JBC RESEARCH ARTICLE

An artificial intelligence optimized hepatic differentiation unveils NR5A2 and AP-1 transcriptional regulation in hepatic maturation

Received for publication, June 5, 2025, and in revised form, February 25, 2026 Published, Papers in Press, April 8, 2026
https://doi.org/10.1016/j.jbc.2026.111435

Zijun Huo[1,2,‡], Jian Tu[2,3,4,‡], Wei-Lei Yang[5], Mo-Fan Huang[2,6], Ruoyu Wang[6,7], Chih-Wei Chu[2,6], An Xu[2], Yao Yu[8], Tara N. Tavakol[9], Mikal Kizilbash[9], Megan E. Fisher[2,6], Yu-Wen Huang[2], Dandan Zhu[2], Trinh T. T. Phan[2], Rachel Shoemaker[2,6], Ya-Wen Chen[9,10,11,12,13], Yang Zhang[14], Chad D. Huff[2,8], Shih-Yu Chen[5], Tien-Jen Liu[5], Haipeng Xiao[1], Dung-Fang Lee[2,6,15,*], and Ruiying Zhao[2,*]

From the [1]Department of Endocrinology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, P. R. China; [2]Department of Integrative Biology and Pharmacology, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, Texas, USA; [3]Department of Musculoskeletal Oncology, and [4]Guangdong Provincial Key Laboratory of Orthopedics and Traumatology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, P. R. China; [5]AIxMed Inc. Santa Clara, California, USA; [6]The University of Texas MD Anderson Cancer Center UTHealth Houston Graduate School of Biomedical Sciences, Houston, Texas, USA; [7]Department of Biochemistry and Molecular Biology, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, Texas, USA; [8]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA; [9]Department of Otolaryngology, [10]Department of Cell, Developmental and Regenerative Biology, [11]Black Family Stem Cell Institute, [12]Institute for Airway Sciences, and [13]Center for Epithelial and Airway Biology and Regeneration, Icahn School of Medicine at Mount Sinai, New York, New York, USA; [14]College of Science, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong, China; [15]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA

Reviewed by members of the JBC Editorial Board. Edited by Todd R. Graham

The generation of hepatocyte-like cells (HLCs) from human pluripotent stem cells (hPSCs) holds great promise for drug discovery and cell-based therapy for liver disease. However, current differentiation protocols are complicated and unstable, and the underlying gene regulatory mechanisms of hepatic differentiation remain incompletely defined. Here, we developed a machine learning-based artificial intelligence (AI) tool using phase-contrast images of hepatic progenitor cells (HPCs), which are essential for generating HLCs. The AI tool significantly improves the success rate of hepatic differentiation without the need for immunostaining or lineage tracing. By optimizing the methodology, we achieved an impressive purity of 90 to 95% for HLCs derived from hPSCs, aided by the AI algorithm. Through further investigating transcriptomes and epigenomic changes, we discovered the pivotal roles of nuclear receptor subfamily 5 group A member 2 and activator protein-1 transcription factors in regulating the maturation of hepatocytes. Single-cell RNA sequencing demonstrated the upregulation of nuclear receptor subfamily 5 group A member 2 and activator protein-1 during hepatic differentiation. Importantly, mutation analysis and tumorigenesis assays confirmed the safety of this modified hepatic differentiation protocol. This work highlights the potential of combining AI algorithm and computational genomics to facilitate development of lineage differentiation and molecular mechanism study.

The hepatocyte is a vital cell type responsible for numerous physiological processes in the body, including protein synthesis, lipid and carbohydrate metabolism, bile acid production, blood clotting factor synthesis, and detoxification of xenobiotic substances and pharmaceuticals (1, 2). While the hepatocyte has the ability to regenerate *in vivo*, primary human hepatocytes (PHHs) have limited ability to be expanded *in vitro* (3), resulting in a shortage of supply. To address this issue, researchers have established protocols for the directed differentiation of human pluripotent stem cells [hPSCs, including human embryonic stem cells and human induced pluripotent stem cells (iPSCs)] into hepatocyte-like cells (HLCs) (4–11). These HLCs can be produced on a large scale, making them an ideal tool for basic and translational science. Many of the current protocols involve a multistep process, where hPSCs are first committed to a definitive endoderm (DE) stage using activin A combined with or without WNT signaling activator CHIR99021, a glycogen synthase kinase 3β inhibitor. The definitive endoderm (DE) cells (DECs) are then induced to differentiate into hepatic progenitor cells (HPCs) using specific factors, such as BMP4 and FGF2, that promote hepatic lineage. Finally, the cells are matured using oncostatin M (OSM) induction to generate HLCs that express hepatocyte-specific markers. These hepatic differentiation protocols developed for hepatic differentiation offer a valuable tool for studying liver physiology, disease pathology, and drug discovery and may lead to the development of more effective cell-based therapies for liver disease.

‡ These authors contributed equally to this work.
* For correspondence: Ruiying Zhao, ruiying.zhao@uth.tmc.edu; Dung-Fang Lee, dung-fang.lee@uth.tmc.edu.

Despite the significant progress achieved in hepatic differentiation protocols for generating albumin (ALB)-positive HLCs, the maturation and purity of these cells remain suboptimal in most published studies. While established protocols can achieve an efficiency of over 70% in generating HLCs (8), the multistep differentiation process often results in a low success rate. To confirm the successful directed differentiation of hPSCs into HPCs and HLCs, researchers typically rely on microscopic observation of morphological changes during *in vitro* hepatic differentiation. However, this method is laborious, time-consuming, and prone to interobserver variability, which can limit the progress of stem cell research and its clinical applications. Moreover, the molecular mechanisms underlying hepatocyte maturation are still not fully understood, further hindering the development of more efficient and reliable differentiation protocols.

Recent advances in machine learning have demonstrated its potential across a wide range of applications in artificial intelligence (AI), particularly in computer vision and image analysis (12). In these approaches, raw image data are provided as input, processed through multiple hidden layers of artificial neurons that extract hierarchical features and ultimately generate an output such as the classification or prediction. Convolutional neural networks (CNNs), a widely used model in image analysis, learn spatial features from training images and are optimized using a training set, with performance fine-tuned on a validation set and independently evaluated on a testing set. The use of such AI-based computational tools holds great promise for advancing modern medicine by offering advantages such as automation, high throughput, accuracy, and reproducibility. In the context of cell lineage differentiation, several studies have shown that AI technology can be leveraged to detect abnormal differentiation at an early stage, thereby saving time, labor, and cost while also promoting reproducible results (13–15). To date, only a limited number of AI-related tools have been developed or validated for the systematic monitoring of hPSC-based hepatocyte differentiation (16, 17).

The activator protein-1 (AP-1) family is a group of transcription factors that includes subfamilies JUN, FOS, ATF, and MAF (18, 19). These factors belong to the class of basic leucine zipper transcription factors that form dimers through hydrophobic leucine residues and interact with DNA promoters of target genes using positively charged amino acids. The JUN family proteins, such as JUN, JUNB, and JUND, can form homodimers within their own family or heterodimers with FOS family proteins, including c-FOS, FOSB, FOSL1, and FOSL2. In contrast, FOS family proteins only heterodimerize with JUN family members to form transcriptionally active complexes (20). AP-1 plays a critical role in various biological activities, such as differentiation, proliferation, apoptosis, cell migration, and transformation (21). Knockout or transgenic mouse models have shown that AP-1 is essential for cell differentiation, as perturbation of AP-1 gene expression in mice leads to impairments in the development of various organs, including bone, brain, liver, heart, eye, T-cell, and testis (22). However, the precise molecular mechanism through which AP-1 influences human hepatic differentiation remains unclear.

Nuclear receptor subfamily 5 group A member 2 (NR5A2), also known as liver receptor homolog 1, operates as a vital transcription factor with diverse cellular roles, maintaining tissue homeostasis and orchestrating the intricate regulatory network governing embryonic development and physiological balance in various organs (23). NR5A2's absence allows embryos to progress normally beyond the 2-cell stage, highlighting its pivotal role in regulating zygotic genome activation and the initial lineage segregation during early mouse development (24). Associated primarily with embryonic development, NR5A2 crucially determines cell lineage, particularly in the formation and functionality of the liver, pancreas, and intestines (25, 26). In the liver, NR5A2 contributes to metabolic regulation by binding DNA as a monomer, influencing hepatic gene expression, including key enzymes in bile acid biosynthesis like cholesterol 7 alpha-hydroxylase, thereby impacting cholesterol and bile acid metabolism (23, 27). While NR5A2's role in liver development is suggested, its essentiality in hepatocyte maturation remains unclear. Notably, the relationship between NR5A2 and AP-1 has not been explored, despite the potential roles of both in liver development in rodent models.

In this study, we employed a deep learning–based (CNN) computer vision algorithm to analyze bright-field microscopy images acquired at various stages of differentiation, with a focus on the HPC stage as a critical checkpoint in hepatic differentiation. Using this AI algorithm, we developed an innovative protocol that produces HLCs from hPSCs safely and efficiently. Using this novel method, we combined parallel sequencing methods, including RNA sequencing (RNA-seq) and assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq), to gain insights into the basic transcriptional regulation of the hepatocyte maturation process and uncover the critical roles of NR5A2 and AP-1 transcription factor in regulating hepatogenesis. In summary, our study optimizes hepatic differentiation using a machine learning-based AI tool, with potential implications for understanding human hepatogenesis and providing valuable resources for future regenerative medicine and pharmacological drug safety testing.

Results

Applying machine learning–based differentiation to assess and enhance the efficiency of hepatic differentiation

To produce clinically and scientifically valuable HLCs from hPSCs, we employ a machine learning–based AI algorithm to enhance hepatocyte generation. This strategy addresses the variability in efficiencies seen in established protocols and mitigates the common challenge of a low success rate during the multistep differentiation process. Initially, we conducted a comprehensive analysis of the existing knowledge from published hepatic differentiation studies (4–10) to establish the foundational protocol for hepatic differentiation. To differentiate hPSCs to DE, we used activin A throughout the entire

endoderm induction process due to its critical role in DE differentiation (28). Previous studies have demonstrated that the use of the glycogen synthase kinase 3β inhibitor CHIR99021 for the first 24 h of differentiation is beneficial for DE differentiation (29). Therefore, CHIR99021 was added to the differentiation medium during the first 24 h. Inhibition of the insulin signal transducer PI3 kinase has also been found to promote DE differentiation (30). Thus, the B-27 supplement without insulin was used throughout the entire DE differentiation process to increase DE differentiation efficiency. In addition, both BMP4 and FGF2 were reported as necessary factors for hepatic specification (31, 32). Thus, we continued culturing these differentiated cells in BMP4/FGF2-containing medium. Also, HGF was reported to increase the success rate of turning HPCs into HLCs (33), which was used in our study. To facilitate HLC maturation, we re-evaluated the chemicals and growth factors in the previous established methods (4–10) and found that OSM (34), an interleukin-6 family cytokine, is the main essential factor beneficial for HLC maturation in our modified method.

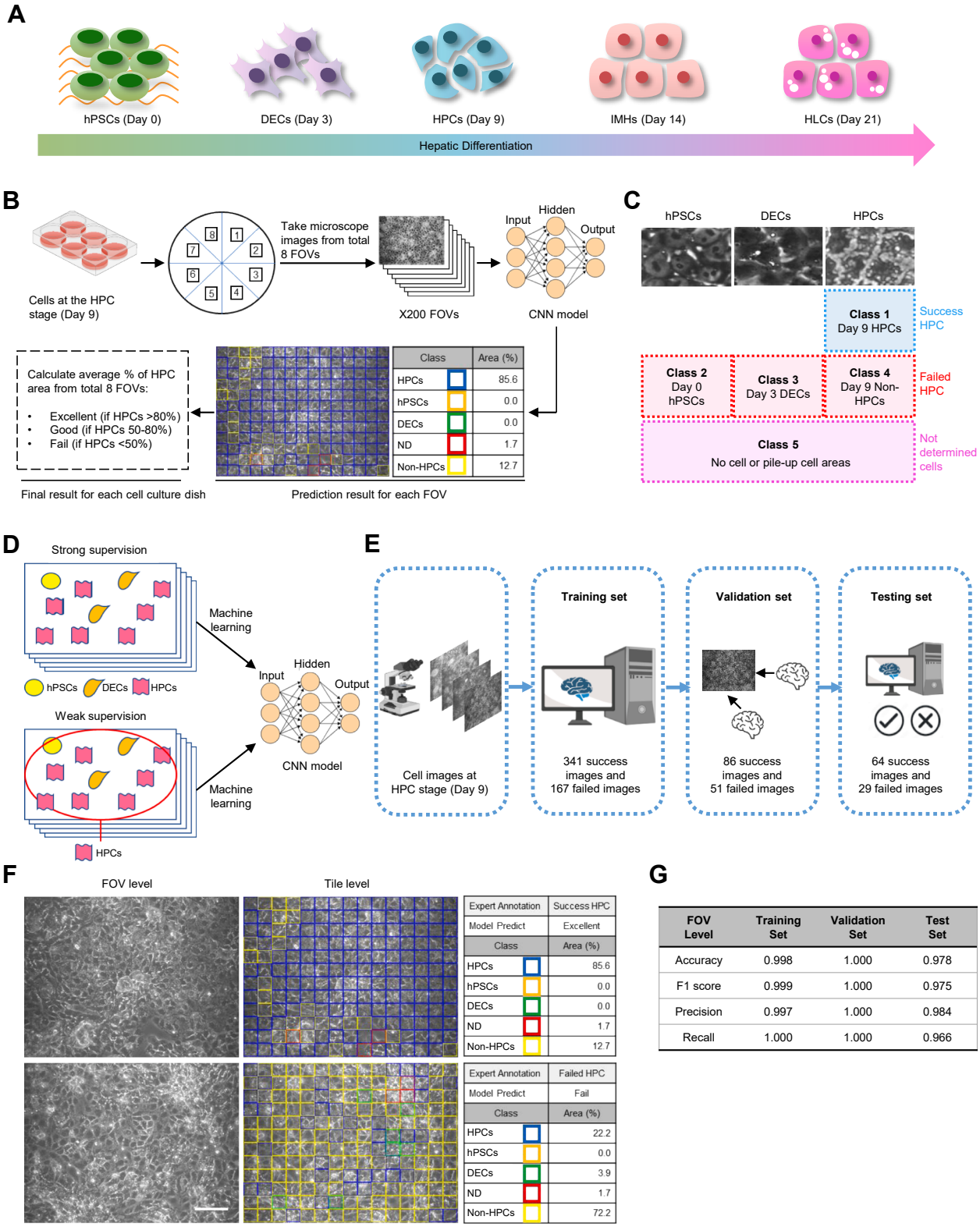
Based on our analysis, we started with the four-stage differentiation protocol (Fig. 1*A*) and then modified and refined medium components using a machine learning-based AI algorithm to monitor cell morphology of hPSC-derived HPCs, which are the most critical differentiation steps for hPSC-based hepatic differentiation. Specifically, we optimized the culture conditions to improve the efficacy and consistency of the differentiation process. The efficiency of the initial stage of hepatic differentiation, from hPSCs to DE, then to HPCs, is critical for the success of subsequent differentiation into HLCs. However, accurately identifying differentiated HPC cells can be challenging due to the heterogeneous nature of the cell population. We notice that a low percentage of HPCs may exhibit the desired molecular and functional characteristics, while others may remain undifferentiated or exhibit an HPC-like morphology. They can compromise the reliability of differentiation results, leading to wasted time and resources in the continued efforts to achieve further differentiation stages, suggesting specialized assays and techniques are needed to accurately evaluate the differentiation status of the cell population. Therefore, the AI algorithm, based on machine learning, assists in analyzing morphological changes of differentiated HPCs in bright-field microscopy images. It was employed to quantitatively assess differentiation efficiency during the established four-stage hepatic differentiation process. While the model was not applied to test multiple experimental conditions, it provided an objective and reproducible measure of differentiation quality that could support future optimization of hepatic induction protocols (Fig. 1*B*). We obtained field-of-view images (FOVs) at three differentiation stages, including 12 FOVs of hPSCs on Day 0, 413 FOVs of DECs on Day 3, and 738 FOVs of HPCs on Day 9 (Fig. 1*B*). FOVs acquired through bright-field microscopy were carefully annotated and classified into successful, failed, or not determined categories to create a training dataset for the CNN algorithm. By systematically classifying the annotated images, the CNN algorithm can effectively learn the distinguishing features and variations associated with different categories, enhancing its ability to make accurate predictions on new, unseen images. To prepare the datasets for developing the AI algorithm, two experienced stem cell researchers annotated and classified the morphologies of hPSCs (Class 2), DECs (Class 3), HPCs (Class 1), and non-HPC cells (Class 4) in a total of 1163 FOVs obtained from the three differentiation stages (Fig. 1*C*).

The CNN algorithm was next trained using both strong and weak supervision models. Strong supervision involves precise annotations, leading to accurate predictions, while weak supervision, with less precise annotations, still contributes to effective learning and predictions despite inherent noise. The independent researchers also determined a differentiation result (successful or failed HPC differentiation) for each image, which served as the ground truth for comparison with the prediction of the AI algorithm. To reduce the researcher's workload for annotation, we applied "weak supervised learning" in the AI algorithm training process, which required annotating areas with dominant HPC distribution instead of every HPC (Fig. 1*D*). The collected images were carefully annotated and classified, either through strong or weak supervision techniques, to create a labeled training dataset to train the CNN algorithm. Next, the trained algorithm was evaluated using a separate validation dataset to assess its accuracy and performance. Finally, the algorithm's performance was evaluated on an independent testing dataset to assess its generalization and robustness. The workflow of the AI algorithm development is shown in Figure 1*E*.

The initial AI algorithm was trained using a "training" dataset comprising 341 images of successful and 167 images of failed HPC differentiation. During the training process, we monitored the performance of a smaller, separate "validation" dataset consisting of 86 successful and 51 failed HPC differentiation images to determine when to terminate the training process and complete the AI algorithm development. Finally, we used a "test" dataset containing 64 successful and 29 failed HPC differentiation images to evaluate the performance of the AI algorithm (Fig. 1, *E* and *F*). In the evaluation of the AI algorithm, the final model demonstrated robust and consistent performance across all three datasets. On the independent testing set, the model achieved an accuracy of 0.978, F1 score of 0.975, precision of 0.984, and recall of 0.966 (Fig. 1*G*). Similarly, high performance was observed on the training set (accuracy = 0.998) and the validation set (accuracy = 1.000), indicating strong model stability and minimal overfitting. Collectively, these performance metrics confirm the reliability of the model in distinguishing among differentiation stages. Overall, these findings support the potential application of an AI-assisted monitoring tool for stem cell culture and differentiation. Furthermore, they demonstrate the feasibility of integrating a machine learning–based algorithm to optimize hepatic differentiation protocols.

Optimizing hepatic differentiation of hPSCs into HLCs

Utilizing our machine learning–based cell morphology analysis, we optimized the four-stage hepatic differentiation process, achieving a success rate of over 90% (Fig. 2*A*). Four hPSC lines (H1, H9, HES2, and RUES2) were cultured and



FOV Level	Training Set	Validation Set	Test Set
Accuracy	0.998	1.000	0.978
F1 score	0.999	1.000	0.975
Precision	0.997	1.000	0.984
Recall	1.000	1.000	0.966

Figure 1. Analysis of hPSC-derived HPCs using a machine learning algorithm. *A*, schematic diagram depicting the four-stage protocol utilized to achieve hepatic differentiation of hPSCs into HLCs. *B*, schematic diagram illustrating the workflow for obtaining bright-field microscopy images of hPSCs, DECs, and HPCs. The large number of images captured during the experiment showcases the distinct morphological characteristics of each cell type and provides valuable input for the CNN algorithm. The researchers determined a differentiation result for each image, which served as the gold standard for comparison with the predictions made by the AI algorithm. *C*, bright-field microscopy images are annotated and categorized into successful, failed, or undetermined classes to create a CNN training dataset. This systematic classification enhances the algorithm's ability to make accurate predictions on new, unseen images by learning distinguishing features and variations. *D*, strong and weak supervision machine learning models employed to train the CNN algorithm. In the strong supervision approach, the training dataset is annotated with precise and detailed labels, providing explicit information on the desired output for each image. This enables the algorithm to learn directly from labeled examples, leading to more accurate predictions. In contrast,

differentiated into HLCs. During the first 3 days of stage 1, we observed a significant enrichment of petal-like DECs expressing SOX17 (Fig. 2*B*). In the subsequent stage, cells quickly differentiated into polygonal HPC expressing high levels of the liver transcription factor hepatocyte nuclear factor 4 alpha (HNF4A) (Fig. 2*B*). In the next two stages, immature hepatocytes (IMHs) and HLCs were identified by the expression of IMH markers alpha-fetoprotein (AFP) and ALB, respectively (Fig. 2*B*). We used qPCR to assess the dynamic changes of stage-specific marker genes and confirmed the success of our modified differentiation protocol (Fig. S1, *A*–*D*). The flow cytometry analysis of ALB-positive cells in HLCs indicated that over 95% of the HLCs expressed ALB (Fig. 2*C*, left panels), a result consistent with high ALB expression in commercial PHHs-OSZ and -JYN, as well as the liver cancer cell line PLC/PRF/5 (Fig. 2*C*, middle and right panels). Consistent with the flow cytometry analysis, ELISA results demonstrated that the levels of secreted ALB in HLCs were comparable to those observed in PHHs (Fig. 2*D*). Examination of urea synthetase activity and CYP3A4 activity, two key indicators of liver function, suggests that HLCs exhibit comparable functional metabolism but lower detoxification abilities to PHHs (Fig. 2, *E* and *F*). Furthermore, compared to PHHs, HLCs exhibit a similar ability to uptake indocyanine green (Fig. S1*E*), indicating that HLCs retain liver function for efficient substrate transport. To assess the maturation potential of the developed HPCs *in vivo*, we conducted renal capsule transplantation by introducing HPCs into the kidneys of immune-compromised NSG mice, allowing them to undergo differentiation into HLCs. Immunofluorescent staining demonstrated significant enrichment of mature hepatocyte markers, such as ALB, α-haptoglobin, and transferrin, in outgrowth transplants (Fig. 2*G*). In summary, our machine learning–based AI algorithm optimizes the efficiency and success rates of the four-stage hepatic differentiation protocol.

Transcriptome analysis identifies gene expression signatures during hepatic differentiation

The differentiation of HLCs is a complex and dynamic process that involves transcriptional and chromatin changes resulting in the development of distinct cellular identities (35). To evaluate the fidelity of our established hepatic differentiation protocol in mirroring human hepatogenesis and its ability to discern crucial gene regulatory networks governing this process, we conducted a comprehensive analysis of transcriptional changes throughout the differentiation of HLCs. mRNA samples were collected at four differentiation time points [Day 0, hPSCs (H1, H9, HES2, and RUES2); Day 3, DECs; Day 9, HPCs; and Day 21, HLCs], and their transcriptomes were analyzed using RNA-seq. Spearman's correlation analysis demonstrated that the transcription profiles of hPSCs, DECs, HPCs, and HLCs were distinctly different from each other (Fig. 3*A*). Principal component analysis (PCA) unveiled notable transcriptomic changes occurring throughout the hepatic differentiation process (Fig. 3*B*). These findings indicate that dynamic and unique transcriptional changes occur during hepatic differentiation.

Human Gene Atlas analysis demonstrated the differentiated HLCs closely resembled both human fetal liver and adult liver compared with DECs and HPCs (Fig. S2*A*), signifying a successful hepatic differentiation. KEGG pathway analysis revealed that pathways related to hepatocyte function, such as fatty acid metabolism, the cytochrome P450 pathway, and the complement and coagulation cascades, were highly enriched in HLCs (Fig. 3*C*, upper and lower panels). In contrast, pathways related to cell cycle, DNA replication, and RNA metabolism were highly enriched in hPSCs or DECs (Fig. 3*C*, lower panel). Consistent with previous studies, pluripotency genes POU5F1, LIN28A, DPPA4, SOX2, ZFP42, NANOG, and IDO1 were significantly enriched at the hPSCs (36); in contrast, transcription factors SOX17, MIXL1, GSC, GATA6, and KIT were significantly enriched at the DECs (37, 38). Moreover, numerous crucial hepatic transcription factors HNF4A, HNF1B, GATA4, FOXA2, CEBPA, TBX3, and FOXA1 were dramatically upregulated in HPCs (39) (Fig. S2*B*). Mature hepatic markers AFP, TTR, APOA2, SERPINA1, ALB, and APOA1 were significantly elevated at the mature HLCs (40) (Fig. S2*B*).

We then identified a set of 9000 genes that exhibited dynamic expression patterns, which we divided into 15 distinct clusters (Fig. 3*D*). Through an analysis of the Human Gene Atlas, we found that clusters C5 to C8 were particularly enriched for biological processes associated with liver formation (Fig. 3*E*). Consistent with this finding, clusters C5 to C7 showed a high level of enrichment for hepatocyte-related pathways, such as lysosome, fatty acid degradation, and metabolism of xenobiotics by cytochrome P450 (Fig. 3*F*, upper panel). By performing transcription factor enrichment analyses, we discovered that the targets of HNF4A, FOSL2, GATA family, and FOX family transcription factors were highly enriched in the C6 cluster during hepatic differentiation (Fig. 3*F*, lower panel), which has been reported previously (41, 42). In contrast, the targets of MYC, YY1, and FOXM1 were highly enriched in clusters C0 to C3, whose genes have been linked to pluripotency (Fig. 3*F*, lower panel).

Heatmap analysis revealed that the expression of transcription factors changed dramatically at different stages of hepatic

weak supervision techniques leverage less precise annotations to train the algorithm. Despite the inherent noise or ambiguity in the annotations, weak supervision approaches can still yield effective learning and prediction capabilities. *E*, schematic diagram illustrating the workflow of the AI algorithm development. FOVs are collected for training, validating, and testing the algorithm. *F*, FOVs and tile-level images used as input data for the CNN algorithm to predict the successful or failed differentiation of HPCs. By feeding both FOVs and tile-level images into the CNN algorithm, it can effectively analyze and learn from the varying scales and resolutions of the images, enabling the algorithm to make predictions regarding the successful or failed differentiation of HPCs based on the distinctive patterns and characteristics observed in the input images. Scale bar: 100 μm. *G*, table summarizing the excellent performance of the AI algorithm across all three datasets, as evidenced by high accuracy and F1 score, which demonstrates the AI algorithm's exceptional performance and reliability in predicting HPC differentiation outcomes. hPSCs, human pluripotent stem cells; DECs, definitive endoderm cells; HPCs, hepatic progenitor cells; IMHs, immature hepatocytes; HLCs, hepatocyte-like cells; FOV, field-of-view images; CNN, convolutional neural networks.

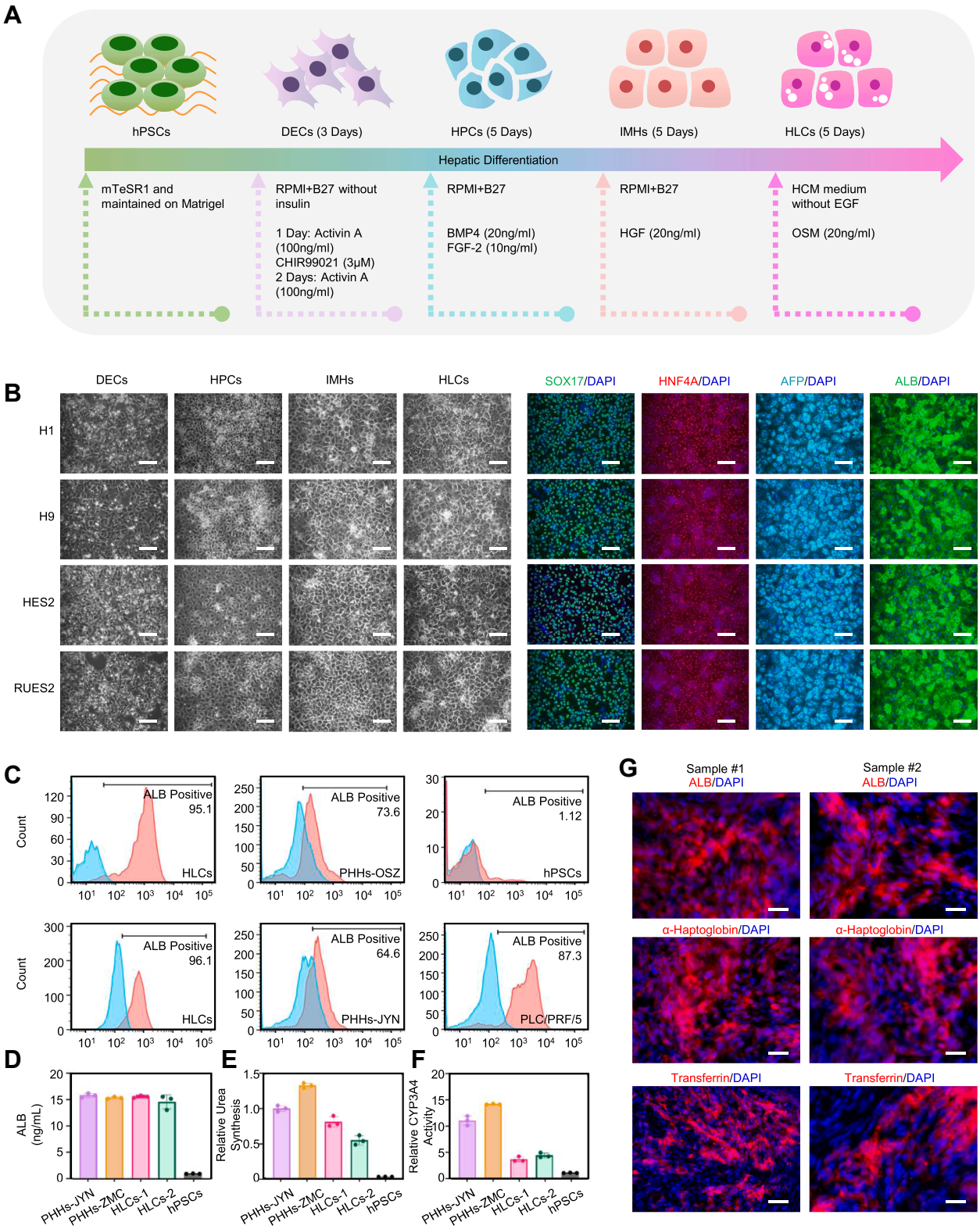


Figure 2. Developing an optimized hepatic differentiation protocol. *A*, schematic diagram illustrating the four-stage protocol employed for the successful differentiation of hPSCs into HLCs. The protocol involves sequential steps to mimic the developmental cues involved in hepatocyte formation, starting with the induction of DE differentiation, followed by specification of HPCs, differentiation into IMHs, and final maturation into functional HLCs. *B*, cell morphology and key marker gene expression during hepatic differentiation for four different hPSC lines (H1, H9, HES2, and RUES2). The left panel exhibits phase-contrast images of DECs, HPCs, IMHs, and HLCs, illustrating morphological transitions throughout each differentiation stage. The right panel demonstrates immunofluorescence staining, indicating high expression of stage-specific marker genes (SOX17 for DECs, HNF4A for HPCs, AFP for IMHs, and ALB for HLCs), validating the successful differentiation and maturation of the cells. Scale bar: 100 μm. *C*, flow cytometry analysis of ALB expression in PHHs, HLCs, H1 hPSCs, and PLC/PRF/5 cells, with ALB-positive cells shown as a percentage. More than 95% of HLCs exhibit ALB expression, comparable to the levels observed in PHHs and PLC/PRF/5 cells. In contrast, H1 hPSCs show minimal ALB expression. *D*, analysis of secreted ALB levels in

differentiation. Significantly, we noted upregulation of the transcription factors NR5A2 and AP-1 family genes in the C5 cluster and the mature HLC stage (Figs. 3*D* and S2*C*). This observation suggests a potential role for these transcription factors in mastering the hepatic differentiation process, although limited knowledge currently indicates their specific functions in HLCs. NR5A2 exhibited upregulation during the HPC and HLC stages compared to the hPSC and DEC stages. In contrast, the AP-1 family genes (JUN, JUNB, FOS, and FOSL2) showed upregulation specifically in the HLC stage compared to other stages (Fig. S2*C*). qRT-PCR confirmed that NR5A2 was enriched in the HPC and HLC stages, while the AP-1 family were enriched in the HLC stage (Fig. S2*D*). Furthermore, Gene Set Enrichment Analysis (GSEA) revealed that the AP-1 targets were enriched in the HLC stage, as compared to the HPC or hPSC stages (Fig. S2*E*). Taken together, these findings provide insights into the genetic mechanisms underlying liver development and differentiation and the potential roles of NR5A2 and AP-1 in hepatogenesis and HLC maturation.

ATAC-seq analyses identify stage-specific transcription factors during hepatogenesis

Our differentiation system provided valuable insights into the sequential differentiation events that resulted in the generation of HLCs. To gain a better understanding of the dynamic changes in the chromatin landscape during hepatic differentiation, we conducted ATAC-seq analysis on H1 and H9 hPSCs and their derived DECs, HPCs, and HLCs (Fig. S3*A*). After removing small peaks that could not be clearly distinguished from the background, we identified 79,926 highly reproducible peaks in the merged ATAC-seq dataset. PCA of ATAC-seq result showed marked changes in chromatin accessibility between different stages (Fig. 4*A*), and Spearman's correlation analysis indicated that gene expression profiles from hPSCs were distinct from those of DECs, HPCs, and HLCs (Fig. S3*B*). Integrative Genomics Viewer snapshotted the changes in chromatin accessibility of representative lineage transcription factors (NANOG for hPSCs, SOX17 for DECs, HNF4A for HPCs, and AFP for HLCs), which confirmed the successful differentiation of HLCs (Fig. S3*C*).

The dynamic changes in chromatin accessibility across different stages of hepatic differentiation were shown in the heat map (Fig. 4*B*, left). We then divided all ATAC-seq peaks into 15 clusters based on their chromatin accessibility profiles. Cluster 5 (C5) consisted of 2887 peaks that were accessible throughout the entire differentiation process, while cluster 6 (C6) contained 4421 peaks that specifically opened during the HPC and HLC stages. Interestingly, we identified a group of peaks (C9) that opened during the HLC stage and may contribute to the maturation of HLCs (Fig. 4*B*, right middle panel; and Fig. S3*D*). Peak annotation revealed that the hepatobiliary system physiology pathway, liver regeneration pathway, and liver physiology pathway were highly enriched in C5 and C9 clusters, while liver function-related pathways were highly enriched in C6 and C9 clusters (Fig. 4*B*, right bottom panel). Peaks in clusters C0, C13, and C14 regulated genes in the hPSC stage and contained binding motifs of pluripotent transcription factors such as SOX2 and OCT4. Peaks in clusters C2–C4 specifically opened during the DE stage and showed enrichment of GATA family transcription factors in *de novo* motif analyses (Figs. 4*C* and S3*D*). This is consistent with previous studies that have identified GATA family genes as regulators of DE differentiation. Importantly, peaks in clusters C6–C8 were related to HPC differentiation. *De novo* motif analyses of peaks in clusters C6–C8 revealed that HNF4A and NR5A2 were involved in HPC differentiation from the DE to HPC stage (Figs. 4*C* and S3, *D* and *E*). Moreover, clusters C9 and C10 were largely defined by marked peaks that opened during the HPC to HLC stage, suggesting that these genes may contribute to the formation of IMHs and HLCs. *De novo* motif analyses of clusters C9 and C10 demonstrated that the AP-1 family may modulate hepatic differentiation from the HPC to HLC stage (Figs. 4*C* and S3, *D* and *E*).

To corroborate the insights gained from transcriptome and ATAC-seq analyses, we investigated ChIP-seq data of JUN and JUND in liver cells (43). We then integrated these multi-omics analyses for a comprehensive understanding and found that JUN and JUND could bind to the transcription start sites of numerous HLC marker genes and liver P450 family genes (Fig. 4*D*), indicating that JUN and JUND serve as master regulators in the progression of liver maturation from HPCs to HLCs. We then examined ChIP-seq data of NR5A2 (44) and discovered NR5A2 could bind to the promoter region of AP-1 family genes. This observation concurs with our RNA-seq analysis, which demonstrated a notable upregulation in the expression of AP-1 family genes during hepatic differentiation (Fig. 4*E*). In line with NR5A2 ChIP-seq analysis, the knockdown of NR5A2 resulted in reduced expression of JUN, JUNB, and JUND in IMH cells (Fig. 4*F*). To further explore the role of NR5A2/AP-1 axis in hepatic differentiation from HPCs to HLCs, we knocked down NR5A2 in HPCs and induced their differentiation into HLCs. Notably, NR5A2-depleted HPCs underwent cell death during the differentiation process (Fig. S4*D*), highlighting the critical role of NR5A2 in hepatic differentiation. Ectopic expression of either FOS or JUN rescued cell death and promoted the further differentiation of HPCs into HLCs, leading to their maturation and ALB production (Fig. S4*D*). This indicates that NR5A2

PHHs, HLCs, and H1 hPSCs is performed using ELISA. The ALB ELISA reveals that the secretion levels of ALB in HLCs are remarkably comparable to those observed in PHHs. H1 hPSCs are used as a negative control. Results are expressed as mean ± SD. *E* and *F*, functional analysis of liver cell activities, including urea synthesis (E) and cytochrome P450 enzyme CYP3A4 activity (*F*), is performed in PHHs, HLCs, and hPSCs. H1 hPSCs are used as a negative control in these assays. Results are expressed as mean ± SD. *G*, xenograft experiments illustrate the further differentiation of transplanted HPCs, resulting in high expression levels of HLC markers *in vivo*. Renal capsule transplantation involved introducing HPCs into the kidneys of immune-compromised SCID mice, leading to their differentiation into HLCs. Immunofluorescence staining indicates a robust expression of HLC markers, including ALB, α-haptoglobin, and transferrin, in xenograft samples. This illustrates the substantial enrichment of these markers within transplanted cells, confirming the acquisition of hepatocyte characteristics *in vivo*. Scale bar: 100 μm. ALB, albumin; DECs, definitive endoderm cells; HPCs, hepatic progenitor cells; IMHs, immature hepatocytes; HLCs, hepatocyte-like cells; hPSCs, human pluripotent stem cells; PHH, primary human hepatocyte; AFP, alpha-fetoprotein; DE, definitive endoderm.

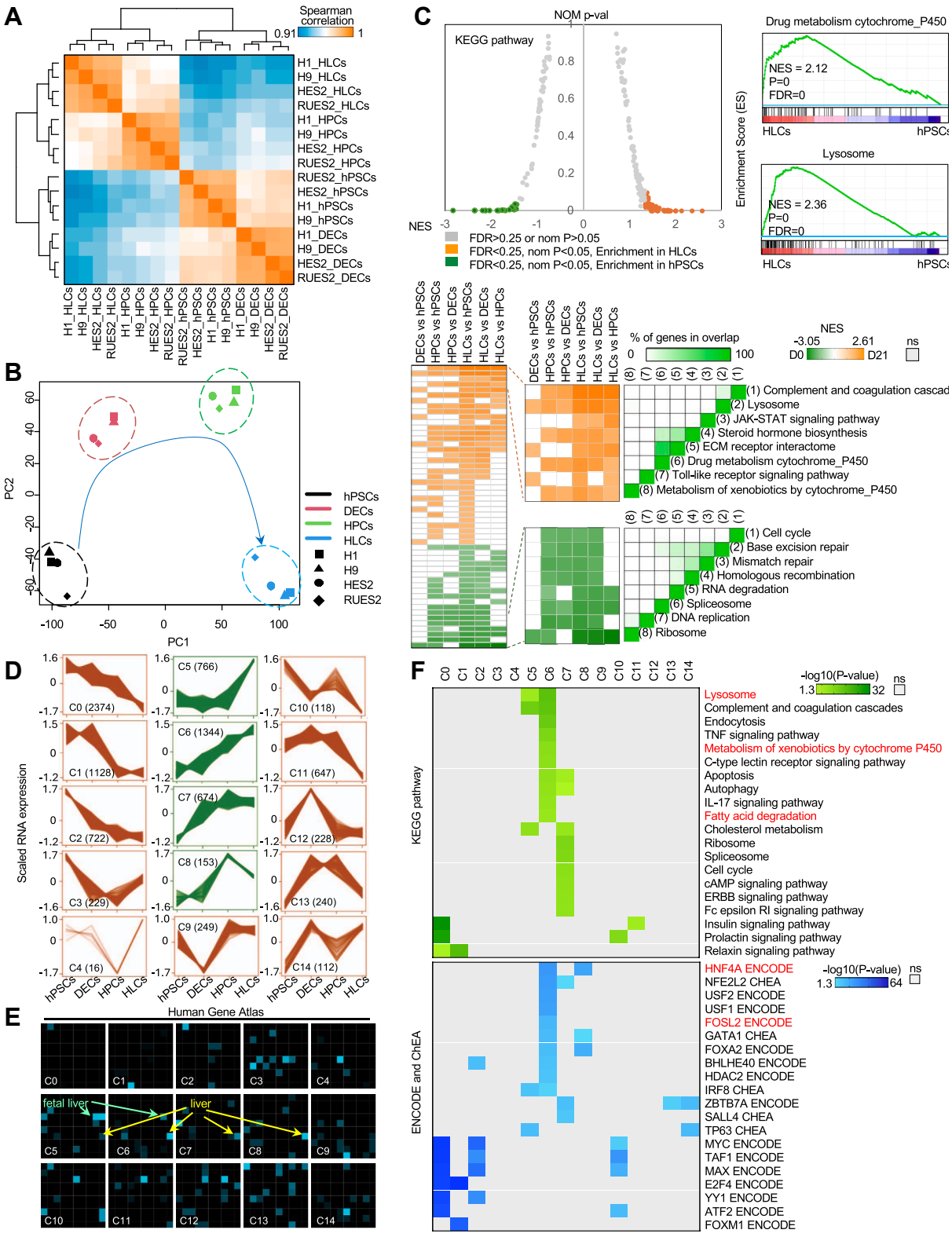


Figure 3. Identification of gene expression signatures during hepatic differentiation. *A*, Spearman's correlation of gene expression levels from RNA-seq analysis of four cell types during hepatic differentiation. The pairwise correlation coefficient is represented by a color gradient ranging from 0.91 (*blue*) to 1 (*orange*). *B*, PCA illustrating significant transcriptomic alterations during the differentiation process of HLCs. Each data point represents a sample at a specific stage of hepatic differentiation. The clear separation and distinct clustering of samples indicate substantial transcriptomic alterations occurring throughout the differentiation process. *C*, pathway analysis of gene sets enriched in HLCs compared to hPSCs. In the upper panel, the identification of KEGG pathways enriched in HLCs compared to hPSCs (p-value < 0.05, FDR q value < 0.25) reveals a high enrichment of pathways associated with hepatocyte functions, including drug metabolism cytochrome P450 and lysosome. The lower panel displays a heatmap illustrating the upregulated and downregulated pathways in HLCs. *D*, dynamic gene expression revealing 15 clusters during hepatic differentiation, representing genes with similar patterns across stages. The figure displays the temporal gene expression patterns observed during hepatic differentiation, classified into 15 distinct gene

facilitates the differentiation of HPCs into HLCs through AP-1 family genes. Collectively, our ATAC-seq data were consistent with RNA-seq data and highlighted the importance of HNF4A and NR5A2 in HPC differentiation and AP-1 family transcription factors in hepatic maturation.

Single-cell RNA sequencing analysis reveals NR5A2 and AP-1 are essential for hepatic differentiation

The utilization of single-cell RNA sequencing (scRNA-seq) has significantly enriched our understanding of gene expression dynamics during hepatic differentiation. In our study, we integrated scRNA-seq data of HLCs from the published studies (45) to analyze and compare dynamic gene expression in our hepatic differentiation samples. Remarkably, our analysis unveiled that HLCs can clearly separate from hPSCs, DECs, HPCs, and IMHs, underscoring the divergent gene expression profiles characterizing these five cell populations (Fig. 5*A*). Specifically, our findings highlighted the robust overexpression of NR5A2 at stages of IMH and HLCs, whereas high expressions of JUN, JUNB, JUND, and FOSL2 were observed at the stage of HLCs (Figs. 5*B* and S4*A*). To verify these results, we also performed immunofluorescent staining in human fetal liver tissues and PHHs, and the result unveiled a notable expression of NR5A2, JUN, FOS, FOSL2, and ALB in human fetal liver and PHHs (Fig. S4*C*). Together, these findings highlight the critical role of NR5A2 and its downstream AP-1 in hepatic differentiation and maturation.

Next, we performed the trajectory analysis of the differentiated cells, leveraging the advanced Monocle algorithm to gain deeper insights into the trajectory of hepatic maturation (Fig. 5, *C* and D). Remarkably, the trajectory of NR5A2 exhibited a distinctive profile, characterized by a remarkable upsurge in expression during both IMH and HLC stages. Conversely, the trajectory of FOS, FOSL2, JUN, and JUNB displayed a different pattern, with a predominant increase in expression at the HLC stage (Fig. 5*E*). These observations aligned seamlessly with our RNA-seq and ATAC-seq results, affirming the relevance of NR5A2 expression dynamics during IMH and HLC stages, and AP-1 expression dynamics during late-stage hepatic maturation. Furthermore, these trajectory-based insights harmonized remarkably with the gene expression patterns observed within the five distinct subclusters identified *via* UMAP analysis. This congruence underscores the robustness and reliability of our findings, as both methods converge to highlight the importance of NR5A2 and AP-1 family genes in hepatic development. Meanwhile, NR5A2, AP-1 family genes, such as CEBPA, FOSL2, HNF1A, HNF1B, HNF4G, and HLC marker genes, such as ALB and SERPINA3 were gradually upregulated along with trajectory differentiation process (Fig. 5*F*). Integrating ChIP-seq data for JUN, JUND, and FOSL2 (43, 45) with the top 100 gene expression profiles from scRNA-seq analysis, we revealed a clear upregulation of HLC marker genes governed by JUN, JUND, and FOSL2 during the HLC stage of hepatocyte differentiation (Fig. 5, *G* and *H*, and S4*B*). Collectively, our multi-omics analyses suggest that NR5A2 and AP-1 transcription factors contribute to the maturation of liver cells, transitioning from HPCs and IMHs to HLCs.

Safety assessment of differentiated HLCs

In response to the growing demand for a reliable source of hepatocytes for both *in vitro* studies and *in vivo* applications, we conducted a safety assessment of hPSCs (H1 and H9)-derived HLCs using our developed protocol. To evaluate the safety of these cells, we initiated a mutation analysis using whole genome sequencing (WGS) data obtained from cell samples collected at various stages of the protocol. The samples included Day 0 hPSCs cultured in mTeSR1 medium, Day 21 hPSCs continuously cultured in mTeSR1 medium, and Day 21 HLCs obtained after differentiation (Fig. S5*A*). The WGS data of the six samples were analyzed by the Genome Analysis Toolkit's best practice workflow (46, 47) (Fig. 6*A*).

We achieved a median depth of 48.5X (range from 45 to 60) for the six samples using the Illumina X10 platform, with high read quality (Q20 $>$ 96%) (Fig. S5*B*). Additional mutation quality criteria were implemented to eliminate false-positive calls, as depicted in Fig. 6*A*. The spectra of *de novo* mutations showed no significant differences between hPSCs culturing and hepatic differentiation processes (Fig. 6*B*). Additionally, both hPSCs culturing and hepatic differentiation process exhibited strong concordance in terms of mutation location (Fig. 6*C*). Furthermore, no significant differences in mutation number and variant allele fractions were observed between hPSCs culturing and hepatic differentiation (Fig. S5, *C* and *D*). *De novo* mutations were identified using M1–M4. We detected one exonic *de novo* mutation in M1, defined in the context of H1 hPSC day 21 culture, and three exonic *de novo* mutations in M4, defined in the context of H9 hPSCs (Fig. 6*D* and Fig. S5*E*). Among these four exonic *de novo* mutations, one nonsynonymous mutation was identified in M1, while the three mutations detected in M4 comprised two nonsynonymous mutations and one synonymous mutation. The Sorting Intolerant From Tolerant prediction (48, 49) revealed that none of the nonsynonymous mutations are predicted to be deleterious and/or significantly damaging (Fig. S5*E*).

clusters. "C0", "C1", *etc.*, refer to "Cluster 0", "Cluster 1", *etc.*, and that the number in parentheses indicates the number of genes assigned to each cluster. *E*, analysis of dynamic gene expression clusters (C0 to C14) during the differentiation of iPSCs into hepatic cells using Enrichr Human Gene Atlas demonstrates that clusters five through eight exhibit strong and specific enrichment for biological processes associated with liver function and fetal liver–enriched gene signatures. These enriched pathways reflect the progressive acquisition of hepatic identity and maturation throughout the differentiation process. *F*, KEGG pathway and transcription factor enrichment analyses indicating the enrichment of gene sets in Cluster 0 to 14. In the upper panel, hepatocyte-related pathways are significantly enriched in gene clusters five to seven during hepatic differentiation. In the lower panel, transcription factor enrichment analyses reveal that the targets of HNF4A, FOSL2, GATA family, and FOX family transcription factors are highly enriched in cluster 6, while the targets of MYC, YY1, and FOXM1 are highly enriched in clusters 0 to 3. FDR, false discovery rate; HLCs, hepatocyte-like cells; hPSCs, human pluripotent stem cell; iPSCs, human induced pluripotent stem cells; HNF4A, hepatocyte nuclear factor 4 alpha; PCA, principal component analysis; RNA-seq, RNA sequencing.

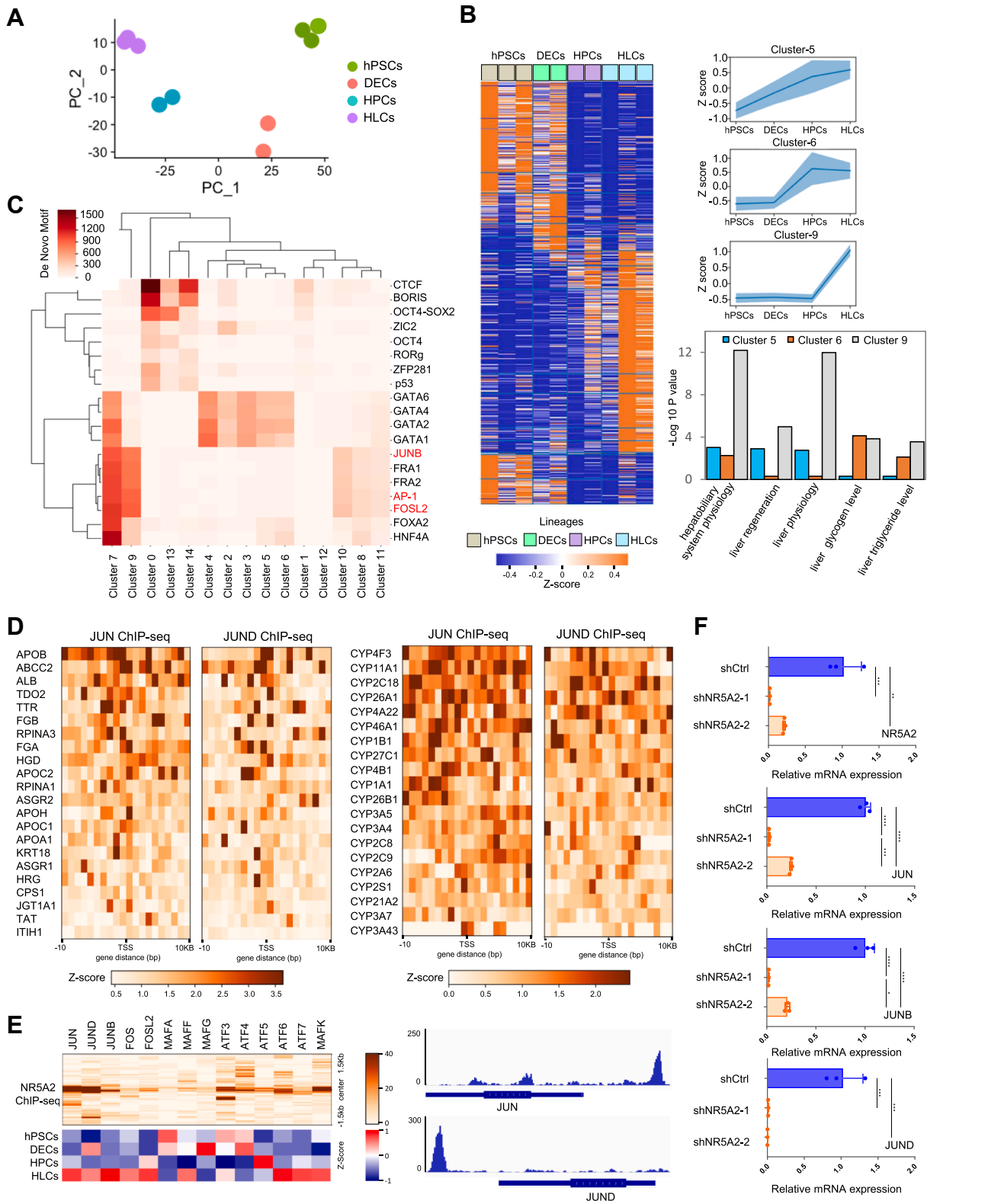


Figure 4. ATAC-seq analyses identify stage-specific transcription factors during hepatic differentiation. *A*, PCA showing significant chromatin alterations during hepatic differentiation. The plot illustrates distinct sample clusters based on chromatin accessibility, indicating substantial alterations of chromatin accessibility throughout differentiation. *B*, the *left panel*, heatmap representing dynamic changes in chromatin accessibility across different stages of hepatic differentiation. The y-axis displays ATAC-seq peaks, with each row corresponding to a distinct genomic region of open chromatin. The *right middle panel*, according to the similarities and differences in chromatin accessibility profiles, the significantly changed peaks can be grouped into 15 distinct clusters. The dynamic changes in chromatin accessibility of Clusters 5, 6, and 9 are observed and illustrated throughout the differentiation of HLCs. The *right bottom* panel, peak annotation reveals liver-related pathways are enriched in Clusters 5, 6, and 9. *C*, *de novo* motif analyses of peaks within each cluster. *De novo* motif analyses identify stage-specific transcription factors, with a notable enrichment of AP-1 family genes in Cluster 9. The colored genes represent key members of the AP-1 transcription factor family. *D*, heatmap of ChIP-seq peak intensities depicting JUN and JUND binding to the

Moreover, anchorage-independent growth (AIG) assay showed that hPSC-derived HLCs failed to form a colony in soft agar (Fig. 6*E*), indicating no *in vitro* tumorigenic ability. In comparison with liver cancer lines PLC/PRF/5 and Hep3B, hPSC-derived HLCs could not engraft a tumor and showed little tumorigenic potential *in vivo* xenograft assays (Fig. 6*F*). Together, these findings collectively indicate that our modified hepatic differentiation protocol for hPSC-derived HLCs is not mutagenic and has limited tumorigenicity. This underscores the safety of our developed HLCs for application in various fundamental, translational, and clinical contexts.

Discussion

In this study, we devised an AI-assisted tool reinforced by a machine learning-based algorithm to monitor morphological changes and refine a multi-step protocol, aiming to enhance the success rate of hepatic differentiation. Our algorithm achieved an efficiency rate of 90 to 95% of hepatic differentiation. It can predict the efficiency of stem cell differentiation at the HPC stage, helping researchers save time and resources in the early stages of hepatic differentiation. Through state-of-the-art high-throughput transcriptome, ATAC-seq, and scRNA-seq analysis, we unexpectedly discovered that NR5A2 and AP-1 family transcription factors play a crucial role in controlling the hepatic differentiation process.

Our user-friendly AI algorithm boasts high accuracy. During its development, we directly trained the algorithm using bright-field microscopy images of differentiated cells at the HPC stage. This approach yielded remarkable F1 scores and accuracy, as evidenced across three diverse datasets. The increasing prevalence of AI algorithms in the realms of stem cell differentiation and cell biology underscores their pivotal role and surging popularity in advancing biomedical research. For example, Kusumoto *et al.* developed an automated deep learning-based AI algorithm to identify endothelial cells derived from iPSCs (15). Similarly, Dursun *et al.* generated a CNN to recognize differentiated tenocytes by learning features directly from phase-contrast unlabeled cell images (50). Schaub *et al.* developed deep neural networks to predict the monolayer transepithelial resistance and tissue function of iPSC-derived retinal pigment epithelial cells using quantitative bright-field images (51). Orita *et al.* trained a CNN to perform quality control of human iPSC-derived cardiomyocytes (14). Guo *et al.* set up a machine learning workflow to improve analyzing the iPSCs-based embryo model and found that BMP4 best promoted the morphogenesis of the iPSCs and trophectoderm stem cells embryo (52). These findings underscore the significant role of AI algorithms in enhancing stem cell research. We recommend including longitudinal microscopy images in the trained set to establish a real-time AI algorithm that can estimate the differentiation efficiency as early as possible.

Ensuring the safety of HLCs produced by our established protocol is a critical consideration. Previous studies have shown that prolonged *in vitro* culturing of hPSCs can lead to genetic alterations, such as karyotypic aberrations, subchromosomal, and copy number variations, and nucleotide point mutations, which can result in culture adaptation. For instance, whole-exome sequencing or RNA sequencing revealed mutations in the TP53 gene within hES cell lines, with the allelic fraction of TP53 mutants escalating proportionally with passage number (53). A high prevalence of acquired BCOR mutations (26.9% of lines) in blood-derived hPSCs were identified using WGS and whole-exome sequencing (54). To assess the genetic stability of HLCs produced by our protocol, we performed WGS to detect any genetic mutations arising from the differentiation process. WGS identifies single nucleotide variations and insertions/deletions (indels) from the full chromosome down to the single nucleotide. Our findings indicate that both hPSC culturing and differentiation during our established hepatic differentiation process do not significantly induce genetic variants over a relatively short period. All of the detected variants were low-risk mutations. Additional examinations into the oncogenic traits of HLCs revealed that our differentiation method does not prompt tumorigenesis either *in vitro* or *in vivo*, reinforcing the assurance that the HLCs generated through this protocol are deemed safe for utilization.

Transcription factors play a pivotal role in lineage differentiation, yet their specific mechanisms in regulating hepatic differentiation, especially during the maturation stage, remain largely unexplored. Previous studies indicated that NR5A2 expression progressively increases during foregut differentiation into the liver (55), and it persists at high levels throughout fetal liver growth, regulating the early expression of AFP (56). This suggests NR5A2's involvement in liver development. Moreover, AP-1 has been implicated in developmental and organogenic processes, as fetuses with mutations in JUN exhibit liver developmental abnormalities (57, 58). However, the mechanism through which NR5A2 regulates the AP-1 family during hepatic differentiation remains unexplored. Through a comprehensive analysis involving RNA-seq, ATAC-seq, and scRNA-seq, we have portrayed the transcriptional dynamics throughout the hepatic differentiation process. Our findings reveal a significant increase in NR5A2 expression during HLC maturation, while AP-1 displays elevated expression primarily during the HLC stage.

transcription start sites of mature HLC marker genes (*left panel*) and liver-associated P450 family genes (*right panel*). ChIP-seq data for JUN and JUND were retrieved from the ENCODE database (GSM935364 and GSM935649). *E*, heatmap of ChIP-seq peak intensities illustrating NR5A2's binding in the promoter region of AP-1 family genes. The *upper panel* displays heatmaps of NR5A2 ChIP-seq peak intensities in the promoter region of AP-1 family genes. The *lower panel* shows a heatmap of AP-1 family gene expression in differentiated cells at various stages by RNA-seq. The *right panel* of IGV snapshots illustrates NR5A2's binding peak occupancy over the promoter region of JUN and JUND. ChIP-seq data for NR5A2 were obtained from GSM846875. *F*, qRT-PCR demonstrating that depletion of NR5A2 leads to downregulation of AP-1 genes in IMH cells. Results are expressed as mean ± SD. $*p< 0.05$, $**p< 0.01$, $***p < 0.001$, $****p < 0.0001$ (n = 3 technical replicates). HLCs, hepatocyte-like cells; ATAC-seq, assay for transposase-accessible chromatin sequencing; IMHs, immature hepatocytes; AP-1, activator protein-1; NR5A2, nuclear receptor subfamily 5 group A member 2; PCA, principal component analysis; RNA-seq, RNA sequencing.

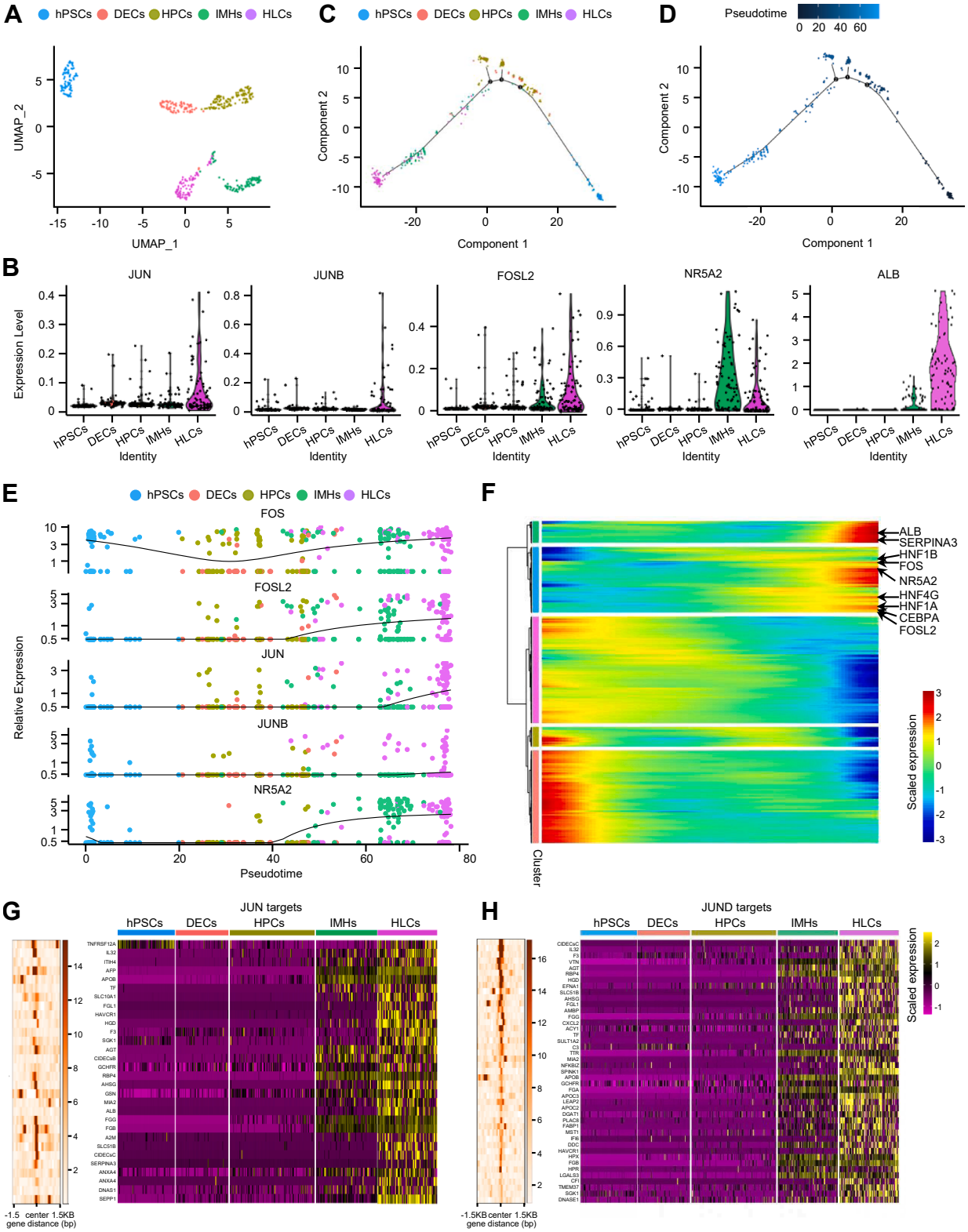


Figure 5. scRNA-seq reveals upregulation of NR5A2 and AP-1 family genes during hepatic differentiation. *A*, UMAP displays distinct subclusters of cells at various differentiation stages, providing insights into dynamic cellular transitions (total cell count: 425; Cluster hPSCs: 80; Cluster DECs: 70; Cluster HPCs: 113; Cluster IMHs: 81; Cluster HLCs: 81). *B*, Violin plots showing the expression levels of JUN, JUNB, FOSL2, and NR5A2 during hepatic differentiation. AP-1 family genes exhibit increasing expression patterns as cells progress through various differentiation stages, with NR5A2 reaching its peak expression in the IMH stage. ALB expression serves as a positive marker of HLCs. *C* and *D*, monocle trajectory plot illustrating the dynamics of subclusters and their pseudotime curve, offering a comprehensive visualization of cellular transitions during hepatic differentiation. The pseudotime trajectory is consistent with the underlying biological transition. *E*, Monocle plot displaying the relative expression change of FOS, FOSL2, JUN, JUNB, and NR5A2 across pseudotime, showing dynamic evolution along the trajectory. AP-1 family genes are primarily upregulated in HLC stages, while NR5A2 begins its upregulation in the IMH stage. *F*, heatmap hierarchical clustering shows the dynamic expression changes of the top 5000 genes along the pseudotime

NR5A2 demonstrates the ability to bind to the promoter region of AP-1 family genes, potentially facilitating their expression activation. Additionally, AP-1 family genes can bind to the promoter regions of HLC marker genes. Depletion of NR5A2 resulted in decreased expression of JUN, JUNB, and JUND, suggesting that NR5A2 may regulate hepatic differentiation and HLC maturation through the AP-1 family. Further investigations are necessary to elucidate the NR5A2-AP-1 regulatory network for a comprehensive understanding of hepatogenesis.

Taken together, we employed a machine learning-based AI algorithm to refine hepatic differentiation, enhancing efficiency and ensuring the safe production of HLCs from hPSCs. Utilizing multi-omics technology, we unraveled the transcriptional regulatory network, shedding light on previously overlooked transcription factors NR5A2 and AP-1 in governing hepatic differentiation and maturation. These findings provide essential insights into the potential applications of hPSC-derived HLCs in regenerative medicine, disease investigation, and drug pharmacology.

Experimental procedures

Antibodies, cytokines, growth factors, chemicals, and shRNAs

Antibodies against SOX17 (R&D Systems, AF1924), HNF4A (Cell Signaling Technology, 3113S), AFP (R&D Systems, MAB1368), ALB (Bethyl Laboratories, A80-129A), α-Haptoglobin (Santa Cruz Biotechnology, sc-365396), Transferrin (Proteintech, 17435-1-AP), NR5A2 (Sigma Aldrich, ABE2867M), JUN (Cell Signaling Technology, 9165), FOS (Cell Signaling Technology, 2250), and FOSL2 (Sigma Aldrich, HPA004817) were purchased from the indicated suppliers. Activin A (R&D Systems; 338-AC-010), B-27(-insulin) supplement (Life Technologies; A1895601), B-27(+insulin) supplement (Life Technologies; 17504044), BMP4 (R&D Systems, 314-BP-050), FGF2 (PeproTech, 100-18B-1 mg), HGF (R&D Systems, 294-HG-025), OSM (R&D Systems, 295-OM-010), and CHIR99021 (Sigma-Aldrich; SML1046-5MG) were also procured from the specified suppliers. The lentiviral plasmids for FOS (Addgene #59140) and JUN (Addgene #142292) expression were obtained from Addgene. The lentiviral shRNAs were designed using the TRC library database and inserted into pLKO.pig (59, 60). The shRNA primers used in this study are listed in the supplemental information, Table S1.

Differentiation of hPSCs to HLCs

hPSC lines H1, H9, HES2, and RUES2 were utilized for hepatic differentiation to generate HLCs, following established protocols for cell maintenance (61, 62). Monitoring of hepatocyte differentiation employed an AI algorithm based on machine learning, tracking morphological changes through bright-field microscopy images. Upon achieving peak pluripotency, hPSCs were dissociated into single cells using Accutase (Stemcell Technology, 07,920), and 1×10^7 cells were seeded onto Matrigel-coated 6-well plates using human embryonic stem cell medium [Dulbecco's modified Eagle's medium (DMEM)/F12 (Corning Cellgro, 10–090-CV) supplemented with 20% (vol/vol) KnockOut Serum Replacement (Invitrogen, 10828028), L-glutamine, non-essential amino acids, β-mercaptoethanol, penicillin-streptomycin antibiotics, and 10 ng/ml bFGF (R&D Systems, 233-FB)].

Next, induction of DECs involved 24 h in endoderm differentiation medium A [RPMI 1640 medium (Life Technologies, 11875093) supplemented with B-27(-insulin) supplement (Life Technologies, A1895601), 100 ng/ml activin A (R&D Systems, 338-AC-010), and 3 μM CHIR99021 (Sigma-Aldrich, SML1046-5MG)] and 48 h in endoderm differentiation medium B ([RPMI 1640 medium supplemented with B-27(-insulin) supplement, 100 ng/ml Activin A], resulting in petal-like DECs visible under the microscope (Fig. 2, *A* and B). Then, the DE cells were transitioned to the HPC medium (RPMI1640 medium supplemented with B-27(+insulin) supplement, 20 ng/ml BMP4 (R&D Systems, 314-BP-050), and 10 ng/ml FGF2 (PeproTech, 100–18B-1 mg) for 5 days. The HPC medium was replaced every other day for 5 days, and polygonal cells were visible under the microscope (Figs. 2, *A* and *B*, and S1*F*).

Subsequent generation of IMHs utilized IMH differentiation medium (RPMI1640 supplemented with B-27(+insulin) supplement and 20 ng/ml HGF (R&D Systems, 294-HG-025)), producing polygonal cells with small cytoplasmic bubbles (Figs. 2, *A* and *B*, and S1*F*). To further differentiate the immature HLCs, we transferred them to HLC medium for 5 days. HLC medium consisted of HCM-based medium (Lonza; CC-3199) supplemented with 20 ng/ml OSM (R&D Systems, 295-OM-050) but not EGF. The medium was changed every other day for 5 days, and polygonal cells with numerous bubbles in the cytoplasm were visible under the microscope (Figs. 2, *A* and *B*, and S1*F*).

Human primary hepatocyte culture

Three different individual clones of PHHs, OSZ, ZMC, and JYN, were purchased from Novabiosis. The PHHs were thawed in Liver Hepatocyte Plating Medium (Novabiosis, 7011), then cultured in collagen type I-coated plates, and

trajectory inferred by Monocle. Each column represents a single cell ordered by pseudotime from early (*left*) to late (*right*) states. Each row corresponds to a gene, and genes are grouped by hierarchical clustering based on similarities in their temporal expression patterns. The resulting gene modules highlight coordinated transcriptional programs activated at distinct stages of the trajectory. NR5A2 and AP-1 family genes are upregulated during hepatic differentiation. *G*, increased expression of JUN-targeted genes during hepatic differentiation. On the left: Heatmaps of ChIP-seq peak intensities in JUN-targeted gene regions. On the *right*: Heatmap of JUN-targeted gene expression in differentiated cells at various stages by scRNA-seq. ChIP-seq data for JUN were extracted from the ENCODE database (GSM935364). *H*, increased expression of JUND-targeted genes during hepatic differentiation. On the *left*: Heatmaps of ChIP-seq peak intensities in JUND-targeted gene regions. On the right: Heatmap of JUND-targeted gene expression in differentiated cells at various stages by scRNA-seq. ChIP-seq data for JUND were extracted from the ENCODE database (GSM935649). IMHs, immature hepatocytes; scRNA-seq, single-cell RNA sequencing; NR5A2, nuclear receptor subfamily 5 group A member 2; HLCs, hepatocyte-like cells; AP-1, activator protein-1; ALB, albumin; HPCs, hepatic progenitor cells.

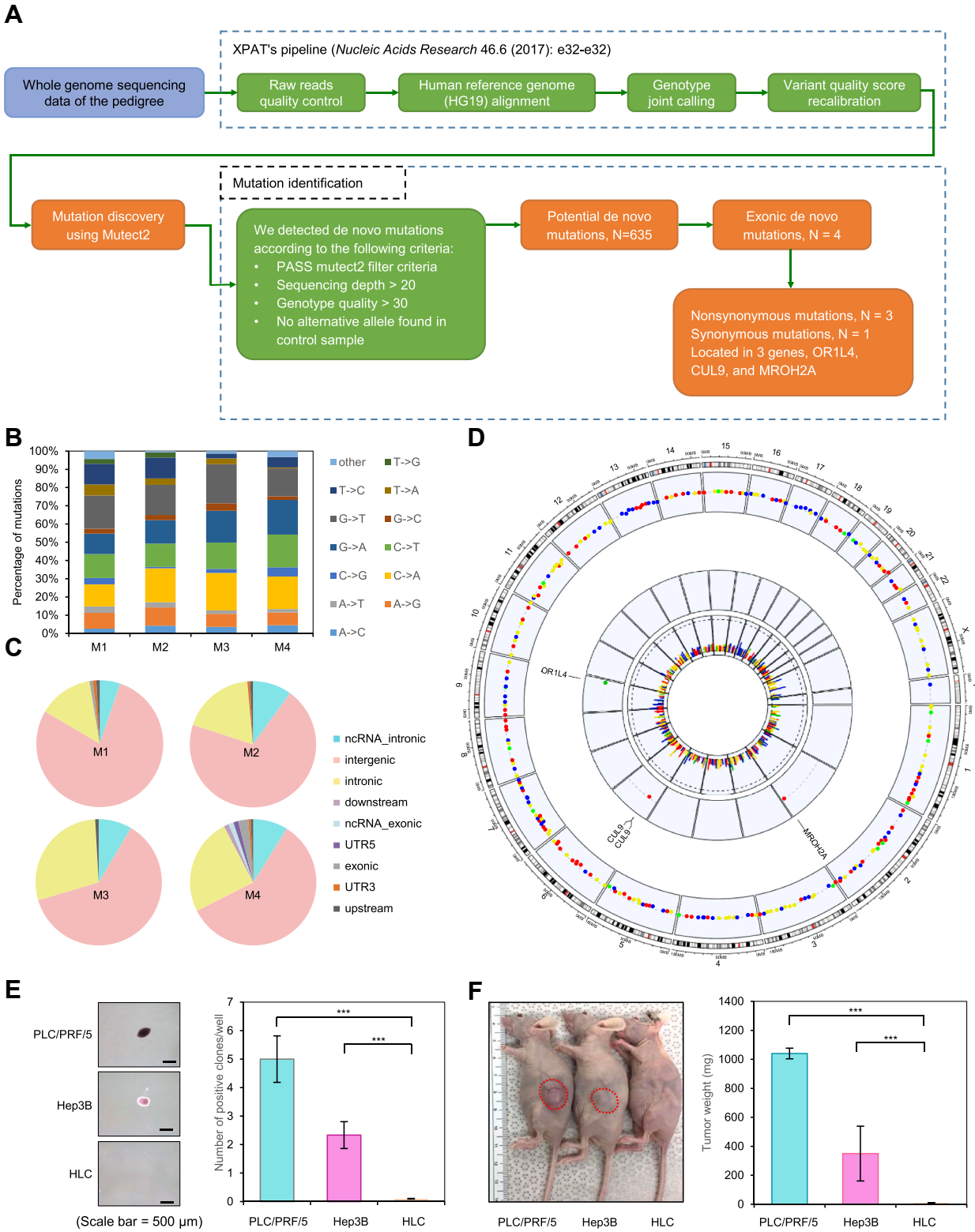


Figure 6. WGS identifies mutations during hepatic differentiation. *A*, diagram of the XPAT pipeline for analyzing WGS data. The flow chart of mutation analysis illustrates the criteria for identifying *de novo* mutations in the XPAT analysis. *B*, bar graph presenting the spectra of mutation types observed in M1 to M4. *C*, Pie chart depicting the distribution of mutations in different locations, such as exonic, intronic, and UTRs. *D*, circle plot depicting *de novo* mutations during either *in vitro* culture or *in vitro* differentiation processes. M1 (*green plots*) and M3 (*yellow plots*) represent *in vitro* culture, and M2 (*blue plots*) and M4 (*red plots*) represent *in vitro* differentiation. From the outermost to the innermost circles, each depicts all *de novo* mutations, genes with coding mutations, exonic mutations (*red*: nonsynonymous, *green*: synonymous), and the variant allele fractions. *E*, AIG assay demonstrating the *in vitro* tumorigenic ability of hPSC-derived HLCs, PLC/PRF/5, and Hep3B. PLC/PRF/5 and Hep3B demonstrate the ability to form a colony, while hPSC-derived HLCs fail to form a colony when cultured in a soft agar medium. Results are expressed as mean ± SD. ***$p < 0.001$ (n = 3 biological replicates). *F*, xenograft assay illustrating the *in vivo* tumorigenic ability of hPSC-derived HLCs, PLC/PRF/5, and Hep3B. PLC/PRF/5 and Hep3B demonstrate the ability to engraft a tumor, while hPSC-derived HLCs cannot engraft a tumor in the *in vivo* xenograft assays. Results are expressed as mean ± SD. ***$p < 0.001$ (n = 4 biological replicates). WGS, whole genome sequencing; HLCs, hepatocyte-like cells; hPSCs, human pluripotent stem cells.

maintained in Hepatocyte Culture Medium (Novabiosis, 7111), following the manufacturer's instructions.

Immunofluorescence staining

The cells were fixed on tissue culture plates using 4% paraformaldehyde in 1 × phosphate buffered saline (PBS) for 10 min at room temperature (RT) and washed with 1× PBS. To enable permeabilization and blocking, cells were treated with 10% donkey serum containing 0.1% Triton X-100 for 1 h at RT. Subsequently, cells were incubated with the primary antibody in the blocking solution overnight at 4 °C (Table S3). After washing with PBST (PBS containing 0.1% Triton X-100) three times, secondary antibodies were incubated for 1 h at RT. DAPI was used to stain cell nuclei for 3 min at RT. Finally, cells were observed under a Leica DMi8 microscope.

RNA isolation and qRT-PCR

The cells were washed with 1 × PBS and lysed in TRIzol (Invitrogen, 15596026) following the manufacturer's protocol. To synthesize cDNA, 1 μg of mRNA was reverse transcribed using the iScript cDNA synthesis kit (Bio-Rad Laboratories, 1,708,891). qRT-PCR was carried out using a CFX96 machine (Bio-Rad Laboratories). A 20 μl PCR reaction solution was prepared by combining 1 μl cDNA, 1 μl each of 10 μM forward and reverse qPCR primers designed using PrimerBank (https://pga.mgh.harvard.edu/primerbank/), 10 μl SYBR Green PCR Master Mix (Bio-Rad Laboratories, 1725124), and 7 μl RT-PCR grade water. All reactions were performed in triplicate and normalized to *GAPDH* mRNA expression. The primer sequences used in this experiment are listed in the Table S2. The relative fold changes in expression were determined using the $2^{-\triangle\triangle CT}$ method.

RNA-seq

The quality of RNA samples was assessed using the 2100 Bioanalyzer (Agilent Technologies), and 1000 ng of total RNA was used to generate cDNA libraries with the Next Ultra II RNA Library Prep Kit (New England BioLabs, E7770) following the manufacturer's instructions. The resulting libraries were sequenced using an Illumina HiSeq X10 sequencer with PE150 high-throughput sequencing after multiplexing. RNA-seq data analyses were conducted as previously described (63).

GSEA and Enrichr

GSEA was conducted using the GSEA software, ranking genes based on the "Signal2Noise" metric and employing a "weighted" enrichment statistic. Significance criteria included a false discovery rate q-value less than 0.25 and a nominal *p*-value less than 0.05. Enrichment analysis of KEGG pathways and the Human Gene Atlas was carried out using Enrichr (64). Default gene set databases, encompassing GO biological processes, KEGG, and TFT motifs, were utilized in the GSEA.

ATAC-seq

After washing with 1 ml of cold 1 × PBS, 50,000 viable cells were pelleted and then resuspended in a 1× PBS containing 0.1% NP40, 0.1% Tween-20, and 0.01% digitonin. The cells were incubated on ice for 3 min, followed by the addition of 1 ml of cold 1 × PBS containing 0.1% Tween-20 and 0.01% digitonin. The nuclei were pelleted and resuspended in 50 μl of transposition mixture and incubated at 37 °C for 30 min. The reaction was stopped, and the products were purified using the Qiagen PCR DNA purification kit (Qiagen, 28106). The eluted products were directly amplified using NEBNext Q5 Hot Start HiFi PCR Master Mix (New England BioLabs, M0543 L) and Nextera primers for library preparation. The resulting libraries were submitted for PE150 high-throughput sequencing. ATAC-Seq datasets were aligned to the human reference genome (hg19) using snap-aligner (https://github.com/amplab/snap). Reads mapped to the mitochondrial chromosome were discarded. Uniquely mapped reads were further converted to bigWig files for visualization with the HOMER toolset. Peak calling for ATAC-Seq was performed with MACS2 (parameters: -f BAMPE -g hs -B -q 0.01).

scRNA-seq analysis

The scRNA-seq data for HLCs were obtained from published studies GSE81252 and GSE96981 (65). Processing of the scRNA-seq data was performed on the NovelBrain Cloud Analysis Platform by NovelBio Bio-Pharm Technology Co., Ltd. Initial data preprocessing involved default parameter application for adapter sequence filtering and removal of low-quality reads, resulting in clean data. The cell barcode whitelist was identified using UMI-tools for Single Cell Transcriptome Analysis. Clean data, based on UMIs, were aligned to the human genome (Ensemble version 91) using STAR mapping with customized parameters from the UMI-tools standard pipeline, enabling UMI quantification for each sample.

Cells passing the quality filter exhibited over 200 expressed genes and a mitochondrial UMI rate below 40%, with mitochondrial genes excluded from the expression table. Normalization and regression were performed on the expression table using the Seurat package (version: 3.1.4, https://satijalab.org/seurat/). This normalization was based on UMI counts and the percentage of mitochondrial content, resulting in scaled data. PCA on the scaled data used the top 2000 highly variable genes, and the top 10 principal components were employed to construct UMAP plots. Unsupervised cell clustering utilized a graph-based method (resolution = 0.8) with the top 10 principal components from PCA. Marker genes were identified using the FindAllMarkers function with the Wilcoxon rank sum test algorithm, applying criteria: 1. Log-fold change > 0.25; 2. p-value < 0.05; 3. Minimum percent of cells expressing the marker gene (min.pct) > 0.1.

Pseudo-time analysis

Single-Cell Trajectories analysis was performed using Monocle (http://cole-trapnell-lab.github.io/monocle-release)

with DDR-Tree and default parameters. Prior to the Monocle analysis, we curated marker genes from the Seurat clustering results and collected raw expression counts from cells meeting the filtering criteria. Our analysis primarily focused on pseudo-time analysis, emphasizing branch expression analysis modeling (BEAM Analysis). This approach enabled a detailed exploration of gene expression patterns crucial for determining branch fates.

AIG assay

The AIG assay was performed according to the previously published (66). Briefly, 10,000 cells were mixed with either HCM medium (for HLCs) or complete DMEM medium (for PLC/PRF/5 or Hep 3B) and 0.5% SeaPlaque low-melting agarose (Lonzo, 50,100) and plated over a base layer of complete DMEM medium (containing 10% FBS and 1× Penicillin-Streptomycin antibiotics) with 1% agar. After 1 month of culture, colonies were fixed and stained with 0.005% crystal violet for 1 h, and positive colonies (>50 μm) were counted using a Leica DMi8 microscope.

Flow cytometry analysis

HLCs, PHHs, H1 hPSCs, and PLC/PRF/5 cells were fixed by 2% paraformaldehyde (PFA) and permeabilized by 1% Triton X-100 followed by 1 h incubation of human ALB PE-conjugated antibody (R&D, IC1455P). The ALB-positive cells were analyzed using an LSR Fortessa analyzer (BD Biosciences), and the data were processed with FlowJo software.

ALB ELISA assay

Human ALB levels were measured using the Human Albumin ELISA Kit (Elabscience, E-UNEL-H0006), following the manufacturer's instructions. The colorimetric readings were taken at 450 nm using a TECAN Infinite M Plex multimode microplate reader.

Urea assay

Urea concentration was quantitatively determined using the QuantiChrom Urea Assay Kit (BioAssay Systems, DIUR-100), following the manufacturer's instructions. Colorimetric readings were taken at 520 nm using a TECAN Infinite M Plex multimode microplate reader.

CYP3A4 activity assay

CYP3A4 activity was measured using the P450-Glo CYP3A4 Assay Kit (Promega, V9001), following the manufacturer's instructions. Luminescence was measured using a TECAN Infinite M Plex multimode microplate reader.

Indocyanine green staining

HLCs and PHHs were stained with 1 mg/ml indocyanine green (MP Biomedicals, 02155020-CF) for 20 min at 37 °C, followed by three washes with PBS. Cell images were captured using a Leica DMi8 microscope.

Stem cell research ethical compliance

The stem cell research adheres to all applicable ethical regulations, as approved by the Institutional Human Embryonic Stem Cell Research Oversight Committee at The University of Texas Health Science Center at Houston (SCRO-24-5).

Dataset of FOVs

A total of 1163 fields of view (FOVs) were acquired from images representing various stages of differentiation, including 12 FOVs of hPSCs (Day 0), 413 FOVs of DE (Day 3), and 738 FOVs of HPCs, (Day 9). To ensure balanced representation across differentiation stages and minimize sampling bias, the dataset was randomly divided into training (70%), validation (20%), and testing (10%) subsets. Specifically, the hPSC dataset was divided into eight FOVs for training, two for validation, and two for testing. The DE dataset comprised 289 training FOVs, 82 validation FOVs, and 42 testing FOVs. The HPC dataset included 516 training FOVs, 148 validation FOVs, and 74 testing FOVs. The model performance was comprehensively evaluated using accuracy, F1 score, precision, and recall as quantitative metrics. All images were acquired at a resolution of 1296 × 966 pixels and stored in 16-level grayscale format.

Development of the AI algorithm

We developed a machine learning-based AI algorithm to facilitate the monitoring of hepatocyte differentiation by analyzing morphological changes in bright-field microscopy images of stem cells. To create datasets for training the AI algorithm, a senior researcher annotated and classified the morphologies of hPSCs (Class 2), DECs (Class 3), HPCs (Class 1), and non-HPCs (Class 4) in a total of 1163 field-of-view images obtained from three differentiation stages. The researcher also determined the differentiation result (success/failed HPC) for each image, serving as the gold standard for comparison with the predictions of the AI algorithm.

To streamline the annotation process and reduce the researcher's workload, we employed "weak supervised learning" in the AI algorithm training. In this approach, the researcher only needed to annotate areas with dominant HPC distribution rather than every instance of HPC. The workflow of the AI algorithm development is illustrated in Figure 1*E*. Initially, the AI algorithm was trained using a dataset consisting of 341 images with successful HPC differentiation and 167 images with failed differentiation. Throughout the training, the performance of a smaller independent "validation" dataset, comprising 86 successful and 51 failed HPC images, was monitored to determine when to conclude the training process. Finally, the "test" dataset, containing 64 successful and 29 failed HPC images, was utilized to assess the performance of the AI algorithm.

Interpretation of prediction results from the AI algorithm

Utilizing the differentiation result criteria established by the senior researcher (considered as the ground truth), the AI algorithm categorizes images based on the percentage of HPC tiles. If an image contains >80% HPC tiles, the AI algorithm

predicts an "Excellent" result. For images with 50% to 80% HPC tiles, the prediction is "Good," and for those with <50% HPC tiles, the prediction is "Failed".

Figure 1*F* presents a representative AI algorithm prediction report. The upper panel showcases images predicted as "success," while the lower panel features images predicted as "failed." The left column displays the original FOVs, and the right column illustrates the AI algorithm's inference results at the tile level. Images are segmented into five classes of cell types, represented by different colored lines (12 × 15 tiles) in the image. The percentage of tiles for each class in one image is calculated and displayed in the right table. In the upper panel, the AI algorithm predicted an "Excellent" result for an image with >80% HPC tiles, while in the lower panel, it predicted a "Failed" result for an image with <50% HPC tiles.

Evaluation of the AI algorithm

To assess the efficacy of the developed AI algorithm, we gauged its performance across three datasets, employing two widely used metrics: accuracy and F1 score. In comparison to the differentiation results determined by a senior researcher (considered as the ground truth), binary accuracy was quantified using the following formula: (TP = True positive; FP = False positive; TN = True negative; FN = False negative).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

We included the F1 score, the harmonic mean of precision and recall, as it provides a balanced measure of performance particularly useful for evaluating AI algorithms in imbalanced medical datasets (67).

$$\text{F1} = 2 \cdot \frac{precision \cdot recall}{precision+recall} = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$$

Renal capsule transplantation

The renal capsule transplantation was approved by the University of Southern California Institutional Animal Care and Use Committee under protocol number 21037 (Y.W.C). Mouse kidney capsule transplantation was conducted using NOD.Cg-Prkdcscid.Il2rgtm1Wjl/SzJ (NSG) mice (The Jackson Laboratory) housed in a specific pathogen-free mouse facility. Mice aged 10 to 13 weeks and of either gender were utilized for the experiment, with seven mice per trial. All experiments and animal care were conducted in compliance with protocols approved by the Icahn School of Medicine at Mount Sinai Institutional Animal Care and Use Committee. Before surgery, 1×10^7day 10 to 15 hPSC-derived HPCs were combined with 5 μl Matrigel and implanted under the kidney capsule. The outgrowths were excised and either fixed in 4% paraformaldehyde for paraffin embedding or embedded freshly in optimal cutting temperature (OCT) for immunostaining with ALB, α-haptoglobin, and transferrin.

Tumor xenograft assay

The tumor xenograft assay adheres to all relevant ethical regulations and was approved by The University of Texas Health Science Center at Houston Institutional Animal Welfare Committee under protocol number AWC-20 to 0005 (D.F.L. and R.Z). Tumor xenograft experiments were conducted following previously described methods (63). Subcutaneous inoculation of 1×10^7 HLCs, PLC/PRF/5, or Hep3B was performed in 8-week-old female nude (*Foxn1*^{nu}) mice (The Jackson Laboratory). Tumor size was measured in two dimensions, and tumor volume was calculated using the formula: half × Length × Width^2, employing calipers. At 4 weeks after injection, all mice were euthanized, and tumor samples were excised and examined.

Human fetal liver tissue collection and processing

All human fetal liver tissues were collected under Institutional Review Board Approval (STUDY-22–00,065) at the Icahn School of Medicine at Mount Sinai. Consent for tissue donation was obtained after the patient had already made the decision for pregnancy termination and was obtained by a different clinical staff member than the physician performing the procedure. All tissues were de-identified, and the only clinical information collected was gestational age and the presence of any maternal or fetal diagnoses. Liver samples ranging in age from 9.0 to 23 weeks of gestation were used for this study.

The dissected fetal liver tissues underwent a 7-day sample processing period. On day 1, the tissues were fixed in 4% PFA and kept at 4 °C overnight. On day 2, the tissues were washed with 1x PBS to remove excess PFA and stored at 4 °C overnight. Samples were then equilibrated with 10% sucrose on day 3, 20% sucrose on day 4, 30% sucrose on day 5, a 1:1 ratio of 30% sucrose, and 100% OCT compound on day 6 and finally, a 100% OCT mixture on the last day of sample processing. After thoroughly equilibrated with 100% OCT, fetal liver tissue specimens were carefully embedded in OCT (Tissue-Tek OCT Compound, Sakura Finetek) in a cryomold. Embedded fetal liver tissue specimens were transferred to a precooled −80 °C freezer for storage. Tissues were sectioned at a thickness of 5 μm using a Leica CM3050 S Cryostat. Immunostaining of NR5A2, JUN, FOS, FOSL2, and ALB was performed using the immunofluorescence staining procedure previously described.

WGS and mutation analysis

On Day 0 and Day 21, hPSCs H1 and H9, as well as Day 21 HLCs derived from H1 and H9, were collected and washed with 1 ml PBS. Genomic DNA was extracted from the cell pellets using the DNeasy Blood and Tissue kit (Qiagen, 69,504). Whole-genome libraries with insertion lengths of 500 bp were constructed, flow cells were prepared, and sequencing clusters were generated following the manufacturer's protocols. The libraries were subsequently sequenced using the Illumina HiSeq X-10 platform with 2*150 bp paired-end reads.

The mutation analysis was conducted as previously described (62). Raw data quality control (QC), human

reference genome (hg19) alignment, genotype joint calling, and variant quality score recalibration were performed following the Genome Analysis Toolkit's Best Practices workflow (47). Single nucleotide variations with VQSR tranche scores higher than 99.95 and INDELs with scores higher than 98.0 were excluded during variant-level QC. *De novo* mutations were identified using Mutect2, with three additional criteria: 1) a minimum of 20 reads covering the mutation site across all samples; 2) a genotype quality exceeding 30, and 3) the absence of an alternative allele in the control sample.

Statistical analysis and reproducibility

Statistical analysis of experimental data was conducted using GraphPad Prism 7.0 software and Microsoft Excel. The sample size was not determined using statistical methods. Continuous data were presented as mean ± SD. For reproducibility, experiments yielding qualitative results were repeated at least in triplicate.

Data availability

The data supporting the findings of this study can be found in the paper and its Supplementary Information. Additionally, the RNA-seq, ATAC-seq, and WGS data have been deposited in the Gene Expression Omnibus repository under GSE246557 and are available for public access. The datasets generated and analyzed for this study are available from the corresponding author upon reasonable request.

Supporting information—This article contains supporting information.

Acknowledgments—We would like to express our gratitude to members of the R. Z. and D.-F. L. laboratory for their valuable technical assistance and insightful discussions. We appreciate Biorender (https://biorender.com/) for assisting in drawing the graphic abstract. The human fetal liver tissue experiment was conducted with the support of the Developmental Origins of Health and Disease Biorepository at Icahn School of Medicine at Mount Sinai. We are grateful to the Biorepository laboratory team and Biorepository participants for their contributions to this research.

Author contributions—W.-L. Y., J. T., T. T. T. P., M.-F. H., M. E. F., Y-W. H., R. S., C.-W. C., A. X., S.-Y. C., and R. W. validation; W.-L.Y., T-.J. L., S.-Y. C., and Y. Y. software; W.-L. Y., T-.J. L., T. N. T., M. K., Y. Z., Y. Y., and Y-W. C. resources; W.-L. Y., D.-.F. L., R. Z., and T-.J. L. conceptualization; Z. H., J. T., W.-L. Y., Y.-W. C., C. D. H., S.-Y. C., T.-J. L., H. X., D.-F. L., C.-W. C., and R. Z methodology; Z. H., J. T., W.-L. Y., M.-F. H., R. W., C.-W. C., A. X., Y. Y., T. N. T., M. K., M. E. F., Y.-W. H., T. T. T. P., R. S., Y.-W. C., D. Z., D.-F. L., and R. Z. investigation; Z.H., J. T., W.-L. Y., R. W., M.-F. H., C.-W. C., A. X., Y. Y., C. D. H., D.-F. L., M. K., T. N. T., Y-W. C., and R. Z. formal analysis; J. T., D. Z., D.-.F. L., Z. H., C.D.H., A. X., and R. W. data curation; D. Z., D.-.F. L., R. Z., T-.J. L., H. X., Y. Z., and C. D. H. funding acquisition; D.-.F. L. and R. Z. writing–review & editing; Z. H., J. T., W.-L.Y., R. W., M.-F. H., Y. Y., D.-F.L., and R. Z. writing–original draft; D.-.F. L., R. Z., and H. X. supervision; S.-Y. C. and Y-W. C. visualization; Y. Z. project administration.

Funding and additional information—D.-F. L. was supported by the Rolanette and Berdon Lawrence Bone Disease Program of Texas and a Pablove Foundation Childhood Cancer Research Grant (690785). J. T. and Z. H. were supported by the Ke Lin Program Fellowship. Z. H. was supported by the National Natural Science Foundation of China (No. 82002510) and Science and Technology Planning Project of Guangzhou, China (No. 202201010937). J. T. was supported by the National Natural Science Foundation of China (No. 82472685), Science and Technology Planning Project of Guangdong Province, China (No. 2023A1515010154), and Science and Technology Planning Project of Guangzhou, China (No. 202201010904), and the Youth S&T Talent Support Programme of Guangdong Provincial Association for Science and Technology.. M.-F. H. was supported by the Rosalie B. Hite Fellowship and the Dr John J. Kopchick Fellowship. A. X. and T. T. T. P. were CPRIT Postdoctoral Fellows in the Biomedical Informatics, Genomics and Translational Cancer Research Training Program (CPRIT Grant RP210045). Y. W. H. and T. T. T. P. were Postdoctoral Fellows of the American Cancer Society (PF-25-1433660-01-PFCBI and PF-25-1434257-01-PFMBB). M. E. F. was supported by a predoctoral fellowship of the Gulf Coast Consortia, on the Training Interdisciplinary Pharmacology Scientists Program (Grant No. T32GM139801) and the Dr John J. Kopchick Fellowship. D. Z. was supported by the Department of Defense Horizon Award (W81XWH-20-1-0389). R. S. was the BIG-TCR Predoctoral Fellow supported by the Biomedical Informatics, Genomics and Translational Cancer Research Training Program (BIG-TCR, CPRIT grant RP210045).

Conflict of interest—Wei-Lei Yang, Shih-Yu Chen, and Tien-Jen Liu are employees of AIxMed, Inc.

Abbreviations—The abbreviations used are: AFP, alpha-fetoprotein; ALB, albumin; AP-1, activator protein-1; ATAC-seq, assay for transposase-accessible chromatin sequencing; AI, artificial intelligence; DE, definitive endoderm; DECs, DE cells; FOVs, field-of-view images; GATK, Genome Analysis Toolkit; GSEA, Gene Set Enrichment Analysis; HLCs, hepatocyte-like cells; HNF4A, hepatocyte nuclear factor 4 alpha; HPCs, hepatic progenitor cells; hPSCs, human pluripotent stem cells; iPSCs, human induced pluripotent stem cells; IMHs, immature hepatocytes; NR5A2, nuclear receptor subfamily 5 group A member 2; OSM, oncostatin M; PCA, principal component analysis; PHHs, primary human hepatocytes; RNA-seq, RNA sequencing; scRNA-seq, single-cell RNA sequencing; WGS, whole-genome sequencing.

References

1. Cardinale, V., Lanthier, N., Baptista, P. M., Carpino, G., Carnevale, G., Orlando, G., *et al.* (2023) Cell transplantation-based regenerative medicine in liver diseases. *Stem Cell Rep.* **18**, 1555–1572
2. Berasain, C., Arechederra, M., Argemi, J., Fernandez-Barrena, M. G., and Avila, M. A. (2023) Loss of liver function in chronic liver disease: an identity crisis. *J. Hepatol.* **78**, 401–414
3. Michalopoulos, G. K., and Bhushan, B. (2021) Liver regeneration: biological and pathological mechanisms and implications. *Nat. Rev. Gastroenterol. Hepatol.* **18**, 40–55
4. Si-Tayeb, K., Noto, F. K., Nagaoka, M., Li, J., Battle, M. A., Duris, C., *et al.* (2010) Highly efficient generation of human hepatocyte-like cells from induced pluripotent stem cells. *Hepatology* **51**, 297–305

5. Schwartz, R. E., Fleming, H. E., Khetani, S. R., and Bhatia, S. N. (2014) Pluripotent stem cell-derived hepatocyte-like cells. *Biotechnol. Adv.* **32**, 504–513
6. Qin, J., Chang, M., Wang, S., Liu, Z., Zhu, W., Wang, Y., *et al.* (2016) Connexin 32-mediated cell-cell communication is essential for hepatic differentiation from human embryonic stem cells. *Sci. Rep.* **6**, 37388
7. Peters, D. T., Henderson, C. A., Warren, C. R., Friesen, M., Xia, F., Becker, C. E., *et al.* (2016) Asialoglycoprotein receptor 1 is a specific cell-surface marker for isolating hepatocytes derived from human pluripotent stem cells. *Development* **143**, 1475–1481
8. Mallanna, S. K., and Duncan, S. A. (2013) Differentiation of hepatocytes from pluripotent stem cells. *Curr. Protoc. Stem Cell Biol.* **26**, 1G.4.1–1G.4.13
9. Hannan, N. R., Segeritz, C. P., Touboul, T., and Vallier, L. (2013) Production of hepatocyte-like cells from human pluripotent stem cells. *Nat. Protoc.* **8**, 430–437
10. Carpentier, A., Nimgaonkar, I., Chu, V., Xia, Y., Hu, Z., and Liang, T. J. (2016) Hepatic differentiation of human pluripotent stem cells in miniaturized format suitable for high-throughput screen. *Stem Cell Res* **16**, 640–650
11. Warren, C. R., O'Sullivan, J. F., Friesen, M., Becker, C. E., Zhang, X., Liu, P., *et al.* (2017) Induced pluripotent stem cell differentiation enables functional validation of GWAS variants in Metabolic disease. *Cell Stem Cell* **20**, 547–557.e547
12. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. (2018) Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510
13. Waisman, A., La Greca, A., Mobbs, A. M., Scarafia, M. A., Santin Velazque, N. L., Neiman, G., *et al.* (2019) Deep learning neural networks highly predict very early onset of pluripotent stem cell differentiation. *Stem Cell Rep.* **12**, 845–859
14. Orita, K., Sawada, K., Koyama, R., and Ikegaya, Y. (2019) Deep learning-based quality control of cultured human-induced pluripotent stem cell-derived cardiomyocytes. *J. Pharmacol. Sci.* **140**, 313–316
15. Kusumoto, D., Lachmann, M., Kunihiro, T., Yuasa, S., Kishino, Y., Kimura, M., *et al.* (2018) Automated deep learning-based system to identify endothelial cells derived from induced pluripotent stem cells. *Stem Cell Rep.* **10**, 1687–1695
16. Marzec-Schmidt, K., Ghosheh, N., Stahlschmidt, S. R., Kuppers-Munther, B., Synnergren, J., and Ulfenborg, B. (2023) Artificial intelligence supports automated characterization of differentiated human pluripotent stem cells. *Stem Cells* **41**, 850–861
17. Lee, S., Tak, E., Choi, J., Kang, S., Lee, K., Namgoong, J. M., *et al.* (2025) Evaluation of hepatic progenitor and hepatocyte-like cell differentiation using machine learning analysis-assisted surface-enhanced Raman spectroscopy. *Biomater. Res.* **29**, 0190
18. Mirzaei, H., Khodadad, N., Karami, C., Pirmoradi, R., and Khanizadeh, S. (2020) The AP-1 pathway; A key regulator of cellular transformation modulated by oncogenic viruses. *Rev. Med. Virol.* **30**, e2088
19. Eferl, R., and Wagner, E. F. (2003) AP-1: a double-edged sword in tumorigenesis. *Nat. Rev. Cancer* **3**, 859–868
20. Lee, S. Y., Yoon, J., Lee, M. H., Jung, S. K., Kim, D. J., Bode, A. M., *et al.* (2012) The role of heterodimeric AP-1 protein comprised of JunD and c-Fos proteins in hematopoiesis. *J. Biol. Chem.* **287**, 31342–31348
21. Shaulian, E., and Karin, M. (2002) AP-1 as a regulator of cell life and death. *Nat. Cell Biol.* **4**, E131–E136
22. Jochum, W., Passegue, E., and Wagner, E. F. (2001) AP-1 in mouse development and tumorigenesis. *Oncogene* **20**, 2401–2412
23. Lee, Y. K., and Moore, D. D. (2008) Liver receptor homolog-1, an emerging metabolic modulator. *Front. Biosci.* **13**, 5950–5958
24. Lai, F., Li, L., Hu, X., Liu, B., Zhu, Z., Liu, L., *et al.* (2023) NR5A2 connects zygotic genome activation to the first lineage segregation in totipotent embryos. *Cell Res.* **33**, 952–966
25. Fayard, E., Auwerx, J., and Schoonjans, K. (2004) LRH-1: an orphan nuclear receptor involved in development, metabolism and steroidogenesis. *Trends Cell Biol.* **14**, 250–260
26. Sun, Y., Demagny, H., and Schoonjans, K. (2021) Emerging functions of the nuclear receptor LRH-1 in liver physiology and pathology. *Biochim. Biophys. Acta Mol. Basis Dis.* **1867**, 166145
27. Lu, T. T., Makishima, M., Repa, J. J., Schoonjans, K., Kerr, T. A., Auwerx, J., *et al.* (2000) Molecular basis for feedback regulation of bile acid synthesis by nuclear receptors. *Mol. Cell* **6**, 507–515
28. Smith, J. C., Price, B. M., Van Nimmen, K., and Huylebroeck, D. (1990) Identification of a potent Xenopus mesoderm-inducing factor as a homologue of activin A. *Nature* **345**, 729–731
29. Hay, D. C., Fletcher, J., Payne, C., Terrace, J. D., Gallagher, R. C., Snoeys, J., *et al.* (2008) Highly efficient differentiation of hESCs to functional hepatic endoderm requires ActivinA and Wnt3a signaling. *Proc. Natl. Acad. Sci. U S A.* **105**, 12301–12306
30. McLean, A. B., D'Amour, K. A., Jones, K. L., Krishnamoorthy, M., Kulik, M. J., Reynolds, D. M., *et al.* (2007) Activin a efficiently specifies definitive endoderm from human embryonic stem cells only when phosphatidylinositol 3-kinase signaling is suppressed. *Stem Cells* **25**, 29–38
31. Teo, A. K., Ali, Y., Wong, K. Y., Chipperfield, H., Sadasivam, A., Poobalan, Y., *et al.* (2012) Activin and BMP4 synergistically promote formation of definitive endoderm in human embryonic stem cells. *Stem Cells* **30**, 631–642
32. Morrison, G. M., Oikonomopoulou, I., Migueles, R. P., Soneji, S., Livigni, A., Enver, T., *et al.* (2008) Anterior definitive endoderm from ESCs reveals a role for FGF signaling. *Cell Stem Cell* **3**, 402–415
33. Schmidt, C., Bladt, F., Goedecke, S., Brinkmann, V., Zschiesche, W., Sharpe, M., *et al.* (1995) Scatter factor/hepatocyte growth factor is essential for liver development. *Nature* **373**, 699–702
34. Kamiya, A., Kinoshita, T., Ito, Y., Matsui, T., Morikawa, Y., Senba, E., *et al.* (1999) Fetal liver development requires a paracrine action of oncostatin M through the gp130 signal transducer. *EMBO J.* **18**, 2127–2136
35. Nell, P., Kattler, K., Feuerborn, D., Hellwig, B., Rieck, A., Salhab, A., *et al.* (2022) Identification of an FXR-modulated liver-intestine hybrid state in iPSC-derived hepatocyte-like cells. *J. Hepatol.* **77**, 1386–1398
36. Liu, X., Wang, M., Jiang, T., He, J., Fu, X., and Xu, Y. (2019) IDO1 maintains pluripotency of primed human embryonic stem cells by promoting glycolysis. *Stem Cells* **37**, 1158–1165
37. Tan, D. S., Holzner, M., Weng, M., Srivastava, Y., and Jauch, R. (2020) SOX17 in cellular reprogramming and cancer. *Semin. Cancer Biol.* **67**, 65–73
38. Zaret, K. S., and Grompe, M. (2008) Generation and regeneration of cells of the liver and pancreas. *Science* **322**, 1490–1494
39. Tachmatzidi, E. C., Galanopoulou, O., and Talianidis, I. (2021) Transcription control of liver development. *Cells* **10**, 2026
40. Peng, W. C., Kraaier, L. J., and Kluiver, T. A. (2021) Hepatocyte organoids and cell transplantation: what the future holds. *Exp. Mol. Med.* **53**, 1512–1528
41. Vargas, R., and Castaneda, M. (1981) Role of elongation factor 1 in the translational control of rodent brain protein synthesis. *J. Neurochem.* **37**, 687–694
42. Strick-Marchand, H., and Weiss, M. C. (2002) Inducible differentiation and morphogenesis of bipotential liver cell lines from wild-type mouse embryos. *Hepatology* **36**, 794–804
43. Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., *et al.* (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402–405
44. Holmstrom, S. R., Deering, T., Swift, G. H., Poelwijk, F. J., Mangelsdorf, D. J., Kliewer, S. A., *et al.* (2011) LRH-1 and PTF1-L coregulate an exocrine pancreas-specific transcriptional network for digestive function. *Genes Dev.* **25**, 1674–1679
45. Gertz, J., Savic, D., Varley, K. E., Partridge, E. C., Safi, A., Jain, P., *et al.* (2013) Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol. Cell* **52**, 25–36
46. Van der Auwera, G. A., and O'Connor, B. D. (2020) *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*, 1st Edition, O'Reilly Media, Inc., Sebastopol, CA

47. Yu, Y., Hu, H., Bohlender, R. J., Hu, F., Chen, J. S., Holt, C., *et al.* (2018) XPAT: a toolkit to conduct cross-platform association studies with heterogeneous sequencing datasets. *Nucleic Acids Res.* **46**, e32
48. Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457
49. Ng, P. C., and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814
50. Dursun, G., Tandale, S. B., Gulakala, R., Eschweiler, J., Tohidnezhad, M., Markert, B., *et al.* (2021) Development of convolutional neural networks for recognition of tenogenic differentiation based on cellular morphology. *Comput. Methods Programs Biomed.* **208**, 106279
51. Schaub, N. J., Hotaling, N. A., Manescu, P., Padi, S., Wan, Q., Sharma, R., *et al.* (2020) Deep learning predicts function of live retinal pigment epithelium from quantitative microscopy. *J. Clin. Invest.* **130**, 1010–1023
52. Guo, J., Wang, P., Sozen, B., Qiu, H., Zhu, Y., Zhang, X., *et al.* (2021) Machine learning-assisted high-content analysis of pluripotent stem cell-derived embryos in vitro. *Stem Cell Rep.* **16**, 1331–1346
53. Merkle, F. T., Ghosh, S., Kamitaki, N., Mitchell, J., Avior, Y., Mello, C., *et al.* (2017) Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations. *Nature* **545**, 229–233
54. Rouhani, F. J., Zou, X., Danecek, P., Badja, C., Amarante, T. D., Koh, G., *et al.* (2022) Substantial somatic genomic variation and selection for BCOR mutations in human induced pluripotent stem cells. *Nat. Genet.* **54**, 1406–1416
55. Rausa, F. M., Galarneau, L., Belanger, L., and Costa, R. H. (1999) The nuclear receptor fetoprotein transcription factor is coexpressed with its target gene HNF-3beta in the developing murine liver, intestine and pancreas. *Mech. Dev.* **89**, 185–188
56. Galarneau, L., Pare, J. F., Allard, D., Hamel, D., Levesque, L., Tugwood, J. D., *et al.* (1996) The alpha1-fetoprotein locus is activated by a nuclear receptor of the Drosophila FTZ-F1 family. *Mol. Cell Biol.* **16**, 3853–3865
57. Hilberg, F., Aguzzi, A., Howells, N., and Wagner, E. F. (1993) c-jun is essential for normal mouse development and hepatogenesis. *Nature* **365**, 179–181
58. Eferl, R., Sibilia, M., Hilberg, F., Fuchsbichler, A., Kufferath, I., Guertl, B., *et al.* (1999) Functions of c-Jun in liver and heart development. *J. Cell Biol.* **145**, 1049–1061
59. Lee, D. F., Su, J., Ang, Y. S., Carvajal-Vergara, X., Mulero-Navarro, S., Pereira, C. F., *et al.* (2012) Regulation of embryonic and induced pluripotency by aurora kinase-p53 signaling. *Cell Stem Cell* **11**, 179–194
60. Lee, D. F., Su, J., Sevilla, A., Gingold, J., Schaniel, C., and Lemischka, I. R. (2012) Combining competition assays with genetic complementation strategies to dissect mouse embryonic stem cell self-renewal and pluripotency. *Nat. Protoc.* **7**, 729–748
61. Lee, D. F., Su, J., Kim, H. S., Chang, B., Papatsenko, D., Zhao, R., *et al.* (2015) Modeling familial cancer with induced pluripotent stem cells. *Cell* **161**, 240–254
62. Tu, J., Huo, Z., Yu, Y., Zhu, D., Xu, A., Huang, M. F., *et al.* (2022) Hereditary retinoblastoma iPSC model reveals aberrant spliceosome function driving bone malignancies. *Proc. Natl. Acad. Sci. U S A.* **119**, e2117857119
63. Xu, A., Liu, M., Huang, M. F., Zhang, Y., Hu, R., Gingold, J. A., *et al.* (2023) Rewired m(6)A epitranscriptomic networks link mutant p53 to neoplastic transformation. *Nat. Commun.* **14**, 1694
64. Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97
65. Camp, J. G., Sekine, K., Gerber, T., Loeffler-Wirth, H., Binder, H., Gac, M., *et al.* (2017) Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546**, 533–538
66. Jewell, B. E., Xu, A., Zhu, D., Huang, M. F., Lu, L., Liu, M., *et al.* (2021) Patient-derived iPSCs link elevated mitochondrial respiratory complex I function to osteosarcoma in Rothmund-Thomson syndrome. *PLoS Genet.* **17**, e1009971
67. Hicks, S. A., Strumke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., *et al.* (2022) On evaluation metrics for medical applications of artificial intelligence. *Sci. Rep.* **12**, 5979