Machine learning-based identification of key genes underlying sex differences in hepatocellular carcinoma and targeted drug screening

ZHEN WANG[1*], JINGHUA NING[1,2*], HONGYUAN ZHANG[1], QINGFEN RUAN[3], YANHONG ZHAO[1], RUN QU[1], CONGCONG LV[1], YUTONG WU[1], WEIDONG LIU[1], XIAOYIN YANG[1], ZIMING LI[1], YI LIANG[4] and YUZHE ZHANG[1,5,6]

[1]Department of Biochemistry and Molecular Biology, College of Basic Medical Sciences, Dali University, Dali, Yunnan 671003, P.R. China; [2]Department of Pathology, Fumin County People's Hospital, Kunming, Yunnan 650400, P.R. China; [3]Department of Gastroenterology, The First Affiliated Hospital of Dali University, Dali, Yunnan 671003, P.R. China; [4]Princess Margaret Cancer Centre, University Health Network, TMDT-MaRS Centre, Toronto, Ontario ON M5G 1L7, Canada; [5]Yunnan Provincial Key Laboratory of Entomological Biopharmaceutical Research and Development, Dali University, Dali, Yunnan 671000, P.R. China; [6]Yunnan Key Laboratory of Screening and Research on Anti-pathogenic Plant Resources from West Yunnan (Cultivation), Dali, Yunnan 671000, P.R. China

Received December 8, 2025; Accepted March 23, 2026

DOI: 10.3892/br.2026.2147

Abstract. Hepatocellular carcinoma (HCC) shows a marked predominance in men, yet the molecular basis for this sex disparity remains unclear. The present study leveraged multi-omics data and machine learning algorithms to identify key genes associated with sex-specific differences in HCC and to screen for putative candidate compounds, aiming to provide new insights for sex-specific therapy. The mRNA expression data of male and female patients with HCC and paracancerous tissues were obtained from the GEO and TCGA databases. To mitigate overfitting, data were partitioned into independent training and testing sets. Candidate genes were screened by differential expression analysis and weighted gene co-expression network analysis. A total of four complementary algorithms, random forest, support vector machines, generalized linear models and extreme gradient boosting were used to identify key genes with high predictive capability. CYP17A1 and IRX3 were identified as the top differentially expressed core genes associated with HCC in men. Pan-cancer analysis showed that CYP17A1 was lowly expressed in the majority of tumors, but significantly highly expressed in HCC, rectal adenocarcinoma and gastric cancer (P<0.001). Functional cell-based assays showed that knockout of CYP17A1 inhibited the proliferation, migration and invasion ability of HCC cells (P<0.001). Immunohistochemistry showed that CYP17A1 protein expression was significantly increased in HCC tissues from male patients when compared with that in paracancerous tissues (P<0.001), whereas there was no significant difference in female patient tissues (P>0.05). Notably, while IRX3 was identified computationally, its functional role remains to be experimentally validated. Molecular docking predicted a potential interaction between the natural compound Saikosaponin A and the CYP17A1 protein, and cellular assays revealed that it dose-dependently inhibits HCC cell malignant phenotypes. The present study suggests that CYP17A1 is associated with sex differences in HCC, potentially via the androgen signaling axis. Furthermore, IRX3 emerges as a novel hypothesis-generating candidate gene. Finally, the findings of the present study highlight Saikosaponin A as a putative therapeutic candidate for male patients with HCC, warranting further target-dependency investigations.

Introduction

Hepatocellular carcinoma (HCC) is one of the leading causes of cancer-related mortalities worldwide. It occupies a dominant position among primary types of liver cancer (1). Due to the insidious onset of HCC in its early stages and the lack of typical symptoms and signs, the disease is delayed, and diagnosis is made at an advanced stage (2). Patients with advanced HCC are frequently unsuitable for surgical treatment due to tumor metastasis, compromised liver function and poor systemic status, leading to a generally poor prognosis (3,4).

Epidemiologic data show that there is a considerable sex difference in the incidence of HCC, with a global prevalence of 2.8-3 times higher in men when compared with women (3,4). In the present study, all analyses regarding these disparities are

Correspondence to: Dr Yuzhe Zhang, Department of Biochemistry and Molecular Biology, College of Basic Medical Sciences, Dali University, 22 Wanhua Road, Dali, Yunnan 671003, P.R. China
E-mail: lzuzyz1568@hotmail.com

*Contributed equally

Key words: hepatocellular carcinoma, sex differences, machine learning, CYP17A1, Saikosaponin A

strictly based on biological sex (men vs. women) rather than sociocultural factors. Studies show that sex hormones carry out an important role in the development of HCC. Estrogens inhibit HCC by regulating the cytokine IL-6, whereas androgens promote HCC by inhibiting lipocalin secretion or directly activating pro-cancer signaling pathways (5,6). Testosterone in particular, leads to reduced levels of lipocalin by inhibiting its secretion, a mechanism that is considered to be one of the most important reasons for the increased incidence of HCC in men (7). However, the effect of these hormones on the sex differences in mechanisms of HCC has not been fully clarified at the molecular level.

Machine learning, as an emerging bioinformatics analysis method, has shown great potential for application in the medical field and has been widely used in the analysis of clinical datasets (8). By constructing robust risk models to aid clinical decision-making and simultaneously redefining patient categories based on data characteristics, precision medicine can be realized in the future (9). Machine learning has been used to analyze high-dimensional transcriptomic data with promising results in identifying key genes for diseases, providing key clues for the study of disease mechanisms (10).

The present study combined differential expression analysis, weighted gene co-expression network analysis (WGCNA) and four machine learning algorithms, including random forest (RF), support vector machine (SVM), generalized linear model (GLM) and extreme gradient boosting (XGBoost), to systematically screen the core genes associated with sex differences in the development of HCC. This was then combined with pan-cancer analysis, functional cell-based assays and virtual drug screening to explore the potential molecular associations of the key genes and identify putative candidate compounds. Specifically, the screening in the present study highlighted Saikosaponin A as a promising therapeutic candidate. This will provide a theoretical basis for understanding the sex disparity in HCC incidence and proposing candidate compounds for further evaluation.

Materials and methods

Cell lines and reagents. Human liver cancer cell lines HepG2 and Huh7, and 293T were purchased from Beijing Beina Chunglian Institute of Biotechnology and The Cell Bank of Type Culture Collection of The Chinese Academy of Sciences. The authenticity of the cell lines used in the present study was verified via short tandem repeat profiling. The plasmids used in the experiments [including the CYP17A1 knockout plasmid (cat. no. L25930), pLenti-Control-sgRNA (cat. no. L00011) and lentiviral packaging vector set A (cat. no. L00002S)] were obtained from Beyotime Biotechnology. According to the manufacturer, the sgRNA incorporated in the commercial CYP17A1 knockout plasmid (cat. no. L25930) was generated using Beyotime's CRISPR/Cas9 sgRNA rapid screening and validation system, and its efficacy had been pre-validated by a T7 endonuclease I (T7EI) assay. The primary reagents included DMEM containing 10% fetal bovine serum (Gibco; Thermo Fisher Scientific, Inc.), 0.25% trypsin-EDTA digest (Beijing Solarbio Science & Technology Co., Ltd.), penicillin-streptomycin solution (100X; Beyotime Biotechnology), BCA Protein Quantification Kit (cat. no. 500T; Beyotime Biotechnology) and puromycin (cat. no. ST551; Beyotime Biotechnology). Cell culture conditions were maintained at 37˚C in a humidified incubator with 5% CO_2.

Data acquisition and pre-processing. The mRNA expression profiling data of HCC and corresponding paracancerous tissues were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/; accession nos. GSE19665, GSE54236, GSE84402 and GSE121248), which included samples of both sexes and were generated in the original studies by Deng *et al* (11), Villa *et al* (12), Wang *et al* (13) and Wang *et al* (14). Batch effects were corrected using Surrogate Variable Analysis [sva package, v3.54.0 (15)] in R (v4.4.2; R Foundation for Statistical Computing), and the expression matrix of HCC vs. normal tissue for sex groupings was extracted. Differentially expressed genes (DEGs) were screened using limma (v3.62.2) based on |log2 fold change (FC)|≥0.5 and an adjusted $P<0.05$ after Benjamini-Hochberg correction.

WGCNA. A WGCN was constructed based on the expression matrix of DEGs, and a dynamic tree-cutting algorithm was used to divide the gene modules and screen for those significantly associated with the HCC phenotype ($P<0.05$ for module significance). The 'WGCNA' R package (v1.72) was utilized to construct the network. The network parameters were determined by setting a soft threshold of $\beta=6$ (scale-free topological fit index $R^2>0.85$). Modules significantly associated with HCC phenotypes were screened (Pearson correlation; $P<0.05$) and the top 10% of genes with high intra-module connectivity were extracted as candidate hub genes.

Machine learning-based core gene screening. The WGCNA hub genes were intersected with sex-specific DEGs (yielding a total of 36 candidate genes). Given the relatively limited sample size and the small, highly pre-filtered feature set (only 36 candidate genes), further dividing the dataset into separate training and testing cohorts would severely compromise the training process and the statistical power. Instead, to maximize data utility while strictly preventing overfitting, the present study relied entirely on a rigorous 10-fold cross-validation strategy across the entire dataset. A total of four machine learning algorithms, RF, SVM, GLM and XGBoost, were used to assess gene importance. During the 10-fold cross-validation process, the data were iteratively partitioned into training and validation folds, ensuring that the predictive performance of each model was robustly assessed on unseen data within each fold. The predictive performance was evaluated by calculating the area under the receiver operating characteristic curve (AUC), with all models achieving an AUC >0.95. The overlapping genes (CYP17A1 and IRX3) ranked in the top ten for importance across all four algorithms were ultimately selected as the core genes. Their sex-specific differential expression profiles were further validated using the GEO database.

Bioinformatics validation. For pan-cancer expression analysis, the TIMER2.0 platform (http://timer.cistrome.org/) was used to analyze the expression differences of the core genes across 30 tumors in The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/; Wilcoxon rank-sum test; with

P<0.05 as the criterion for determining statistically significant differences). For survival and clinical correlation analyses, TCGA clinical data were integrated and the association between core gene expression and patients' overall survival was evaluated using the Kaplan-Meier method (R package 'survival'; v3.3.1). Correlations with tumor grade and stage were analyzed using the Spearman rank correlation test. Gene set enrichment analysis was performed to identify the biological pathways associated with the core genes. For immune infiltration analysis, the CIBERSORT algorithm (R package 'CIBERSORT'; v1.06) was applied to quantitatively analyze the infiltration levels of 22 immune cells and explore the correlation between core gene expression and the immune microenvironment (P<0.05).

Collection of samples. Tumors and corresponding paracancerous tissues of liver cancer patients (n=18; 12 men and 6 women; age range, 40-75 years) were included in the present study. The samples were collected between January 2019 and June 2023. Specifically, samples dating from January 2019 to November 2022 were retrospectively accessed from historical archival formalin-fixed paraffin-embedded blocks, while samples from November 2022 to June 2023 were prospectively collected. Based on the study design, the precise inclusion criteria were: i) Patients who underwent radical resection for liver cancer; ii) pathologically confirmed primary liver cancer; iii) availability of matched paraffin-embedded tumor and paracancerous tissues (with paracancerous tissue strictly defined as being >10 mm away from the tumor margin). The exclusion criteria were: i) Patients who received preoperative radiotherapy, chemotherapy or targeted therapy; ii) presence of other primary malignancies. Following surgical resection, fresh tissues were snap-frozen and stored at -80˚C within 30 min, while matched tissue blocks were routinely formalin-fixed and paraffin-embedded for subsequent analysis. The present study was conducted with the approval of the Ethics Committee of the First Affiliated Hospital of Dali University (Approval no. OFY20221122001), which explicitly covered both the retrospective use of archived material and the prospective collection of new samples. Written informed consent was obtained from all patients prior to their participation; notably, for the newer prospectively collected samples, dedicated written informed consent for research use was strictly obtained prior to surgery. Strict confidentiality of patient identities was maintained. Detailed clinicopathological characteristics, including age, biological sex, tumor stage and viral hepatitis status, were retrospectively collected from medical records.

Immunohistochemistry (IHC). Paraffin sections were deparaffinized with xylene and rehydrated with descending gradient ethanol. Antigen retrieval was performed in sodium citrate buffer (pH 6.0) at 100˚C for 20 min. After blocking with goat serum blocking solution (cat. no. C0265; Beyotime Biotechnology) at room temperature for 10 min, the sections were incubated in a humidified chamber with rabbit anti-CYP17A1 primary antibody (1:200; cat. no. bsm-54306R; Bioss) overnight at 4˚C, followed by an HRP-labeled goat anti-rabbit secondary antibody (1:1,000; cat. no. G1213-100UL; Wuhan Servicebio Technology) at 37˚C for 30 min. Visualization was achieved using DAB, followed by hematoxylin counterstaining at room temperature for 15 sec. Staining intensity (0-3) and the percentage score of positive cells (1-4) were independently evaluated by two pathologists in a double-blind manner; the final IHC score was calculated as intensity x percentage score, and discordant cases were reviewed jointly to reach consensus.

Knockout cell line construction and functional experiments. To construct stable CYP17A1 knockout cells, 293T cells were transfected at 37˚C using a third-generation lentiviral packaging system consisting of the CYP17A1 knockout plasmid (cat. no. L25930; Beyotime Biotechnology), pMDLg, Rev and VSV-g plasmids. For each 10-cm dish, 10 μg transfer plasmid, 6.5 μg pMDLg, 2.5 μg Rev and 3.5 μg VSV-g were mixed in 500 μl serum- and antibiotic-free DMEM and transfected using 40 μl Lipo6000 (cat. no. C0526; Beyotime Biotechnology) at 37˚C for 6 h. The medium was replaced after 4-6 h, and viral supernatants were collected at 24 and 48 h, filtered through a 0.45-μm membrane, and used to infect HepG2 cells. Because a fixed MOI was not determined, transduction conditions were optimized empirically using 0.5-2.0 ml viral supernatant/well according to transduction efficiency; 1 ml viral supernatant plus Polybrene (8 μg/ml; cat. no. ST1380; Beyotime Biotechnology) was used for routine infection in 6-well plates. Cells were transduced for 48 h, after which puromycin selection was initiated at 2 μg/ml and maintained at the same concentration for >1 week until stable knockout cells were obtained. The CYP17A1 sgRNA sequence was 5'-GCACCA GGGCACCTTCTCTT-3', and an empty vector was used as the negative control (NC).

Reverse transcription-quantitative PCR (RT-qPCR). Total RNA was extracted from cells using TRIzol® reagent (cat. no. 15596026; Invitrogen; Thermo Fisher Scientific, Inc.). Reverse transcription was performed using HiScript III All-in-one RT SuperMix Perfect for qPCR (cat. no. R333-01; Vazyme Biotech Co, Ltd.) according to the manufacturer's instructions (37˚C for 15 min and 85˚C for 5 sec). qPCR was performed using ChamQ Universal SYBR qPCR Master Mix (cat. no. Q311-02; Vazyme Biotech Co, Ltd.) on cDNA derived from the indicated cells, and relative mRNA expression was calculated using the $2^{-\Delta\Delta Cq}$ method (16) with GAPDH as the internal reference. The thermocycling conditions were 95˚C for 30 sec, followed by 40 cycles of 95˚C for 10 sec and 60˚C for 30 sec. The forward and reverse sequences of all primers used were as follows: CYP17A1 Forward: 5'-AGAAGTTAT CATCAATCTGTGGGC-3', Reverse: 5'-CTGCTCCGAAGG GCAAATA-3'; GAPDH Forward: 5'-GGAAGCTTGTCA TCAATGGAAATC-3', Reverse: 5'-TGATGACCCTTTTGG CTCCC-3'.

Western blotting. Total protein was extracted from cells using RIPA lysis buffer (cat. no. R0010; Beijing Solarbio Science & Technology) supplemented with PMSF. Protein concentrations were determined using a BCA Protein Assay kit. Equal amounts of protein (20 μg/lane) were separated by 10% SDS-PAGE and transferred onto PVDF membranes (cat. no. IPVH00010; MilliporeSigma) at 200 mA for 90 min. The membranes were blocked with 5% non-fat milk in TBST (0.1% Tween-20) at room temperature for 2 h and incubated with

rabbit anti-CYP17A1 antibody (1:1,000; cat. no. bsm-54306R; Bioss) and mouse anti-GAPDH antibody (1:1,000; cat. no. GB12002-100; Wuhan Servicebio Technology) overnight at 4˚C, followed by HRP-conjugated goat anti-rabbit IgG (1:5,000; cat. no. bs-0295G-HRP; Bioss) or HRP-conjugated goat anti-mouse IgG (1:5,000; cat. no. bs-0296G-HRP; Bioss) for 1 h at room temperature. Signals were visualized using BeyoECL Plus chemiluminescence reagent (cat. no. P0018S; Beyotime Biotechnology).

Cell proliferation. Cell Counting Kit-8 (CCK-8) was used to measure cell proliferation (Dojindo Laboratories, Inc.). Cells were seeded in 96-well plates ($5x10^3$ cells/well; six replicates per group) and cell viability was assessed at 0, 24, 48 and 72 h by measuring the OD450. In the functional cell-based assays, the KO-CYP17A1-HepG2 cell line was used as the experimental group, and the KO-Control-HepG2 cell line was used as the control group, to explore the effects of CYP17A1 knockout on the proliferation, migration and invasion ability of the cells.

Cell migration. Cell migration was evaluated using a scratch wound-healing assay. For CYP17A1 knockout experiments, HepG2 NC and KO cells were used; for Saikosaponin A experiments, HepG2 and Huh7 cells were treated with 20 μM Saikosaponin A (Aladdin Biochemical Technology Co., Ltd.). Cells were seeded in 6-well plates and allowed to reach ~100% confluence, scratched with a 200-μl pipette tip, washed three times with PBS, and then cultured in complete medium. Images were captured at 0 and 24 h under an inverted light microscope (magnification, x100), and wound-healing area was quantified using ImageJ (v1.53; National Institutes of Health). For invasion assays, Matrigel (cat. no. M8370; Beijing Solarbio Science & Technology Co., Ltd.) was diluted to ~200 μg/ml, 100 μl was added to each upper chamber and allowed to polymerize at 37˚C for 4 h. Transwell inserts (8-μm pore size; cat. no. WG3422; Wuhan Servicebio Technology) were used. Cells were resuspended at $5x10^5$ cells/ml in serum-free DMEM, and 200 μl cell suspension was added to the upper chamber, while 600 μl complete medium containing 10% FBS was added to the lower chamber. After incubation for 24 h at 37˚C, cells on the lower membrane surface were fixed with 4% paraformaldehyde for 20 min at room temperature, stained with 0.1% crystal violet for 10 min at room temperature, and counted in five random fields under an inverted light microscope (magnification, x200).

Virtual drug screening and molecular docking. The crystal structure of the CYP17A1 protein was retrieved from the PDB database (https://www.rcsb.org/). Water molecules and co-crystallized ligands were removed using PyMOL (v2.5.2; https://pymol.org/2/). A Traditional Chinese Medicine monomer library containing 4,080 compounds was obtained from TargetMol (https://www.targetmol.com/compound-library/traditional-chinese-medicine-monomers-library.html) and virtually screened using AutoDock Vina (v1.2.3). For compounds demonstrating binding energies ≤-10 kcal/mol, molecular docking interactions with the CYP17A1 protein were further analyzed using PyMOL and PLIP software (https://plip-tool.biotec.tu-dresden.de/plip-web/plip/index). Saikosaponin A (Aladdin Biochemical Technology Co., Ltd.) exhibited the highest binding affinity for CYP17A1 and was therefore selected for subsequent *in vitro* validation. The concentration gradient of Saikosaponin A was 0-30 μM. In each experiment, the changes in cell proliferation, migration and invasion ability of HepG2 and Huh7 cells were observed under different concentrations of Saikosaponin A, and the dose-dependent inhibitory effects were analyzed.

Statistical analysis. Experimental data were analyzed using R (v4.4.2; R Foundation for Statistical Computing), GraphPad Prism (v6.0; Dotmatics) and ImageJ (v1.53; National Institutes of Health). Data are presented as mean ± SD. Two-group comparisons were performed using unpaired, two-tailed Student's t-tests. For >2 groups, one-way ANOVA followed by Tukey's post hoc test was used. Time-course proliferation data (CCK-8) were analyzed using two-way ANOVA with Bonferroni post hoc correction. Differential expression analyses were performed using limma (v3.62.2) with Benjamini-Hochberg false discovery rate correction. Pearson correlation analysis was used for WGCNA module-trait associations, Wilcoxon rank-sum tests were used for pan-cancer expression comparisons, and Spearman rank tests were used for correlations with tumor grade and stage. Survival analyses were performed using Kaplan-Meier curves and two-sided log-rank tests. Because late-stage crossover/convergence was observed in Fig. 3C, sensitivity analyses were additionally conducted using i) administrative truncation at 7 years, ii) weighted Gehan-Breslow testing (rho=1) and iii) a two-stage test [TSHRC package (v0.1-6)]. $P<0.05$ was considered statistically significant.

Results

Identification of core genes for sex differences in HCC. A total of 36 candidate genes were screened from samples from male and female patients with HCC by differential expression analysis combined with WGCNA (Fig. 1A and B). The evaluation results showed that the AUC of all models was >0.95 (Fig. 1C) and the residual analysis showed that the model residual values were <0.25 (Fig. 1D), which suggested that the constructed models achieved high discriminatory ability within the cross-validation framework. An intersectional analysis of the top ten genes ranked in importance by the four algorithms identified CYP17A1 and IRX3 as the core genes differentially expressed in HCC samples from men (Fig. 1E).

Validation of sex-specific expression of CYP17A1 and IRX3. To verify the differences in the expression of CYP17A1 and IRX3 in HCC tissues of different sexes, their mRNA expression levels were further evaluated using the comprehensive GEO datasets. The results showed that CYP17A1 and IRX3 mRNA expression levels in HCC tissues from men were significantly higher when compared with those in paired paracancerous tissues ($P<0.001$, Fig. 2A and C). However, there was no significant difference in CYP17A1 and IRX3 mRNA expression between HCC tissues from women and their paired paracancerous tissues ($P>0.05$; Fig. 2B and D). Subgroup analysis further strictly compared tumor tissues to their sex-matched normal controls. The results showed that the expression levels of

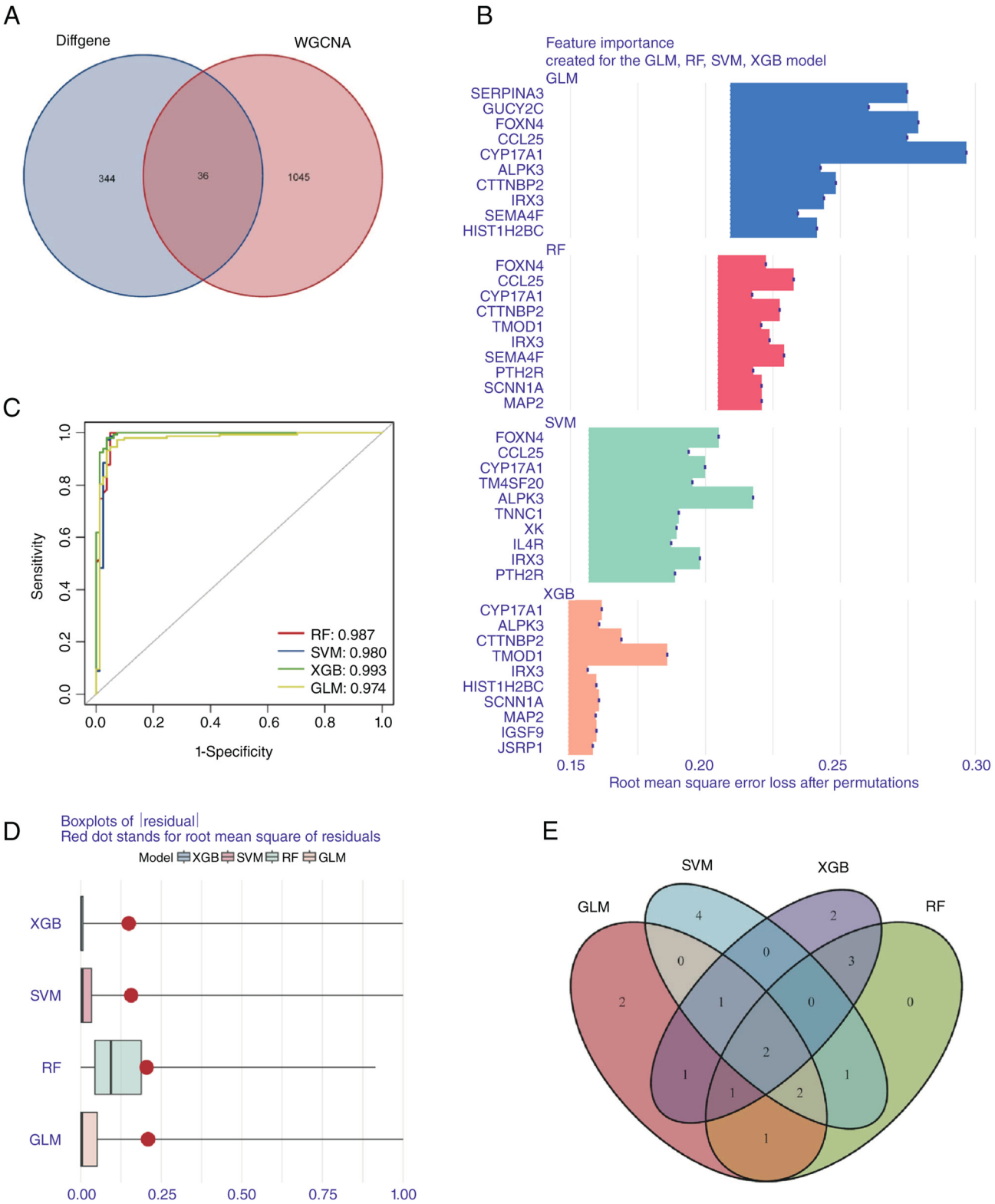

Figure 1. Screening for genes central to sex differences in hepatocellular carcinoma incidence by four machine learning algorithms. (A) Genes that are differentially expressed only in male or female patients intersect with the hub genes of WGCNA. (B) A total of four machine learning algorithms score the importance of 36 candidate core genes. (C) Receiver operating characteristic curve of four machine learning algorithms on model prediction accuracy. (D) Residual values of predictive accuracy of four machine learning algorithms. (E) Venn diagram of the top 10 most important genes for four machine learning algorithms. WGCNA, weighted gene co-expression network analysis; Diffgene, differentially expressed genes; RF, random forest; SVM, support vector machine; XGB, extreme gradient boosting; GLM, generalized linear model.

CYP17A1 and IRX3 were elevated in the male patient HCC group compared with the normal male group (both P<0.001, Fig. 2E and F). These data verified that CYP17A1 and IRX3 were differentially expressed only in male patient HCC, suggesting a sex-specific regulatory pattern. Notably, while CYP17A1 was subjected to further experimental validation

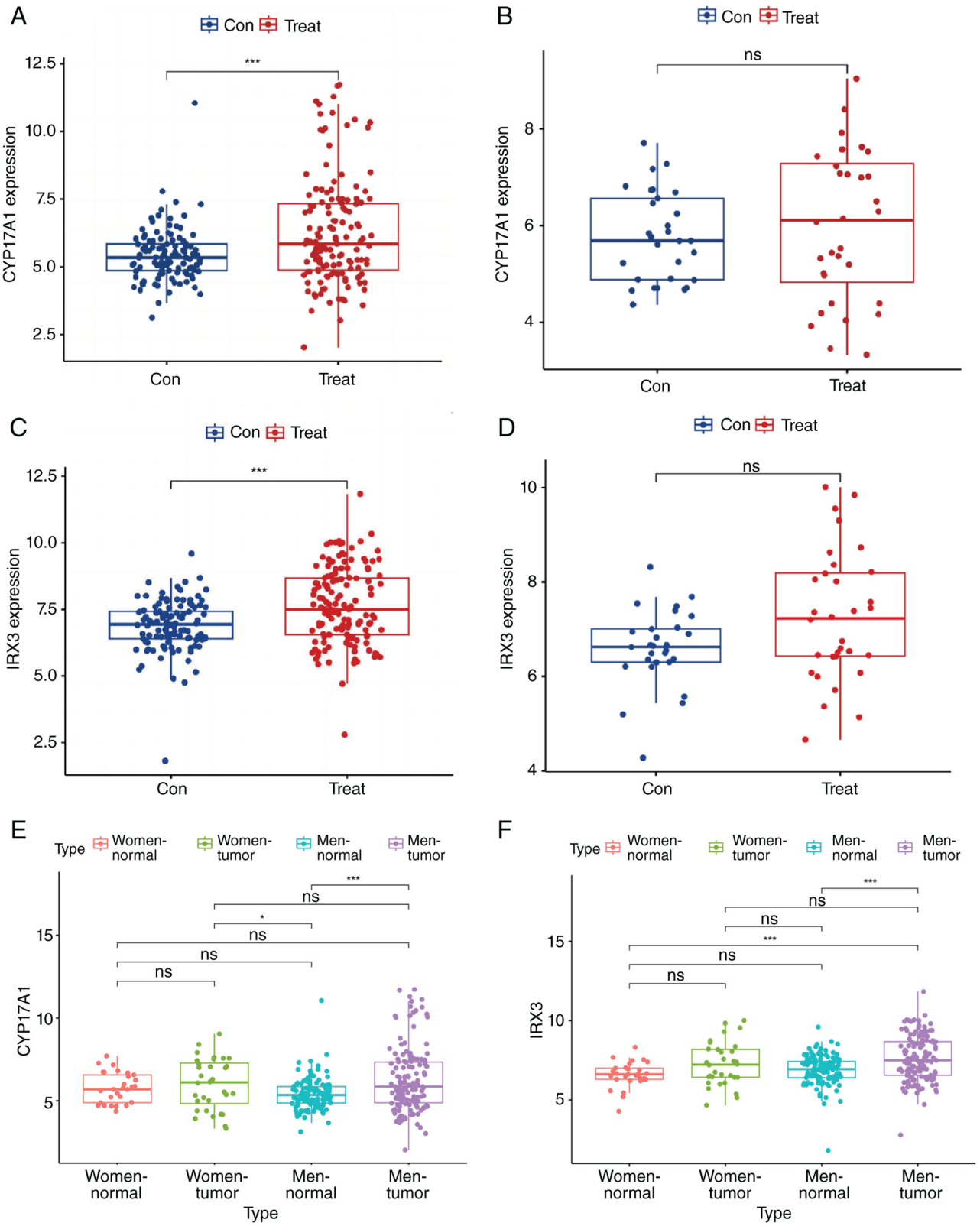

Figure 2. Validation of sex-specific expression of CYP17A1 and IRX3. (A) Expression of CYP17A1 in male patients with HCC. (B) Expression of CYP17A1 in female patients with HCC. (C) Expression of IRX3 in male patients with HCC. (D) Expression of IRX3 in female patients with HCC. (E) Differences in CYP17A1 expression among four groups. (F) Differences in IRX3 expression among four groups. Con, control; Treat, treated; HCC, hepatocellular carcinoma; ns, not significant; *P<0.05; ***P<0.001.

in the present study, IRX3 lacks functional validation and currently serves as a hypothesis-generating candidate gene warranting future investigation.

Pan-cancer exploratory analysis and HCC-specific clinical relevance of CYP17A1. To explore the broader context of CYP17A1, a pan-cancer expression analysis was performed.

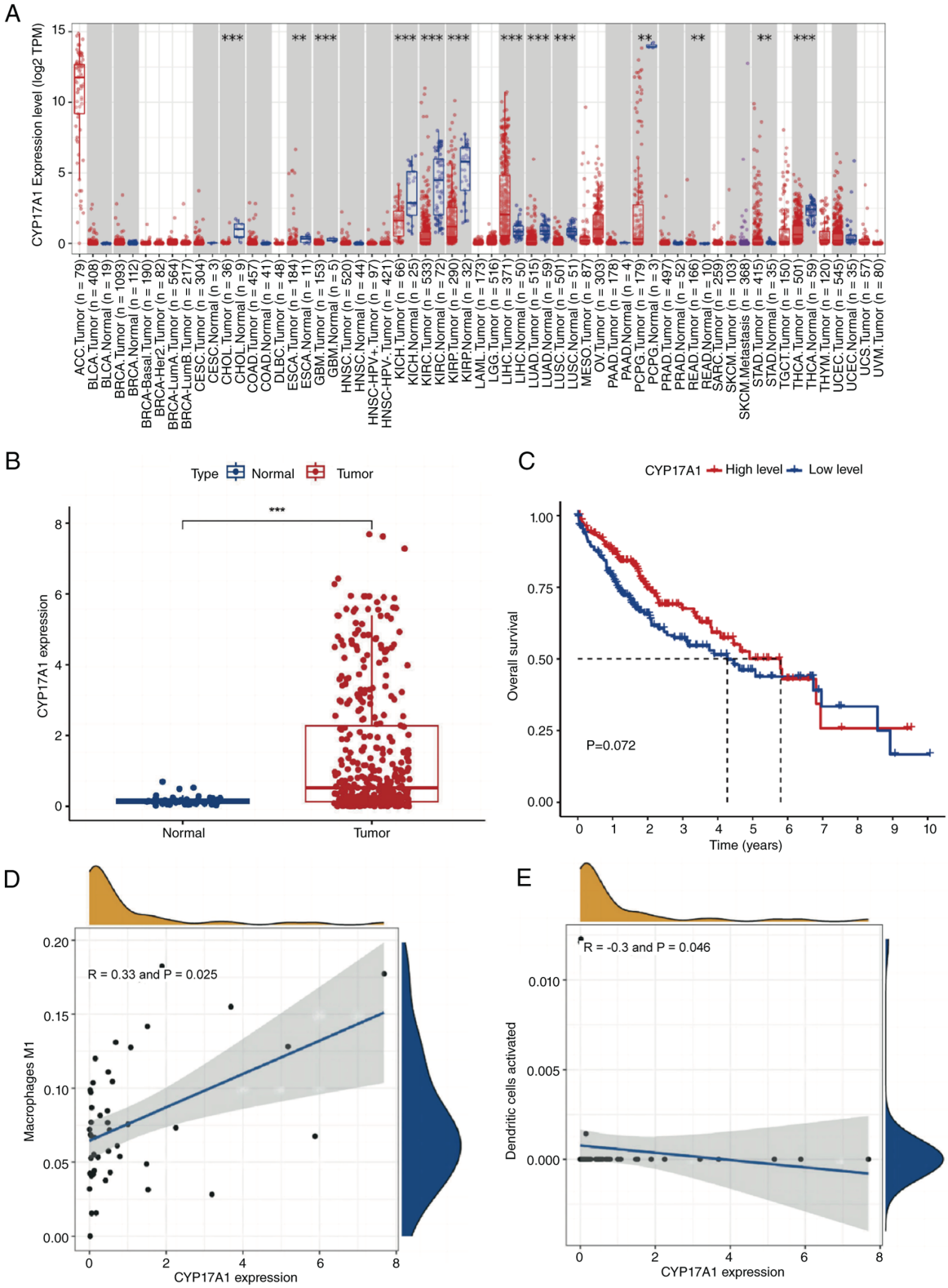

Figure 3. Pan-cancer analysis of CYP17A1 and single-gene analysis in HCC. (A) Expression of CYP17A1 in various tumors. (B) Expression of CYP17A1 in HCC and paracancerous tissues. (C) Kaplan-Meier survival curve of CYP17A1 high- and low-expression groups. (D) Correlation of CYP17A1 expression with M1 macrophages. (E) Correlation of CYP17A1 expression with activated dendritic cells. HCC, hepatocellular carcinoma; **P<0.01; ***P<0.001.

It showed that CYP17A1 was generally lowly expressed in the majority of tumors, such as cholangiocarcinoma and renal cancer, but exhibited specific high expression in HCC, rectal adenocarcinoma and gastric cancer (P<0.001; Fig. 3A). In a subsequent, strictly independent HCC-specific validation using the TCGA-LIHC dataset, the expression of CYP17A1 in HCC tissues was significantly increased compared with that in normal tissues (P<0.001; Fig. 3B). The survival analysis with the standard log-rank test indicated no statistically significant association with overall survival (log-rank

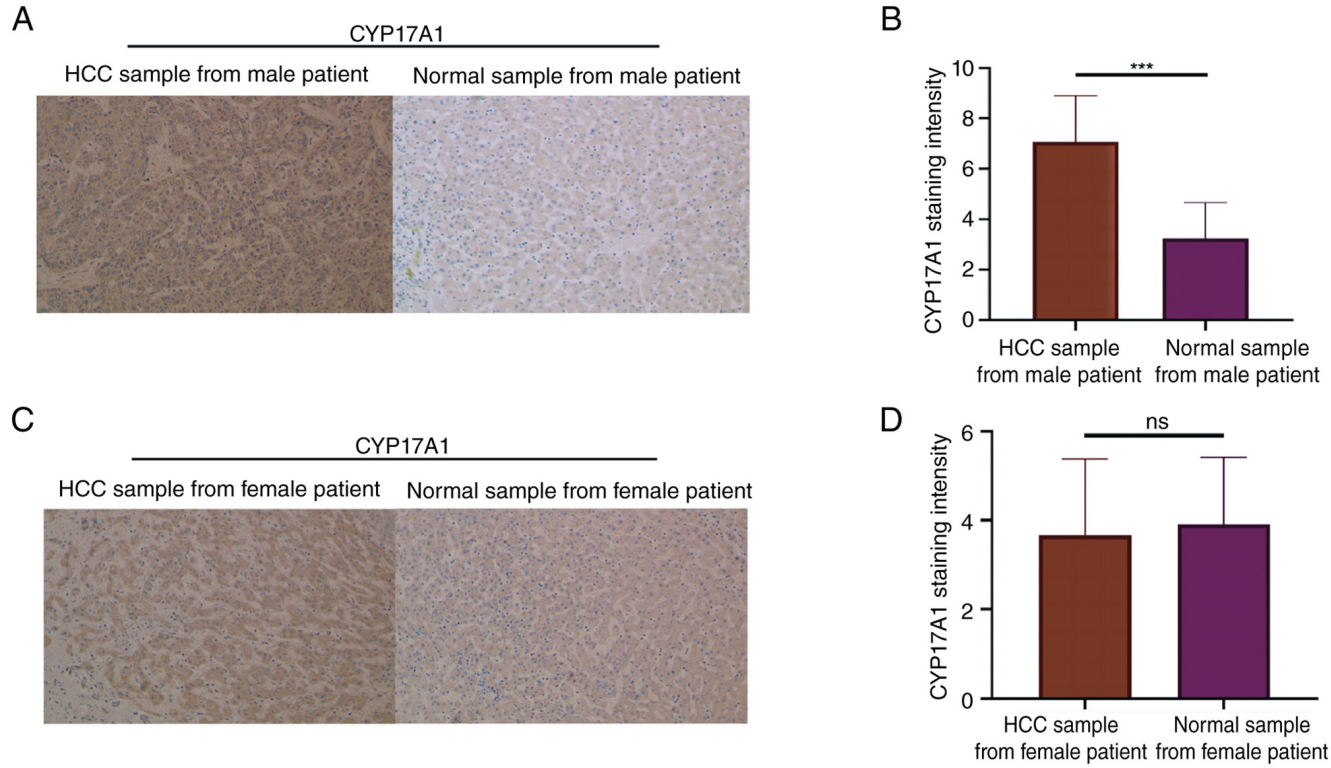

Figure 4. Immunohistochemical validation of CYP17A1 protein in HCC tissues. (A) Immunohistochemical images from HCC and paired paracancerous tissues from a male patient. (B) Difference in CYP17A1 expression between HCC and paired paracancerous tissues from a male patient. (C) Immunohistochemical images from HCC and paired paracancerous tissues from a female patient. (D) Difference in CYP17A1 expression between HCC and paired paracancerous tissues from a female patient. HCC, hepatocellular carcinoma; ns, not significant; ***P<0.001.

P>0.05; Fig. 3C). Considering the late-stage curve crossover, additional sensitivity analyses were performed: A 7-year truncated log-rank (P>0.05), weighted Gehan-Breslow (P<0.05) and two-stage test (P=>0.05), indicating a trend-level but non-robust prognostic association. Furthermore, the correlation analysis of CYP17A1 expression and immune cell infiltration showed a positive correlation with M1 macrophage infiltration (R=0.33; P<0.05; Fig. 3D) and a negative correlation with activated dendritic cell infiltration (R=-0.3; P<0.05; Fig. 3E).

Validation of protein expression and functional mechanisms for CYP17A1. Validation of CYP17A1 protein expression by IHC suggested that CYP17A1 protein was significantly higher in male patient HCC tissues when compared with that in paracancerous tissues (P<0.001; Fig. 4A and B). By contrast, there was no significant difference in female patient samples (P>0.05; Fig. 4C and D). This protein-level result was consistent with the mRNA expression patterns. To investigate the functional mechanisms of CYP17A1 *in vitro*, stable knockout of CYP17A1 was established using lentivirus specifically in the HepG2 cell line. The mRNA (Fig. 5A) and protein (Fig. 5B and C) expression levels of HepG2 cells were significantly reduced following CYP17A1 knockout (P<0.001). Functional cell-based assays demonstrated that knockout of CYP17A1 inhibited HepG2 cell proliferation at 72 h (P<0.001; Fig. 5D; Table SI), reduced migration capacity by 78% in the scratch healing assay (P<0.001; Fig. 5E and F) and attenuated invasive capacity by 65% in the Transwell assay (P<0.001; Fig. 5G and H). These results suggest that CYP17A1 potentially contributes to the malignant phenotypes of HCC cells *in vitro*.

Virtual screening and in vitro validation of putative CYP17A1-targeting compounds. In a virtual screening of 4,080 herbal monomers using the CYP17A1 protein as a receptor, molecular docking predicted that Saikosaponin A possessed the highest binding affinity for the CYP17A1 protein, with a binding energy of -12.5 kcal/mol and the formation of 6 pairs of hydrogen bonds (Table I; Fig. 6A). Based on these predictive docking data, Saikosaponin A was selected as a putative compound targeting CYP17A1 for further cellular validation. *In vitro* experiments evaluated the effects of Saikosaponin A on HCC cell phenotypes. Cell proliferation assays demonstrated that treatment with different concentrations of Saikosaponin A significantly inhibited the proliferation of HepG2 and Huh7 cells in a dose- and time-dependent manner (Fig. 6B and C; Tables SII and SIII). Furthermore, after treatment with 20 μM of Saikosaponin A for 24 h, the scratch healing rates of HepG2 and Huh7 cells were reduced by 63 and 58%, respectively (P<0.001; Fig. 6D-G). In response to the image-quality query, representative wound-healing control images were replaced with fields showing continuous near-confluent monolayers outside the wound area, and quantification was recalculated from qualified fields (Fig. 6F). Similarly, Transwell assays revealed that the number of invading HepG2 and Huh7 cells was reduced by 52 and 48%, respectively (P<0.01; Fig. 6H-J). These inhibitory effects on migration and invasion were also

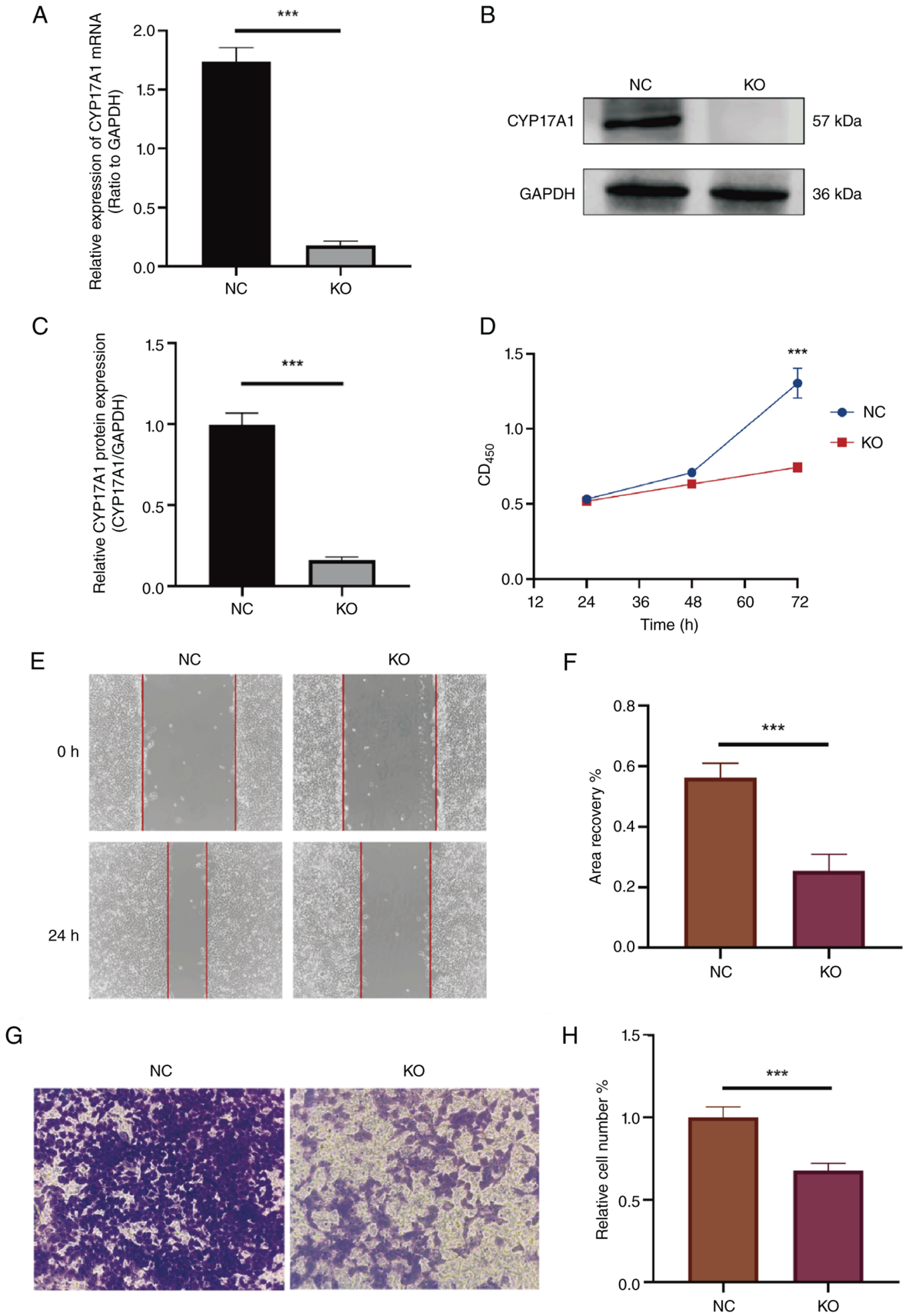

Figure 5. A preliminary investigation of the role of CYP17A1 in hepatocellular carcinoma based on CYP17A1 knockout and functional cell-based assays. (A) Relative expression of CYP17A1 mRNA in the control group and the KO group. (B) CYP17A1 protein expression in the control group and the KO group. (C) Difference analysis of CYP17A1 relative expression levels in the control group and the KO group. (D) Proliferation of the control group and the CYP17A1 gene KO group over 72 h. (E) Effect of CYP17A1 gene knockout on HepG2 migration ability under the microscope at 0 and 24 h (magnification, x100). (F) Difference analysis of scratch healing rates between the control group and the KO group. (G) Transwell results of the control group and the KO group under a microscope. (H) Analysis of the difference in the relative number of invasive cells in the control group and the KO group. NC, the control group; KO, the knockout group; OD, optical density; ***P<0.001.

Table I. Virtual screening of interactions between the top 10 drugs with the highest affinity and CYP17A1 protein.

Ligand	Chemical formula	Pubchem no.	Molecular weight	Affinity (kcal/mol)	Interactions
Wilforlide A	$C_{30}H_{46}O_3$	158477	454.7	-13.7	5 hydrophobic interactions
Epifriedelanol	$C_{30}H_{52}O$	119242	428.7	-13.4	7 hydrophobic interactions
δ-Amyrenone	$C_{30}H_{48}O$	5318261	424.7	-13.2	6 hydrophobic interactions
Arnidiol	$C_{30}H_{50}O_2$	10478550	442.7	-13.1	7 hydrophobic interactions
β-Amyrin	$C_{30}H_{50}O$	73145	426.7	-12.8	7 hydrophobic interactions
Taraxasterol	$C_{30}H_{50}O$	115250	426.7	-12.7	7 hydrophobic interactions
Saikosaponin A	$C_{42}H_{68}O_{13}$	167928	781	-12.5	4 hydrophobic interactions, 6 H-bonds
Taraxastery acetate	$C_{32}H_{52}O_2$	13889352	468.8	-12.4	7 hydrophobic interactions
Glycyrrhetinic acid	$C_{30}H_{46}O_4$	10114	470.7	-12.3	6 hydrophobic interactions
Saikosaponin D	$C_{42}H_{68}O_{13}$	107793	781	-12.3	4 hydrophobic interactions, 5 H-bonds

dose-dependent. This suggests that Saikosaponin A exhibits a significant inhibitory effect on the malignant phenotypes of HCC cells, highlighting it as a potential therapeutic candidate, although direct CYP17A1-dependency requires further rescue validations.

Discussion

HCC is considerably more prevalent in men than in women (3,4), and exploring the underlying molecular mechanisms of this sex difference is a prominent research topic in oncology. In the present study, by integrating multi-omics data and machine learning algorithms, CYP17A1 and IRX3 were computationally identified as key candidate genes associated with sex differences in HCC. The present study explored their potential mechanisms in influencing the progression of HCC, possibly through the modulation of the androgen signaling pathway. Virtual drug screening and *in vitro* functional validation revealed that the traditional Chinese medicine monomer Saikosaponin A could inhibit the malignant phenotypes of HCC cells, highlighting a putative new candidate for the treatment of HCC.

Cytochrome P450 enzyme (CYP17A1) is a key steroidogenic enzyme that carries out an important role in the development of several tumors. It catalyzes the conversion of pregnenolone to dehydroepiandrosterone and promotes testosterone synthesis (17,18). Niu *et al* (19) found that CYP17A1 knockout *in vitro* considerably inhibits the proliferation and invasive ability of glioma cells, while promoting apoptosis. Moreover, in women, it has also been hypothesized that the androgen receptor (AR) can substitute for estrogen-dependent signaling to stimulate transcription of steroid-responsive genes that drive breast cancer (20). Altered expression levels of genes involved in the androgen expression pathway have been suggested as one of the key factors that may be associated with sex differences in the development of these tumors (21-23). The present study found that CYP17A1 was specifically highly expressed in male patient HCC tissues. This result aligns with the mechanism that CYP17A1 drives tumor progression by promoting androgen synthesis in prostate cancer. At the same time, previous studies have found that cell culture medium containing certain levels of androgens promotes the proliferative capacity of glioma cells and that average serum testosterone levels are markedly higher in patients with glioma (24,25). Taken together, the present study speculates that the higher incidence of HCC in men may be associated with CYP17A1-mediated elevations of local androgen levels, which potentially promote tumor cell proliferation through the activation of AR signaling pathways. Notably, while IRX3 was identified alongside CYP17A1 as a core gene, to the best of our knowledge, its association with sex differences has not yet been previously reported. IRX3 regulates cell differentiation in embryonic development, and its function in HCC may involve aberrant activation of the Wnt/β-catenin pathway. However, because no functional validation was performed for IRX3 in the present study, it must be viewed as a hypothesis-generating candidate gene that requires rigorous future experimental validation.

In the field of tumor therapy, natural products have been an important source for the development of chemotherapeutic drugs (26). In order to explore potential interventions, the present study found that Saikosaponin A, an important bioactive triterpene glycoside compound extracted from Chaihu (Radix Bupleuri), has a predicted binding affinity with CYP17A1 (27). Du *et al* (28) found that Saikosaponin A triggered apoptosis and inhibited the PI3K/Akt signaling pathway, exerting an anti-cervical cancer effect. Wang *et al* (29) showed that Saikosaponin A may induce apoptosis in gastric cancer cells. Zhang *et al* (30) found that Saikosaponin A has potent anti-angiogenic activity and inhibits tumor growth mainly by blocking the VEGFR2 signaling pathway. In the present study, molecular docking predicted a high affinity (binding energy=-12.5 kcal/mol) with the CYP17A1 protein, and *in vitro* experiments confirmed that it dose-dependently inhibited HCC cell migration and invasion. However, molecular docking alone does not establish direct enzymatic inhibition, presenting a limitation of the present study. Without target-dependency or rescue experiments, it remains inconclusive whether CYP17A1 is the primary functional target mediating the observed anti-migration and anti-invasion effects, as Saikosaponin A may act through pleiotropic mechanisms. In the future, enzymatic activity

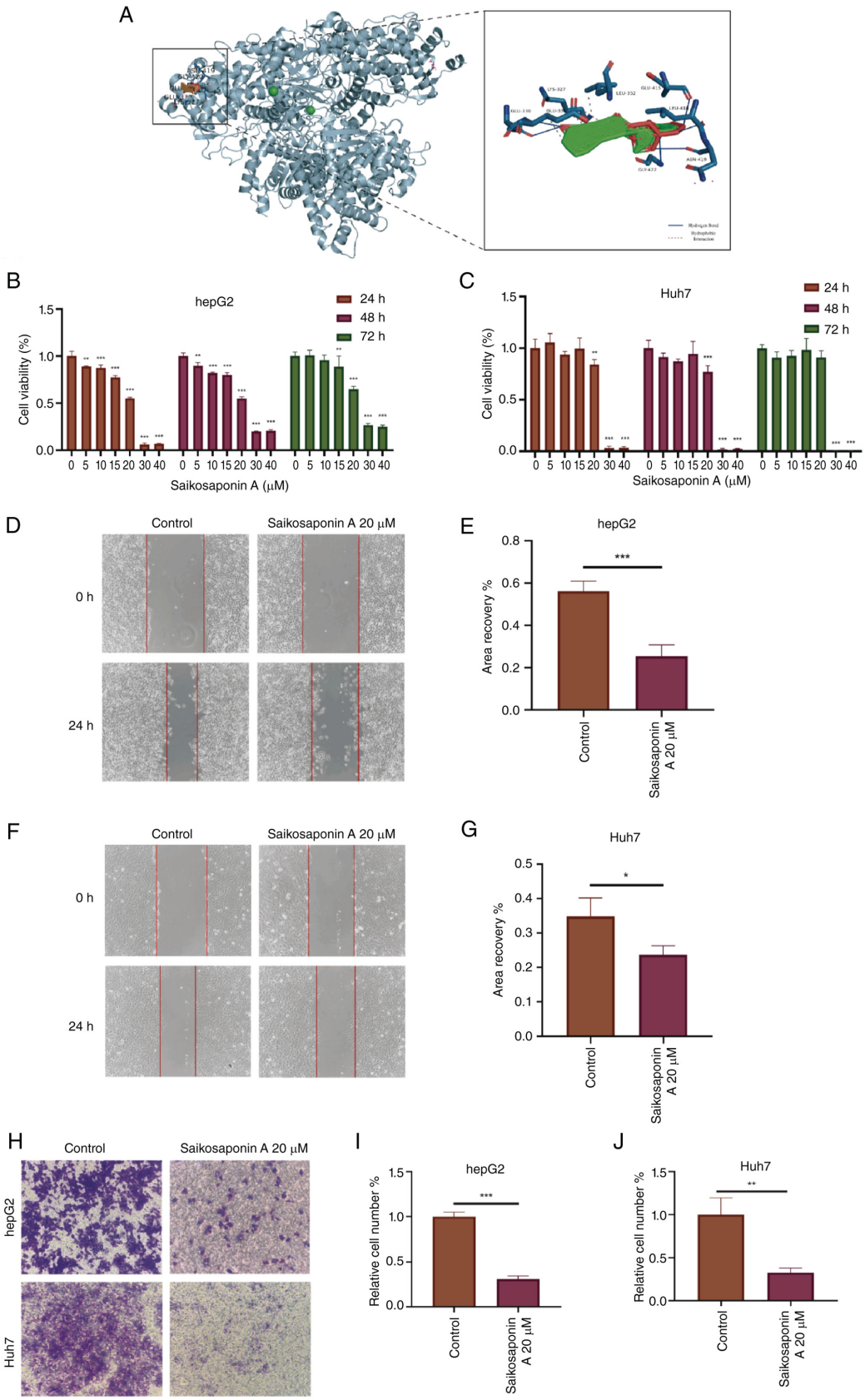

Figure 6. Screening and *in vitro* validation of drug candidates targeting CYP17A1. (A) Saikosaponin A interacts with the CYP17A1 protein. (B) Cell proliferation of HepG2 cells treated with different concentrations of Saikosaponin A for 24, 48 and 72 h. (C) Cell proliferation of Huh7 cells treated with different concentrations of Saikosaponin A for 24, 48 and 72 h. (D) Scratch healing of HepG2 control and Saikosaponin A-treated cells at 0 and 24 h under the microscope (magnification, x100). (E) Difference analysis of the scratch healing rate between the Control group and the drug-treated group. (F) Scratch healing of Huh7 Control and Saikosaponin A-treated cells at 0 and 24 h under the microscope (magnification, x100). (G) Difference analysis of scratch healing rates between the Control group and the drug-treated group. (H) Cell invasion of HepG2 and Huh7 cells in the Control group and the drug-treated group under the microscope. (I) Difference analysis of the relative cell number of HepG2 cells in the Control group and the drug-treated group. (J) Difference analysis of relative cell numbers of Huh7 cells in the Control group and the drug-treated group. *P<0.05; **P<0.01; ***P<0.001.

assays and rescue experiments in CYP17A1-knockout cells are necessary to comprehensively reveal its precise antitumor mechanism.

The sex-specific expression of CYP17A1 in HCC provides a new direction for the development of personalized treatment regimens for male patients. Abiraterone, an inhibitor of CYP17A1, has achieved significant efficacy in the treatment of prostate cancer, suggesting that targeting this gene axis might also be applicable to HCC (31). Despite these promising results, several limitations must be addressed. First, the clinical IHC sample size was small (n=18), and the subgroup analyses did not adequately control for potential confounding factors such as age, viral hepatitis status and alcohol consumption. Thus, it is premature to attribute the observed expression differences of CYP17A1 solely to biological sex without validation in a larger, multifactorial clinical cohort. Second, as indicated by the primary log-rank analysis (P=0.0726) and additional crossover-sensitive analyses (7-year truncated log-rank P=0.0933; weighted Gehan-Breslow P=0.0334; two-stage P=0.0908), the association between CYP17A1 expression and overall survival was not robust across methods, suggesting its role may be mechanistic or therapeutic rather than acting as a robust prognostic biomarker. Third, owing to the scarcity of comprehensive multi-omics HCC datasets with perfectly matched biological sex annotations, we were unable to construct a fully independent external validation cohort, which implies that the risk of overfitting cannot be entirely ruled out. Finally, the involvement of the androgen pathway is inferred from previous literature and our bioinformatics findings, rather than directly demonstrated through androgen level measurements or rescue experiments in this study. The *in vivo* efficacy and safety of Saikosaponin A also require systematic evaluation using humanized HCC mouse or patient-derived xenograft models in future studies.

In conclusion, the present study suggests that CYP17A1 is a key gene associated with HCC in men, potentially driving malignant phenotypes via the androgen signaling axis, while IRX3 emerges as a novel hypothesis-generating candidate. Furthermore, Saikosaponin A is highlighted as a putative therapeutic candidate. Future research efforts will focus on constructing a CYP17A1 conditional knockout mouse model to validate its necessity in HCC occurrence, performing target-dependency assays for Saikosaponin A and resolving the interaction network of IRX3 with sex-related signaling pathways. These studies are expected to provide a more solid theoretical basis for sex-specific strategies and personalized treatment in HCC.

Acknowledgements

Not applicable.

Funding

The author(s) declare that financial support was received for the research, authorship and/or publication of this article. The present study was supported by the following projects: General Project of the Joint Special Project of Local Universities in Yunnan Province (grant nos. 202001BA070001-064 and 202101BA070001-102), Dali University Doctoral Research Start-up Fund Project (grant no. KYBS2018012), Open Project of Yunnan Provincial Key Laboratory of Entomological Biopharmaceutical R&D (grant no. AG2024002), Clinical Medicine Discipline Team Building Project of the First Affiliated Hospital of Dali University (grant no. DFYYB2024026), Open Project of Key Laboratory of Screening and Research of Resistant Plant Resources in West Yunnan, Yunnan Province (grant no. APKL2101) and Yunnan College Students Innovation and Entrepreneurship Training Project Fund (grant no. S202410679068).

Availability of data and materials

The data generated in the present study may be requested from the corresponding author. The datasets analyzed for this study can be found in the GEO database (https://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE19665, GSE54236, GSE84402 and GSE121248, and in the TCGA-LIHC dataset via the TIMER2.0 platform.

Authors' contributions

ZW and JN contributed to conceptualization, data curation, formal analysis, investigation, methodology, project administration, supervision, CYP17A1 protein expression and functional validation, visualization, writing of the original draft, and review and editing. HZ and WL contributed to conceptualization, formal analysis, investigation, supervision, visualization, and review and editing. ZL and XY contributed to methodology, bioinformatics data analysis, visualization and writing of the original draft. YZhao and YW contributed to data curation, formal analysis and writing of the original draft. RQ contributed to data processing, conducting experiments, visualization and writing of the original draft. CL contributed to bioinformatics data analysis, visualization and writing of the original draft. QR contributed to the acquisition of clinical data, interpretation of data, and review of the manuscript. YL and YZhang contributed to conceptualization, data curation, funding acquisition, investigation, methodology, project administration, resources, supervision, visualization, writing of the original draft and review and editing.

Ethics approval and consent to participate

The study of human tumor cells was approved by the ethics committee of the First Affiliated Hospital of Dali University, Dali, China (approval no. OFY20221122001). The present study was conducted in accordance with the local legislation and institutional requirements.

Patient consent for publication

All patients provided written informed consent for the publication of their anonymized clinical data obtained and analyzed in the present study. The consent process ensured that participants were aware that their data would be used for research purposes and potentially included in scientific publications, with all personally identifiable information removed to protect their privacy.

Competing interests

The authors declare that they have no competing financial interests.

References

1. Petrick JL and McGlynn KA: The changing epidemiology of primary liver cancer. Curr Epidemiol Rep 6: 104-111, 2019.
2. Kim E and Viatour P: Hepatocellular carcinoma: Old friends and new tricks. Exp Mol Med 52: 1898-1907, 2020.
3. Kulik L and El-Serag HB: Epidemiology and management of hepatocellular carcinoma. Gastroenterology 156: 477-491.e1, 2019.
4. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA and Jemal A: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68: 394-424, 2018.
5. Nakatani T, Roy G, Fujimoto N, Asahara T and Ito A: Sex hormone dependency of diethylnitrosamine-induced liver tumors in mice and chemoprevention by leuprorelin. Jpn J Cancer Res 92: 249-256, 2001.
6. Naugler WE, Sakurai T, Kim S, Maeda S, Kim K, Elsharkawy AM and Karin M: Gender disparity in liver cancer due to sex differences in MyD88-dependent IL-6 production. Science 317: 121-124, 2007.
7. Manieri E, Herrera-Melle L, Mora A, Tomás-Loba A, Leiva-Vega L, Fernández DI, Rodríguez E, Morán L, Hernández-Cosido L, Torres JL, *et al*: Adiponectin accounts for gender differences in hepatocellular carcinoma incidence. J Exp Med 216: 1108-1119, 2019.
8. Lip GYH, Nieuwlaat R, Pisters R, Lane DA and Crijns HJGM: Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: The euro heart survey on atrial fibrillation. Chest 137: 263-272, 2010.
9. O'Mahony C, Jichi F, Pavlou M, Monserrat L, Anastasakis A, Rapezzi C, Biagini E, Gimeno JR, Limongelli G, McKenna WJ, *et al*: A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (HCM risk-SCD). Eur Heart J 35: 2010-2020, 2014.
10. Bogard N, Linder J, Rosenberg AB and Seelig G: A deep neural network for predicting and engineering alternative polyadenylation. Cell 178: 91-106.e23, 2019.
11. Deng YB, Nagae G, Midorikawa Y, Yagi K, Tsutsumi S, Yamamoto S, Hasegawa K, Kokudo N, Aburatani H and Kaneda A: Identification of genes preferentially methylated in hepatitis C virus-related hepatocellular carcinoma. Cancer Sci 101: 1501-1510, 2010.
12. Villa E, Critelli R, Lei B, Marzocchi G, Cammà C, Giannelli G, Pontisso P, Cabibbo G, Enea M, Colopi S, *et al*: Neoangiogenesis-related genes are hallmarks of fast-growing hepatocellular carcinomas and worst survival. Results from a prospective study. Gut 65: 861-869, 2016.
13. Wang H, Huo X, Yang XR, He J, Cheng L, Wang N, Deng X, Jin H, Wang N, Wang C, *et al*: STAT3-mediated upregulation of lncRNA HOXD-AS1 as a ceRNA facilitates liver cancer metastasis by regulating SOX4. Mol Cancer 16: 136, 2017.
14. Wang SM, Ooi LLPJ and Hui KM: Identification and validation of a novel gene signature associated with the recurrence of human hepatocellular carcinoma. Clin Cancer Res 13: 6275-6283, 2007.
15. Leek JT, Johnson WE, Parker HS, Jaffe AE and Storey JD: The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28: 882-883, 2012.
16. Livak KJ and Schmittgen TD: Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods 25: 402-408, 2001.
17. Abul-Hajj YJ, Iverson R and Kiang DT: Metabolism of pregnenolone by human breast cancer. Evidence for 17 alpha-hydroxylase and 17,20-lyase. Steroids 34: 817-827, 1979.
18. Parker C and Sartor O: Abiraterone and increased survival in metastatic prostate cancer. N Engl J Med 365: 767, 2011.
19. Niu WX, Zhou CX, Cheng CD, Bao DJ, Dong YF, Li DX, Yang Y, He H and Niu CS: Effects of lentivirus-mediated CYP17A1 gene silencing on the biological activity of glioma. Neurosci Lett 692: 210-215, 2019.
20. Robinson JL, Macarthur S, Ross-Innes CS, Tilley WD, Neal DE, Mills IG and Carroll JS: Androgen receptor driven transcription in molecular apocrine breast cancer is mediated by FoxA1. Embo J 30: 3019-3027, 2011.
21. Camargo MC, Goto Y, Zabaleta J, Morgan DR, Correa P and Rabkin CS: Sex hormones, hormonal interventions, and gastric cancer risk: A meta-analysis. Cancer Epidemiol Biomarkers Prev 21: 20-38, 2012.
22. Marcinkiewicz K, Scotland KB, Boorjian SA, Nilsson EM, Persson JL, Abrahamsson PA, Allegrucci C, Hughes IA, Gudas LJ and Mongan NP: The androgen receptor and stem cell pathways in prostate and bladder cancers (review). Int J Oncol 40: 5-12, 2012.
23. Chen PJ, Yeh SH, Liu WH, Lin CC, Huang HC, Chen CL, Chen DS and Chen PJ: Androgen pathway stimulates microRNA-216a transcription to suppress the tumor suppressor in lung cancer-1 gene in early hepatocarcinogenesis. Hepatology 56: 632-643, 2012.
24. Ceccarelli I, Rossi A, Maddalena M, Weber E and Aloisi AM: Effects of morphine on testosterone levels in rat C6 glioma cells: Modulation by anastrozole. J Cell Physiol 221: 1-4, 2009.
25. Bao D, Cheng C, Lan X, Xing R, Chen Z, Zhao H, Sun J, Wang Y, Niu C, Zhang B and Fang S: Regulation of p53wt glioma cell proliferation by androgen receptor-mediated inhibition of small VCP/p97-interacting protein expression. Oncotarget 8: 23142-23154, 2017.
26. Mann J: Natural products in cancer chemotherapy: Past, present and future. Nat Rev Cancer 2: 143-148, 2002.
27. Li X, Li X, Huang N, Liu R and Sun R: A comprehensive review and perspectives on pharmacology and toxicology of saikosaponins. Phytomedicine 50: 73-87, 2018.
28. Du J, Song D, Cao T, Li Y, Liu J, Li B and Li L: Saikosaponin-A induces apoptosis of cervical cancer through mitochondria- and endoplasmic reticulum stress-dependent pathway in vitro and in vivo: Involvement of PI3K/AKT signaling pathway. Cell Cycle 20: 2221-2232, 2021.
29. Wang C, Zhang R, Chen X, Yuan M, Wu J, Sun Q, Miao C and Jing Y: The potential effect and mechanism of Saikosaponin A against gastric cancer. BMC Complement Med Ther 23: 295, 2023.
30. Zhang P, Lai X, Zhu MH, Long M, Liu XL, Wang ZX, Zhang Y, Guo RJ, Dong J, Lu Q, *et al*: Saikosaponin A, a triterpene saponin, suppresses angiogenesis and tumor growth by blocking VEGFR2-mediated signaling pathway. Front Pharmacol 12: 713200, 2021.
31. Fujita K and Nonomura N: Role of androgen receptor in prostate cancer: A review. World J Mens Health 37: 288-295, 2019.