideline that is in accordance with the principles of the Declaration of Helsinki. Hospital medical records and pathology reports of patients were reviewed by internal pathologist. The clinical features and the information of samples used for NGS analysis were provided in Supplementary Table 1 and Supplementary Table 2. For whole genome sequencing (WGS) and whole exome sequencing (WES), DNA was extracted from tissues and blood from MASLD patients. WGS libraries were generated using TruSeq Nano DNA (350), and 150-bp paired-end reads were sequenced on the Illumina platform. WES libraries were prepared using the SureSelectXT Library Prep Kit, and 100-bp paired-end reads were sequenced on the Illumina platform. For whole genome bisulfite sequencing (WGBS), samples were prepared using the Accel-NGS Methyl-Seq DNA Library Kit and the EZ DNA Methylation-Gold Kit. Then, 150-bp paired-end reads from the resulting libraries were sequenced on the Illumina platform. For total RNA-seq, RNA was extracted from the tissues of MASLD patients. Libraries were generated using the TruSeq Stranded Total RNA LT Sample Prep Kit, and 100-bp paired-end reads were sequenced on the Illumina platform (All sequencing was carried out by Macrogen, Inc., Seoul, Korea).

Detailed experimental procedures for histological diagnosis, genomic and epigenomic analysis, transcriptome analysis, machine learning, open chromatin accessibility analysis, statistics, high-fat diet (HFD) mouse model, hematoxylin and eosin (H&E) and with periodic acid schiff (PAS) staining, hepatocyte organoid culture, free fatty acid (FFA) treatment and Oil Red O staining and qRT-PCR are provided in supplementary information.

RESULTS

Identification of MASLD-associated somatic variants

To discover MASLD-associated markers, we took a multi-omics approach, looking at genomic, epigenomic, and transcriptomic data from WGS, WES, WGBS, and total RNA-seq using pathologically-proven biopsy tissue samples obtained from 134 MASLD patients (Fig. 1A). First, to limit our exploration to somatic markers that offer insights into genetic changes occurring in diseased cells, enhancing our understanding of the molecular basis of the disease, we eliminated any germline mutations by comparing WGS data obtained from liver biopsies with those obtained from blood samples (Fig. 1B). By integrating WES data screening for somatic variants in exon regions likely to affect the function of genes, we narrowed our search to 3,888 somatic variant callings. Of these, 79% (3,054) were classified as type of missense mutations. The most common type of somatic variant was the SNP, specifically the SNV in which a T nucleotide was altered to a C (Supplementary Fig. 1). Then, we focused on 504 different genes with 861 somatic variant sites detected in more than two of the 120 MASLD patient samples (Supplementary Table 3). These 504 genes with MASLD-associated somatic variants were broadly distributed throughout all chromosomes (Fig. 1C). Next, we asked whether the variants in these 504 genes were exclusive mutations (Fig. 1D). We found that 346 of 504 genes (69%) with the variants were exclusive, but the remaining 158 genes (31%) showed multiple, non-exclusive variants. Among them, genes mostly showed two variation sites. When we classified the various exclusive or non-exclusive variants in individual genes (Fig. 1E), we found missense mutations were the most common in both genes with exclusive and non-exclusive variants. In genes with non-exclusive variants, we observed cases in which two of the same type of variation appeared along with cases showing combinations of two or more different types. To determine the contribution of these MASLD-associated somatic variants to gene expression, we analyzed the expression levels of the 504 genes between their altered and non-altered groups

https://doi.org/10.3350/cmh.2023.0449

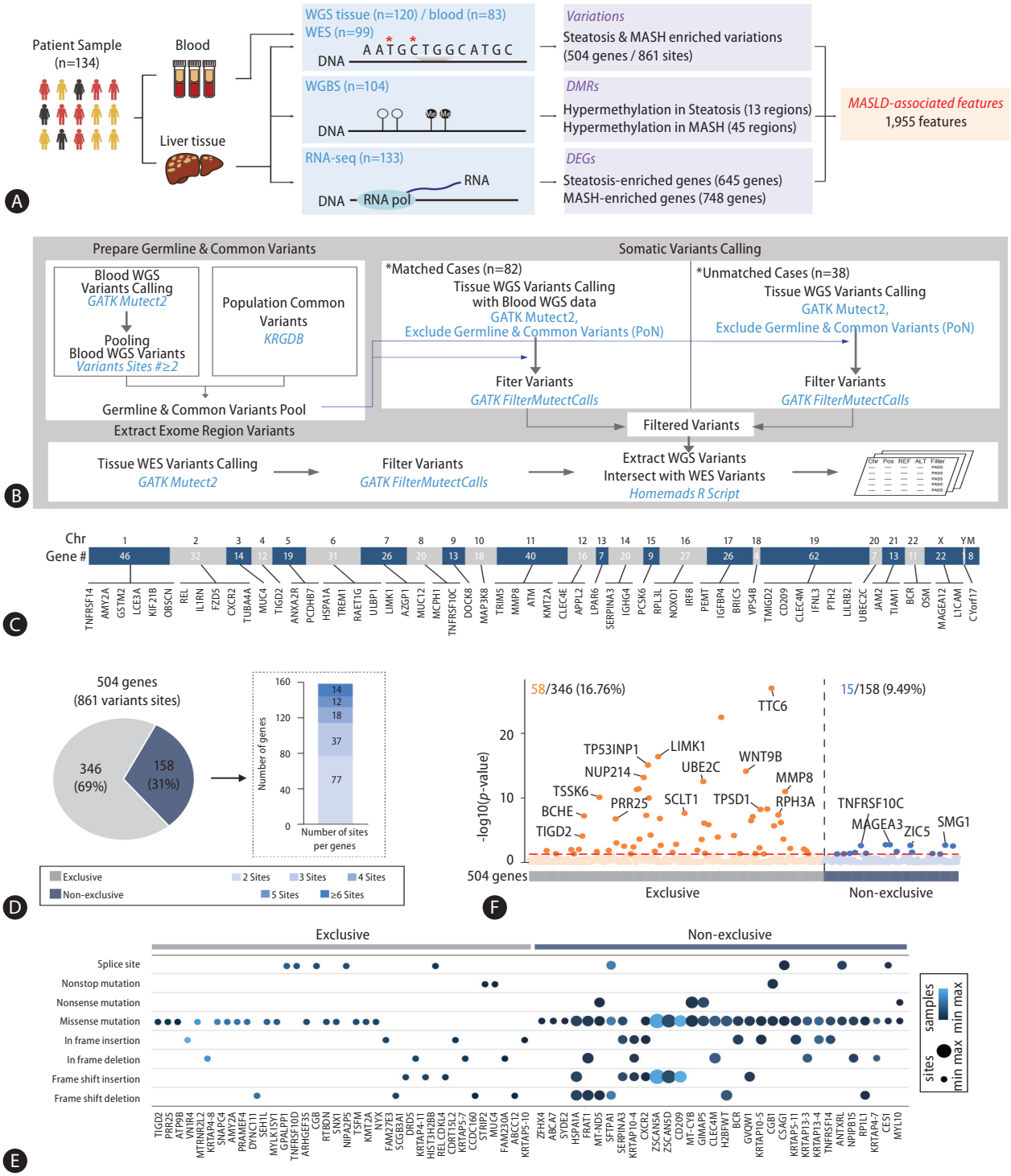


Figure 1. MASLD-associated somatic variants identified through comprehensive WGS and WES analysis. (A) Overall research strategy for identifying MASLD-associated features via a multi-omics approach. (B) Pipeline for calling somatic variants. (C) Distribution of genes with MASLD-associated somatic variations across the chromosomes. (D) Pie chart showing genes with exclusive or non-exclusive variants. (E) Dot plot presenting the types of mutations in genes with exclusive or non-exclusive variants. (F) Dot plot showing gene expression changes between the altered and non-altered groups. MASLD, metabolic dysfunction-associated steatotic liver disease; WES, whole exome sequencing; WGBS, whole genome bisulfite sequencing; WGS, whole genome sequencing; DEGs, differentially expressed genes; DMRs, differentially methylated regions.

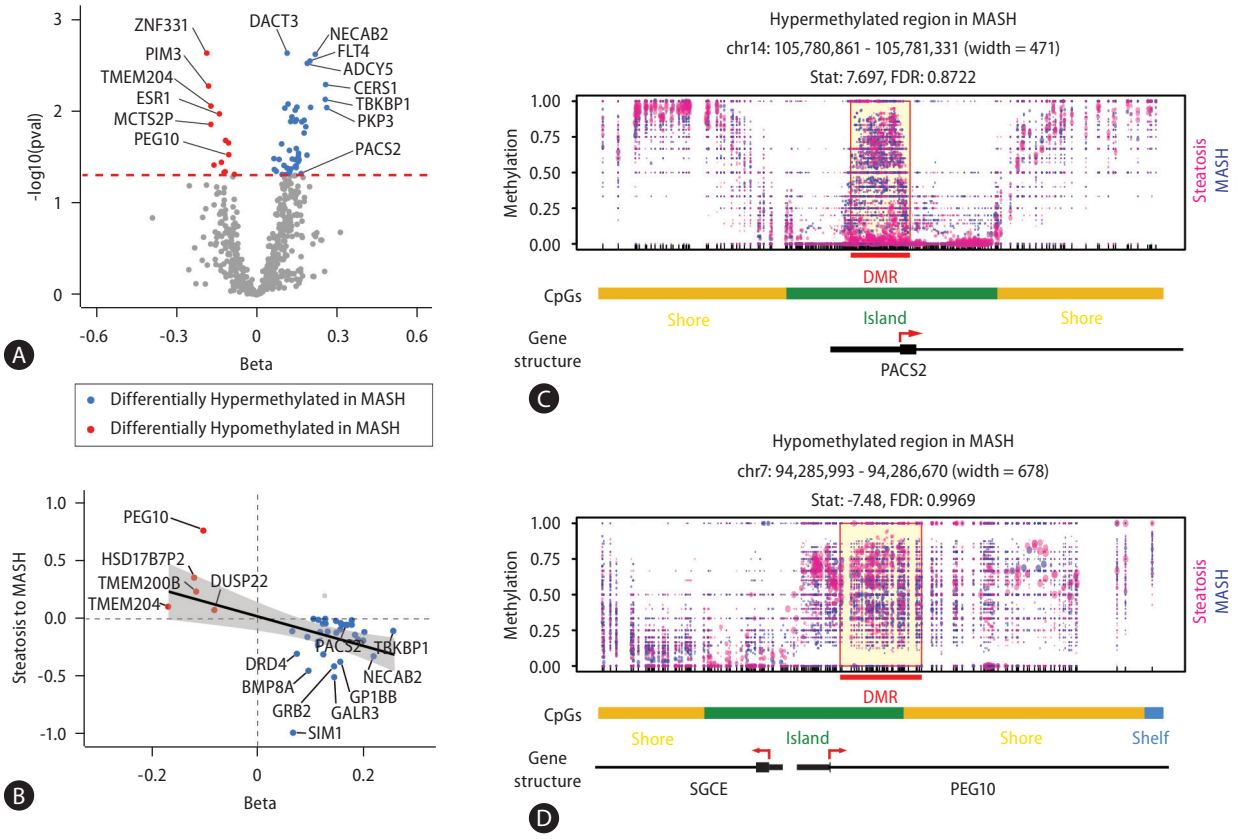


Figure 2. Identification of differentially methylated regions associated with MASLD progression. (A) Scatter plot showing genes with a methylation ratio that is significantly different between steatosis and MASH samples. (B) Correlation between DNA methylation status and gene expressions. (C) Representative loci showing hypermethylation in the *PACS2* and hypomethylation in the *PEG10* promoter. MASLD, metabolic dysfunction-associated steatotic liver disease; MASH, metabolic dysfunction-associated steatohepatitis.

(Fig. 1F). We found 16.76% (58) of the 346 genes with exclusive variations showed a statistically significant differences in their expression level, while only 9.49% (15) of the 158 genes with non-exclusive variants showed statistically significant expression changes.

Since variations of *PNPLA3*, *TM6SF2,* and *AGXT2* have all been reported as genetic factors contributing to MASLD, we also examined the variations on these genes and detected in the WGS results from both liver tissues and in blood, suggesting they are instead germline mutations (Supplementary Fig. 2). In our cohort, a *PNPLA3* variation (rs738409 C>G), a *TM6SF2* variation (rs58542926 C>T), and an *AGXT2* variation (rs2291702 T>C) were detected in 76.67%, 25.83%, and 67.50% of the samples, respectively (Supplementary Fig. 2A). We confirmed diminished expression levels of these genes in the steatosis and MASH altered groups, with remarkable reductions in homozygous variants (Supplementary Fig. 2B and 2C). Thus, these MASLD-related genetic variations were common in our cohort, but they were excluded because we were searching specifically for somatic mutations. From these results, we suggest MASLD-associated somatic variations in 504 genes.

Differentially methylated regions in MASLD

Next, to identify DMRs associated with MASLD, we performed WGBS in 104 MASLD patients (Fig. 2). We identified 87 DMRs with p-values less than 0.05 in the comparison between steatosis and MASH samples. 68 of 87 DMRs (78%) were located within known CpG regions and 58 of these DMRs (66.7%) were annotated to reference genes (Supplementary Table 4). Of these 58 DMRs, 13 DMRs were hypo-

https://doi.org/10.3350/cmh.2023.0449

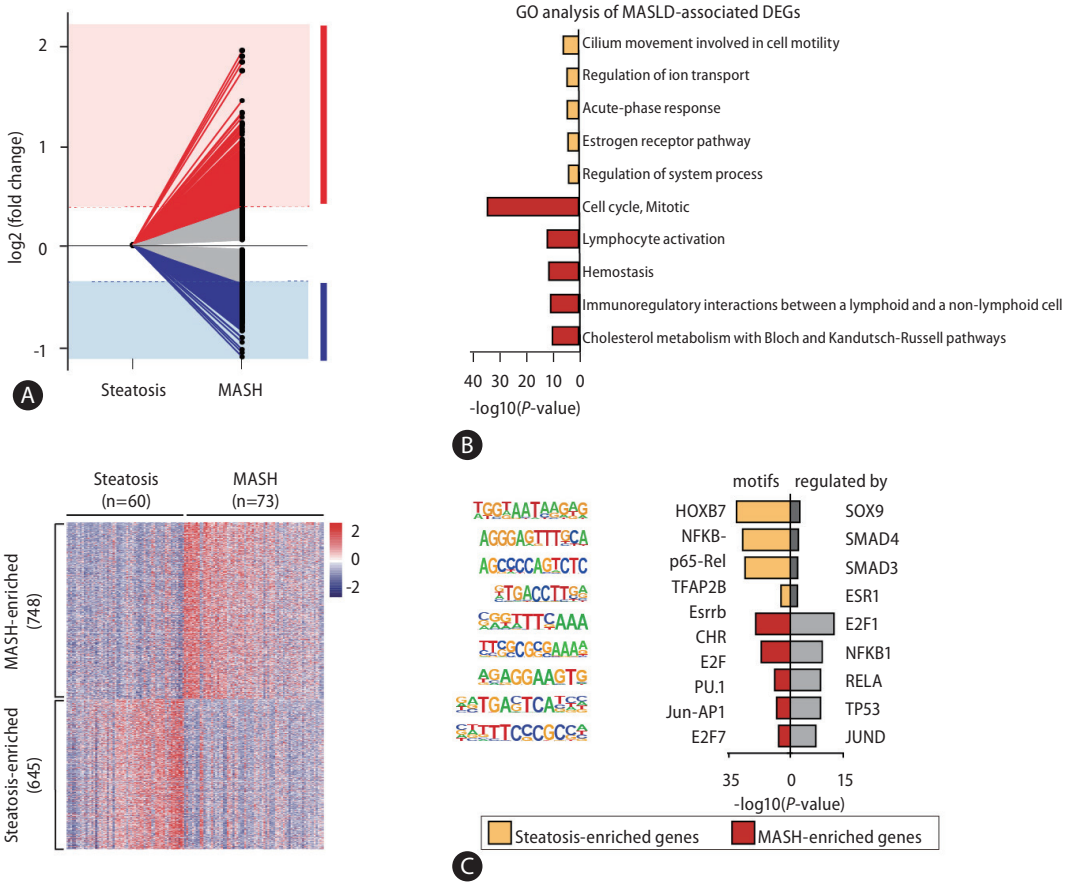


Figure 3. Transcriptomic profiling of MASLD progression. (A) Line plot representing gene expression fold change in the comparison between steatosis and MASH samples. (red, MASH-enriched genes; blue, steatosis-enriched genes). Heat map showing the expression levels of 1,393 genes in 133 MASLD patients. (B) Bar plots showing the results of GO analysis for steatosis- and MASH-enriched genes. (C) Representative results of a motif search analysis and TRRUST analyses. Left bar, motif search results based on known or de novo motif sequences; right bar, the results of the enrichment analysis by TRRUST. MASLD, metabolic dysfunction-associated steatotic liver disease; MASH, metabolic dysfunction-associated steatohepatitis; GO, Gene ontology; DEGs, differentially expressed genes.

methylated and 45 DMRs were hypermethylated in MASH (Fig. 2A). We next asked whether the differential methylation associated with MASLD progression also contributed to gene expressions (Fig. 2B). As results, of the 13 genes with hypomethylated CpGs and the 45 genes with hypermethylated CpGs, 38.4% (5) and 68.8% (31) showed inverse correlations with gene expression, respectively. Indeed, the correlation coefficients comparing methylation status and gene expression were statistically significant (*P*-value=3.07E-03). Figure 2C shows the hypermethylated promoter region of PACS2 and the hypomethylated promoter of PEG10 as examples of altered genes associated with MASH. Together, our results of epigenomic analysis provided MASLD-associated DMRs that could affect disease progression through the regulation of gene expression.

Genes related to MASLD progression

Next, to investigate genes related to MASLD progression, we performed a total RNA-seq analysis and found 1,393 DEGs in the comparison of steatosis and MASH (Supplementary Table 5). Among these, 645 steatosis- and 748 MASH-enriched genes were defined by a 1.3-fold or greater change in expression level between MASLD stages (Fig. 3A). To understand the function of steatosis- and MASH-enriched genes, we performed analysis of GO (Fig. 3B), motif search, and TRRUST enrichment (Fig. 3C). Results of these analysis showed DEGs were involved in terms of cell-cell adhesion, metabolic pro-

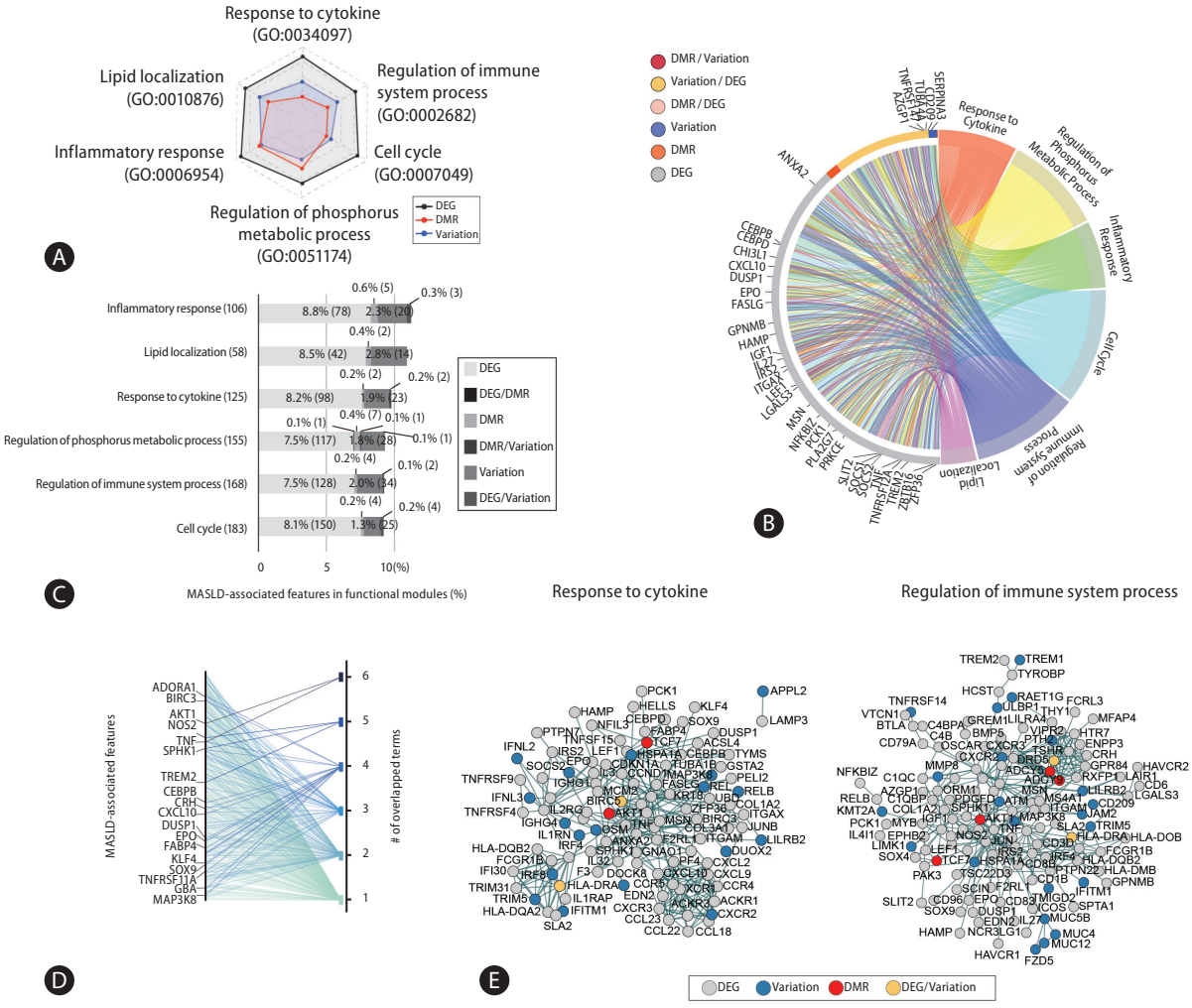


Figure 4. Comprehensive networks of MASLD-associated features within functional modules. (A) Representative MASLD-associated functional modules. The proportion of genes with MASLD-associated somatic variants (blue), DMRs (red), and DEGs (black) in each individual functional module (The range for black is 0–20%, for blue is 0–5%, and for red is 0–1%). (B) Circular plot indicating that individual functional modules included genetic, epigenetic, and transcriptomic features. (C) Bar plot showing the proportion of MASLD-associated somatic variations, DMRs, and DEGs assigned to functional modules. (D) Line plot showing that MASLD-associated features were simultaneously related with one another in functional modules. (E) Maps of the PPI networks of MASLD-associated features involved in the response to cytokines and regulation of immune system processes modules. MASLD, metabolic dysfunction-associated steatotic liver disease; DMRs, differentially methylated regions; DEGs, differentially expressed genes; PPI, protein-protein interaction.

cess, and cytokine signaling and were regulated by transcription factors such as NFκB, JUN, and SMAD3/4 have already been associated with MASLD progression.

Integrated networks of MASLD-associated features within functional modules

From MASLD-associated somatic variations, DMRs, and DEGs we identified, we designated 1,955 as MASLD-associated features. We next investigated whether these MASLD-associated features collaborate in functional modules (Fig. 4). Considering the proportion of MASLD-associated features in each module, frequency represented in terms for steatosis- or MASH-enriched genes, we found dominant 6 functional modules such as response to cytokine, regulation of immune system process, cell cycle, regulation of phosphorus metabolic process, inflammatory response, and lipid localization (Fig. 4A). MASLD-associated features accounted for about 10% of the list corresponding to genes annotated from the public database of functional modules. Since MASLD-associ-

https://doi.org/10.3350/cmh.2023.0449

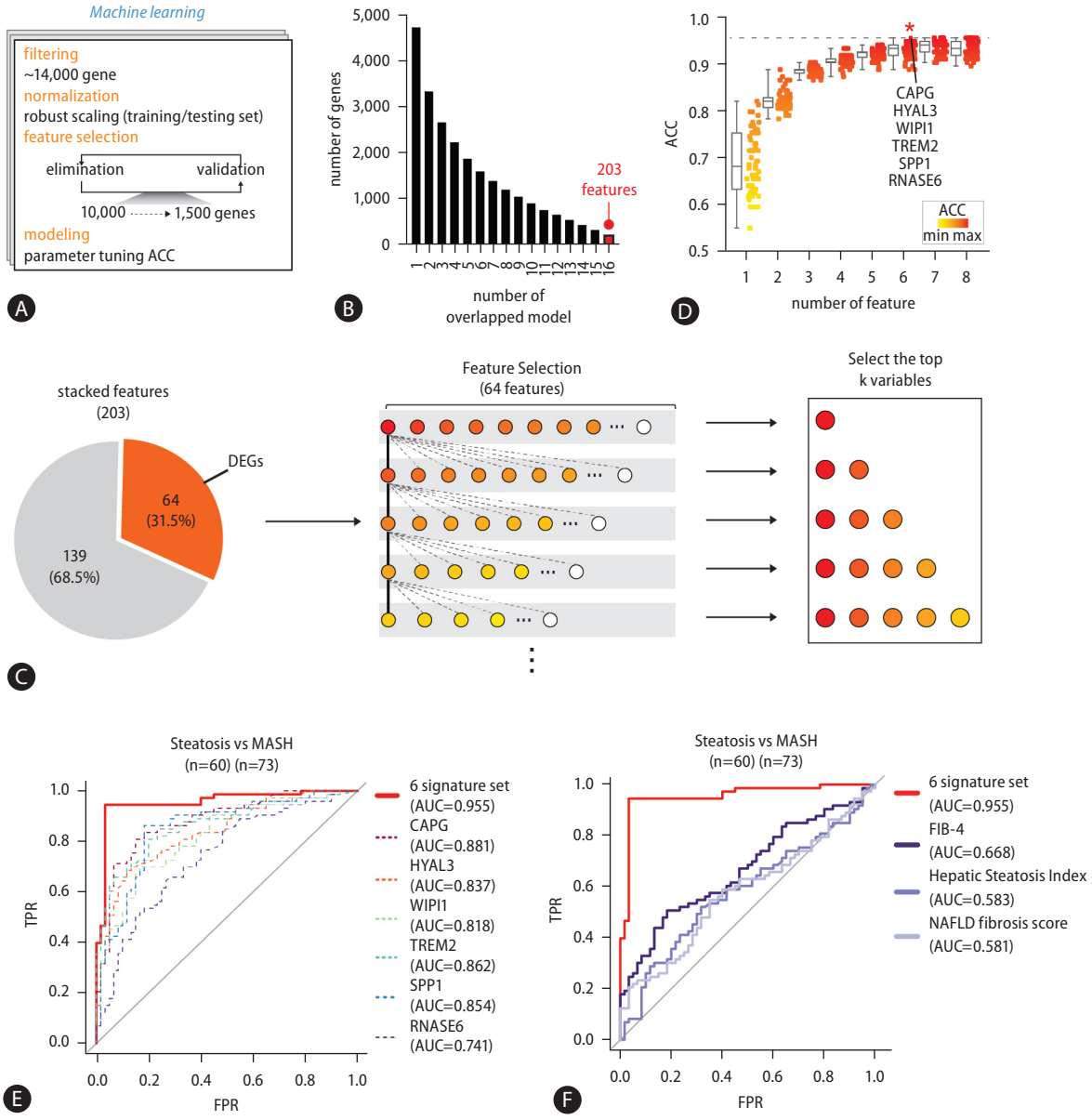


Figure 5. Using machine learning modeling to select features that permit MASLD stage discrimination. (A) Feature selection via machine learning modeling. (B) 203 stacked features obtained from 16 independent models. (C) Designing the signature gene set consisting of the top-ranked genes that provided the highest accuracy. (D) Dot plot of signature gene sets of various sizes against their accuracy in discriminating MASLD stages. The chosen gene set is indicated (ACC=0.955). (E) ROC curve plots showing the accuracy of the 6 signature gene set and individual genes (6 signature set *P*-value=1.04E-19; CAPG *P*-value=2.48E-14; HYAL3 *P*-value=1.26E-11; WIPI1 *P*-value=1.57E-10; TREM2 *P*-value=3.64E-13; SPP1 *P*-value=1.28E-12; RNASE6 *P*-value=9.19E-07). (F) ROC curve plots indicating the accuracy of non-invasive indices and the signature gene set (6 signature set *P*-value=1.04E-19; FIB-4 *P*-value=4.48E-04; Hepatic Steatosis Index *P*-value=5.11E-02; NAFLD fibrosis score *P*-value=5.50E-02). MASLD, metabolic dysfunction-associated steatotic liver disease; ACC, accuracy; ROC, receiver operating characteristic.

ated features may simultaneously be MASLD-associated variations, DMRs, or DEGs, we categorized them in detail. Individual functional modules included genomic, epigenetic, and transcriptomic features (Fig. 4B). As an example, the 125 MASLD-associated features related to cytokine responses included 98 DEGs, 2 DMRs, 23 genes with variations, and 2 genes involved in DEGs/variations (Fig. 4C). We found MASLD-associated features appearing in one or more func-

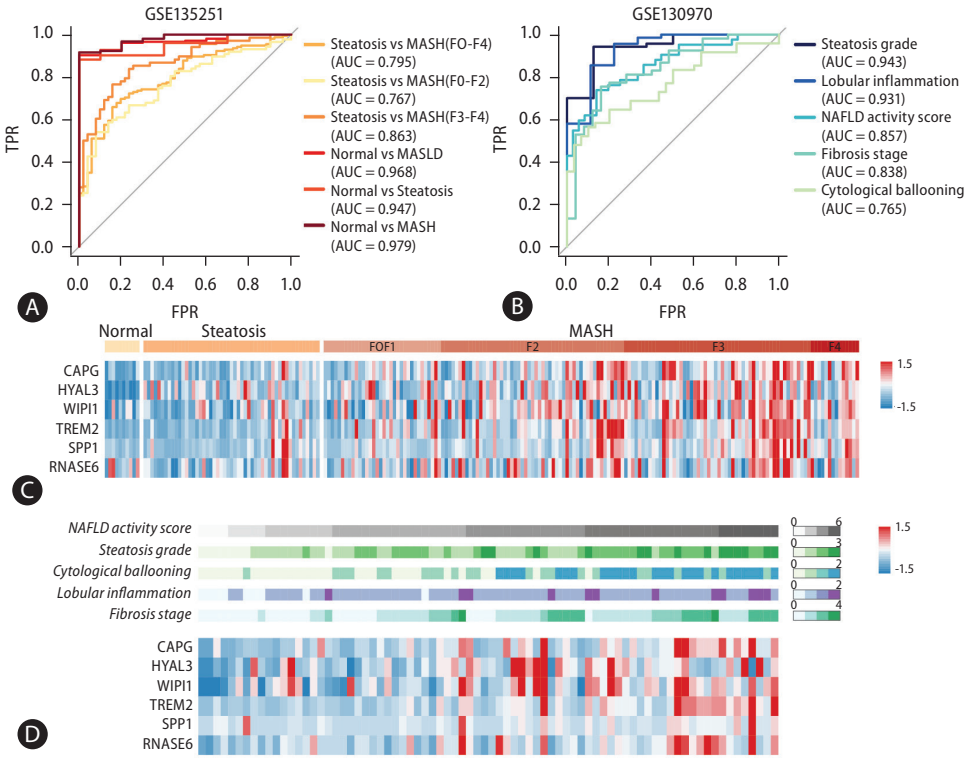


Figure 6. Application of the signature gene set to MASLD progression. (A) ROC curve plots describing the ratio of the true positive rate (TPR) and false positive rate (FPR) for the GLM designed using the signature gene set when predicting results from an independent cohort of normal (n=10), steatosis (n=51) and MASH (n=155) samples (Steatosis vs. MASH(F0-F4) *P*-value=1.28E-10; Steatosis vs. MASH(F0-F2) *P*-value=9.03E-08; Steatosis vs. MASH(F3-F4) *P*-value=7.00E-12; Normal vs. MASLD *P*-value=3.07E-07; Normal vs. Steatosis *P*-value=4.67E-06; Normal vs. MASH *P*-value=2.00E-07). (B) Validation of the accuracy of the signature gene set between various histological features related to MASLD (Steatosis grade *P*-value=2.29E-05; Lobular inflammation *P*-value= 1.48E-05; NAFLD activity score *P*-value=4.31E-09; Fibrosis stage *P*-value=8.51E-07; Cytological ballooning *P*-value=2.75E-05). (C) Heatmap showing the expression levels of the signature genes from normal, steatosis, and MASH samples. (D) The expression levels of signature genes in subgroups of histological features related to MASLD. (E) H&E and PAS staining showing liver morphology changes in an in vivo model fed an HFD compared to an LFD (Top). Expression levels of the signature genes in an in vivo model measured by qRT-PCR (Bottom). (F) Representative bright-field images showing morphology changes in mouse hepatic organoids treated with 1 mM FFA. Oil red O staining showed lipid accumulation in organoids treated with 1 mM FFA, mimicking hepatic steatosis (Top). Relative mRNA expression levels of the signature genes in mouse hepatic organoids treated with 1 mM FFA (Bottom). (Student's *t*-test, *P*-value; *<0.05, **<0.01, ***<0.001). MASLD, metabolic dysfunction-associated steatotic liver disease; ROC, receiver operating characteristic; GLM, generalized linear regression model; MASH, metabolic dysfunction-associated steatohepatitis; FFA, free fatty acid.

tional modules, both specifically and sometimes redundantly (Fig. 4D). They also showed strong cooperation within MASLD-associated functional modules (Fig. 4E and Supplementary Fig. 3). Thus, MASLD-associated features were connected primarily to representative functional modules related to MASLD, collaborating with one another within their functional modules. This suggests MASLD-associated variations, differential methylated regions, and expression changes have a close relationship with one another.

Identification of signature genes through feature selection

To ascertain the signature gene set for diagnosis of MASLD stages, we established machine learning modeling with feature selection (Fig. 5A). To prevent bias in the feature selection process, we randomly divided the samples in our cohort: 70% were assigned to a training set and 30% to a testing set. We started with 14,396 genes and robust scaling was processed to individually normalize the expressions (Supplementary Fig. 4). Then, redundant features were eliminated by

https://doi.org/10.3350/cmh.2023.0449

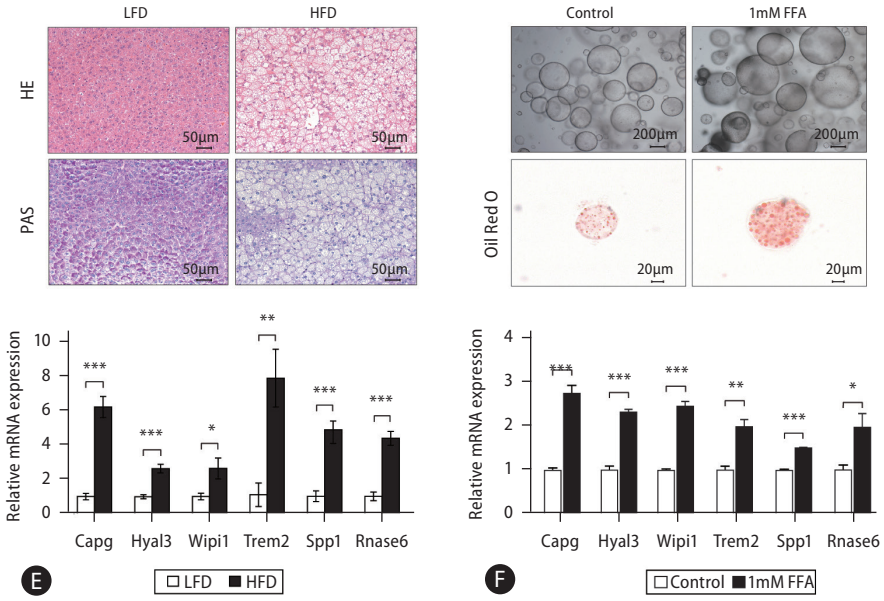


Figure 6. Continued.

repeating linear SVM modeling until less than 1,500 features with optimum coefficient scores remained. The ~1,500 selected features were placed in a testing set to evaluate their accuracy through RBF kernel model with optimal parameters. Finally, we established 20 machine learning models for MASLD-stage discrimination and found 16 models with over 80% accuracy (ACC) (Supplementary Fig. 5A).

Next, to identify signature genes from the selected features through machine learning modeling, we first asked whether there are similarities between the selected features (Fig. 5B). We found that 203 features were shared across 16 individual models, and we designated these "stacked features". Then, we looked at the features shared between the 203 stacked features and the 1,955 MASLD-associated features obtained from multi-omics analysis. We selected 64 features for further analysis and used them to discover an optimal combination of signature genes by generalized linear regression model (GLM) (Fig. 5C). First, after measuring the accuracy of the 64 features independently, we ranked them by an accuracy score. *CAPG* had the highest accuracy score (ACC=0.82) (Supplementary Fig. 5B and Supplementary Table 6). Then, we tried to identify the genes that gave the highest accuracy when paired with *CAPG*. This process was repeated with one feature after another, considering only those features that maintained a combined accuracy as high as the other models with more features (Fig. 5D and Supplementary Table 7). We found that the accuracy of combined gene set increased as features were added to it, but a combination of 6 genes saturated at the highest level of accuracy. In this way, we identified a set of 6 signature genes—*CAPG*, *HYAL3*, *WIPI1*, *TREM2*, *SPP1*, and *RNASE6*—that yielded the highest accuracy in MASLD stage discrimination. We improved the discriminability of steatosis and MASH samples by applying only the 6 signature genes compared with either data from whole transcriptome or 1,393 DEGs (Supplementary Fig. 6). Moreover, we confirmed that utilizing a set of genes enhances the ability to distinguish MASLD stages compared to individual genes (Fig. 5E), as well as non-invasive markers, such as non-alcoholic fatty liver disease (NAFLD) fibrosis score, FIB-4, and Hepatic Steatosis Index (HSI) (Fig. 5F).[20-22] Together, we propose that the 6 signature genes identified using machine learning modeling are essential molecular markers for assessing MASLD progression.

Application of the signature gene set to liver disease

To determine whether signature genes can be applied to the full spectrum of MASLD-associated disease and related histological features, we calculated its accuracy in diagnosing

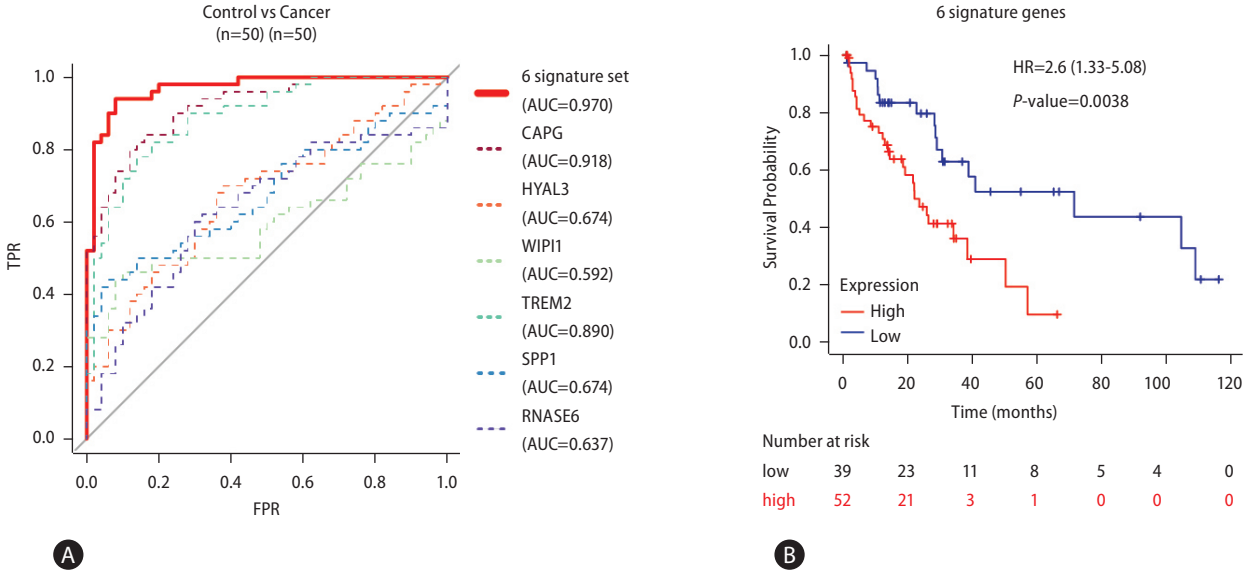


Figure 7. Validation of the signature gene set in HCC. (A) ROC curve plots illustrating the ratio between the TPR and FPR of GLM designed with the signature genes in predicting the status of an independent cohort of samples for control (n=50) and liver cancer (n=50) (6 signature set *P*-value=2.66E-16; CAPG *P*-value=3.00E-13; HYAL3 *P*-value=1.34E-03; WIPI1 *P*-value=5.68E-02; TREM2 *P*-value=9.64E-12; SPP1 *P*-value=1.34E-03; RNASE6 *P*-value=9.91E-01). (B) Kaplan-Meier survival plots showing the survival rates according to the expression levels of the signature genes in liver cancer. HCC, hepatocellular carcinoma; ROC, receiver operating characteristic; TPR, true positive rate; FPR, false positive rate; GLM, generalized linear regression model.

an independent cohort of 216 samples comprising 10 normal, 51 steatosis, and 155 MASH samples (Fig. 6A, GSE135251) and a cohort of 78 samples comprising 6 normal and 72 MASLD samples, providing information on steatosis, inflammation, ballooning hepatocyte, and liver fibrosis stage (Fig. 6B, GSE130970).[10,17] When we plotted ROC curve plots, we found that the signature gene set discriminates between steatosis and MASH (F0-F4) with an AUC score of 0.795 (Fig. 6A). Since MASH samples were graded F0 to F4 according to disease progression, we further analyzed the early-stage MASH groups (F0-2) and the late-stage MASH groups (F3-4). When predicting the groups, steatosis samples from the early-stage MASH (F0-2) samples that were relatively close in disease progression, the performance was still high accuracy (AUC=0.767). Further, in distinguishing steatosis from late-stage MASH (F3-4), which show significantly different levels of disease progression, the signature gene set predicted with high AUC score (AUC=0.863). Next, we were interested on the possibility to extend the coverage of signature gene set from normal to whole spectrum of MASLD. By applying the combination gene set of signature genes, it was possible to distinguish between normal and whole MASLD with very precisely (AUC=0.968). Also, normal and steatosis tissues (AUC=0.947), and normal and MASH tissues (AUC=0.979) showed highly accurate results. Furthermore, we confirmed that the signature gene set accurately distinguished the degree of lobular inflammation (AUC=0.931) and steatosis grade (AUC=0.943) (Fig. 6B). Although the AUC for distinguishing cytological ballooning was about 0.765, the accuracy between fibrosis stages was over 0.838, with an AUC value of 0.857 confirmed for the NAFLD activity score. The expression level of each signature genes was confirmed for normal, steatosis, and MASH and as expected, their expression levels significantly increased with progression through the various stages of MASLD (Fig. 6C) and subgroups based on histological features (Fig. 6D). These results suggest that the discriminatory capacity of the signature gene set for distinguishing different stages of MASLD is comparable to that of histological features.

We further examined the expression levels of the signature genes in an *in vivo* model fed a HFD (Fig. 6E) and in hepatic organoids treated with 1 mM FFA (Fig. 6F). In the *in vivo* model fed a HFD, signature gene levels were significantly increased, consistent with a remarkable accumulation of fat in the liver when compared to controls (Fig. 6E). Moreover, we observed a similar increase in signature gene expression in

https://doi.org/10.3350/cmh.2023.0449

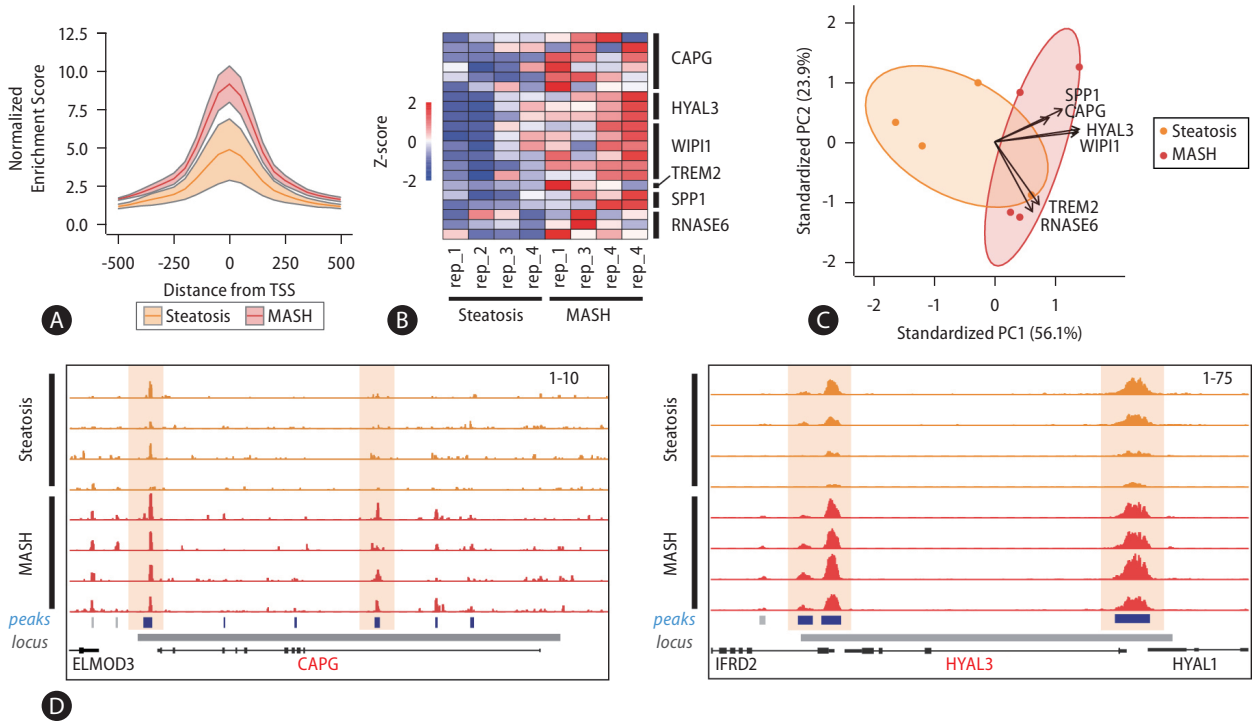


Figure 8. Altered chromatin accessibility of signature genes in MASLD progression. (A) Density plot of chromatin accessibility in the promoter regions of MASH-enriched genes. (B) Heatmap showing enrichment of open chromatin structures in regions associated with the signature genes scaled according to their z-score. (C) PCA plot representing the ability of chromatin accessibility status to discriminate MASLD stages. (D) Snapshots showing increased chromatin accessibility at open chromatin regions annotated to *CAPG* and *HYAL3* in MASH samples compared to steatosis samples. MASLD, metabolic dysfunction-associated steatotic liver disease; MASH, metabolic dysfunction-associated steatohepatitis; PCA, principal component analysis.

organoids induced to accumulate lipid by treatment with 1 mM FFAs (Fig. 6F and Supplementary Fig. 7). These results demonstrate that our signature gene set not only differentiates steatosis from MASH in MASLD progression, but also normal tissue from steatosis. This means it can be used in the early-stage detection of MASLD.

Next, we asked whether the signature gene set could be applied to the detection of liver cancer—which often follow MASLD (Fig. 7). RNA-seq data from liver cancer patients reported in a previous study (GSE77314) were re-analyzed to validate the combination set of signature genes.[23] The accuracy to distinguish between control and cancer was calculated by GLM. Soundingly, the accuracy between control and liver cancer was exceedingly high (ACC=0.970, Fig. 7A). In addition, we found high expression of signature genes showed correlation with poor overall liver cancer survival (Fig. 7B). Taken together, changes in signature gene expression can distinguish not only MASLD progression but also normal tissue from liver cancer. This indicates that our set of 6 signature genes can be used as biomarkers for the full spectrum of MASLD-associated disease.

Chromatin accessibility of signature genes

Since chromatin accessibility contributes strongly to gene expression, we further examined changes in chromatin accessibility at signature gene loci by analyzing ATAC-seq on representative steatosis (n=4) and MASH (n=4) samples (PRJNA725028, Fig. 8).[24] Because the signature genes are also MASH-enriched genes, we first investigated the accessibility status for the promoters of MASH-enriched genes. We found accessibility enrichment at these promoters was significantly increased in MASH samples (Fig. 8A). Furthermore, we confirmed that the enrichment of open chromatin regions at signature gene loci was remarkably increased in MASH (Fig. 8B). We also estimated the combination of the chromatin accessibility scores for the signature genes using PCA and found that also could predict disease progression (Fig. 8C). Figure

8D illustrated the increased enrichment of open chromatin regions at *CAPG* and *HYAL3* loci in MASH compared to steatosis samples. These results indicate both signature gene expression and chromatin accessibility can act as biomarkers for MASLD progression.

DISCUSSION

This study identified MASLD-associated features through integrative genomic, epigenomic, and transcriptomic analyses of samples from 134 MASLD patients. We used machine learning modeling to select from these MASLD-associated features those that could be used to accurately distinguish the stages of MASLD, and then we validated this signature gene set in independent cohorts of MASLD and liver cancer patients. Thus, our results provide diagnostic biomarkers that can accurately discriminate the various stages of MASLD-associated disease.

As big data technologies continue to emerge, machine learning and artificial intelligence (AI) are increasingly being applied to diagnose various human diseases and make decisions regarding their treatment.[25] In previous studies, histological images and/or clinical information have been used as inputs for deep learning or machine learning models designed to predict disease progression. In addition, MASLD researchers have used histological images to predict fibrosis scores in MASH patients and clinical information to distinguish healthy patients from those suffering from MASH.[26,27] Recently, the researchers tried to apply machine learning for MASLD study.[28] One study used lipidomics data and machine learning to detect MASLD and other study used public data (NIDDK NAFLD data and Optum data) to predict MASH.[29,30] Although they had delivered interesting results, the possibility of clinical application may be limited because of either limited data source (lipidomics only or no omics data) or poor AUC (model with AUC 0.82 or 0.76). In our machine learning modeling approach using molecular features, we identified a signature gene comprising *CAPG, HYAL3, WIPI1, TREM2, SPP1,* and *RNASE6*, which can discern the various stages of MASLD with high accuracy (Fig. 6A, normal vs MASLD AUC=0.968; normal vs. MASH AUC=0.979). Additionally, the signature gene set demonstrated high accuracy in discriminating between histological feature-based subgroups related to MASLD, achieving effective performance (Fig. 6B, AUC=0.931 for lobular inflammation; AUC=0.943 for steatosis grade; AUC=0.838 for fibrosis stage). We also confirmed that the diagnostic performance of the signature gene set could accurately distinguish disease stages with high accuracy across various subgroups associated with MASLD, including obesity, PNPLA3 mutation, and diabetes, which are known to have close connections with MASLD (Supplementary Fig. 8). These results indicate that the signature gene set identified in this study could be applied to various subgroups related to MASLD, demonstrating its potential as a diagnostic marker.

There is no uncertainty regarding the distinct functional roles of individual genes within the signature gene set in human diseases, as their expression escalates with the progression of MASLD. However, the degree of expression alterations for individual genes exhibits variability among different subgroups associated with MASLD (Fig. 6C and 6D), and the diagnostic capabilities of individual genes diverge (Fig. 5E). This emphasizes the need for a signature gene set, rather than relying on individual genes, to diagnose the stage of the disease. Using the signature gene set to assess disease stages in diverse subgroups of MASLD yielded the highest accuracy (Supplementary Fig. 9), surpassing that of non-invasive assessments (Fig. 5F). Furthermore, we explored the potential use of the signature genes as non-invasive markers and confirmed their ability to discriminate with an AUC of 0.76 between normal and MASLD in cell-free RNAs in blood (Supplementary Fig. 10). This highlights the superior precision achieved with the signature gene set in evaluating MASLD progression.

In summary, using a multi-omics approach coupled with feature selection via machine learning modeling, we identified a signature gene set that can accurately predict the stages of MASLD. We found this signature gene set can be applied to the full MASLD spectrum, from normal tissue to MASLD-related cancer. Our current understanding of this signature gene set has provided markers for the diagnosis of MASLD, but further study will be required for clinical application with larger patient cohort and functional analysis of signature genes.

Authors' contribution

S.O. contributed to the analysis of WGS, WES, WGBS, and total RNA-seq and establishment of machine learning modeling and interpretation of "omics" data; Y-H.B., S-Y.H., J-S.J., J-H.C., Y-H.R., S-W.L., G-B.C. contributed to interpretate clini-

https://doi.org/10.3350/cmh.2023.0449

cal information and patients enrollment; S.J. and S.Y. contributed to the analysis and interpretation of the data; Y-S.L., Y-H. B., Y.S.L., and G.P. contributed to sample acquisition, sample processing, quality control, and data generation; S.H. contributed to the functional analysis; B.K. and W.K. contributed to the ATAC-sequencing; K.H.Y., R.H.S., Y-S.L., and J.H.P. conceived and designed the study; S.O., S.J., S.Y., and K.H.Y. drafted the manuscript; all authors read and approved the manuscript.

Acknowledgements

We would like to thank Dr. Keun Il Kim (Sookmyung Women's University) for kindly providing tissues from in vivo model with high fat diet (HFD).

This research was supported by the Collaborative Genome Program for Fostering New Post-Genome Industry under the National Research Foundation (NRF) and funded by the Ministry of Science and ICT (MIST) (NRF-2017M3C9A6044519 to K.H.Y, NRF-2017M3C9A6044517 to Y-SL, NRF-2017M3C9A6044199 to R.H.S, 2022M3A9B6017654 to K.H.Y and J.H.P).

Conflicts of Interest

The authors have no conflicts to disclose.

SUPPLEMENTARY MATERIAL

Supplementary material is available at Clinical and Molecular Hepatology website (http://www.e-cmh.org).

The data are available from the Korean Nucleotide Archive (https://www.kobic.re.kr/kona/; Accession ID: PRJKA210057).

REFERENCES

1. Eslam M, Sanyal AJ, George J; International Consensus Panel. MAFLD: A consensus-driven proposed nomenclature for metabolic associated fatty liver disease. Gastroenterology 2020;158:1999-2014.e1.
2. Badmus OO, Hillhouse SA, Anderson CD, Hinds TD, Stec DE. Molecular mechanisms of metabolic associated fatty liver disease (MAFLD): functional analysis of lipid metabolism pathways. Clin Sci (Lond) 2022;136:1347-1366.
3. Yew KC, Wong SH, Wong VW, Oon HH. Letter regarding "Waiting for the changes after the adoption of steatotic liver disease". Clin Mol Hepatol 2024;30:118-120.
4. Nassir F, Rector RS, Hammoud GM, Ibdah JA. Pathogenesis and prevention of hepatic steatosis. Gastroenterol Hepatol (N Y) 2015;11:167-175.
5. Mazzolini G, Sowa JP, Atorrasagasti C, Kücükoglu Ö, Syn WK, Canbay A. Significance of simple steatosis: An update on the clinical and molecular evidence. Cells 2020;9:2458.
6. Kim GA, Moon JH, Kim W. Critical appraisal of metabolic dysfunction-associated steatotic liver disease: Implication of Janus-faced modernity. Clin Mol Hepatol 2023;29:831-843.
7. Younossi ZM. Non-alcoholic fatty liver disease - A global public health perspective. J Hepatol 2019;70:531-544.
8. Maurice J, Manousou P. Non-alcoholic fatty liver disease. Clin Med (Lond) 2018;18:245-250.
9. Feng G, Valenti L, Wong VW, Fouad YM, Yilmaz Y, Kim W, et al. Recompensation in cirrhosis: unravelling the evolving natural history of nonalcoholic fatty liver disease. Nat Rev Gastroenterol Hepatol 2024;21:46-56.
10. Hoang SA, Oseini A, Feaver RE, Cole BK, Asgharpour A, Vincent R, et al. Gene expression predicts histological severity and reveals distinct molecular profiles of nonalcoholic fatty liver disease. Sci Rep 2019;9:12541.
11. Alkhouri N, McCullough AJ. Noninvasive diagnosis of NASH and liver fibrosis within the spectrum of NAFLD. Gastroenterol Hepatol (N Y) 2012;8:661-668.
12. Li G, Zhang X, Lin H, Liang LY, Wong GL, Wong VW. Non-invasive tests of non-alcoholic fatty liver disease. Chin Med J (Engl) 2022;135:532-546.
13. Oh S, Jo Y, Jung S, Yoon S, Yoo KH. From genome sequencing to the discovery of potential biomarkers in liver disease. BMB Rep 2020;53:299-310.
14. Sliz E, Sebert S, Würtz P, Kangas AJ, Soininen P, Lehtimäki T, et al. NAFLD risk alleles in PNPLA3, TM6SF2, GCKR and LYPLAL1 show divergent metabolic effects. Hum Mol Genet 2018;27:2214-2223.
15. Yoo T, Joo SK, Kim HJ, Kim HY, Sim H, Lee J, et al. Disease-specific eQTL screening reveals an anti-fibrotic effect of AGXT2 in non-alcoholic fatty liver disease. J Hepatol 2021;75:514-523.
16. Atanasovska B, Rensen SS, Marsman G, Shiri-Sverdlov R, Withoff S, Kuipers F, et al. Long non-coding rnas involved in progression of non-alcoholic fatty liver disease to steatohepatitis. Cells 2021;10:1883.
17. Govaere O, Cockell S, Tiniakos D, Queen R, Younes R, Vacca M, et al. Transcriptomic profiling across the nonalcoholic fatty liver disease spectrum reveals gene signatures for steatohepatitis

and fibrosis. Sci Transl Med 2020;12:eaba4448.
18. Loomba R, Gindin Y, Jiang Z, Lawitz E, Caldwell S, Djedjos CS, et al. DNA methylation signatures reflect aging in patients with nonalcoholic steatohepatitis. JCI Insight 2018;3:e96685.
19. Ma J, Nano J, Ding J, Zheng Y, Hennein R, Liu C, et al. A peripheral blood DNA methylation signature of hepatic fat reveals a potential causal pathway for nonalcoholic fatty liver disease. Diabetes 2019;68:1073-1083.
20. Sterling RK, Lissen E, Clumeck N, Sola R, Correa MC, Montaner J, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology 2006;43:1317-1325.
21. Lee JH, Kim D, Kim HJ, Lee CH, Yang JI, Kim W, et al. Hepatic steatosis index: a simple screening tool reflecting nonalcoholic fatty liver disease. Dig Liver Dis 2010;42:503-508.
22. Angulo P, Hui JM, Marchesini G, Bugianesi E, George J, Farrell GC, et al. The NAFLD fibrosis score: a noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology 2007;45:846-854.
23. Liu G, Hou G, Li L, Li Y, Zhou W, Liu L. Potential diagnostic and prognostic marker dimethylglycine dehydrogenase (DMGDH) suppresses hepatocellular carcinoma metastasis in vitro and in vivo. Oncotarget 2016;7:32607-32616.
24. Kang B, Kang B, Roh TY, Seong RH, Kim W. The chromatin accessibility landscape of nonalcoholic fatty liver disease progression. Mol Cells 2022;45:343-352.
25. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2:230-243.
26. Heinemann F, Birk G, Stierstorfer B. Deep learning enables pathologist-like scoring of NASH models. Sci Rep 2019;9:18454.
27. Palekar NA, Naus R, Larson SP, Ward J, Harrison SA. Clinical model for distinguishing nonalcoholic steatohepatitis from simple steatosis in patients with nonalcoholic fatty liver disease. Liver Int 2006;26:151-156.
28. Castañé H, Baiges-Gaya G, Hernández-Aguilera A, Rodríguez-Tomàs E, Fernández-Arroyo S, Herrero P, et al. Coupling machine learning and lipidomics as a tool to investigate metabolic dysfunction-associated fatty liver disease. A general overview. Biomolecules 2021;11:473.
29. Hou C, Feng W, Wei S, Wang Y, Xu X, Wei J, et al. Bioinformatics analysis of key differentially expressed genes in nonalcoholic fatty liver disease mice models. Gene Expr 2018;19:25-35.
30. Docherty M, Regnier SA, Capkun G, Balp MM, Ye Q, Janssens N, et al. Development of a novel machine learning model to predict presence of nonalcoholic steatohepatitis. J Am Med Inform Assoc 2021;28:1235-1241.