scientific reports

OPEN

Differentially expressed discriminative genes and significant meta-hub genes based key genes identification for hepatocellular carcinoma using statistical machine learning

Md. Al Mehedi Hasan[1,2], Md. Maniruzzaman[1,3] & Jungpil Shin[1]✉

Hepatocellular carcinoma (HCC) is the most common lethal malignancy of the liver worldwide. Thus, it is important to dig the key genes for uncovering the molecular mechanisms and to improve diagnostic and therapeutic options for HCC. This study aimed to encompass a set of statistical and machine learning computational approaches for identifying the key candidate genes for HCC. Three microarray datasets were used in this work, which were downloaded from the Gene Expression Omnibus Database. At first, normalization and differentially expressed genes (DEGs) identification were performed using limma for each dataset. Then, support vector machine (SVM) was implemented to determine the differentially expressed discriminative genes (DEDGs) from DEGs of each dataset and select overlapping DEDGs genes among identified three sets of DEDGs. Enrichment analysis was performed on common DEDGs using DAVID. A protein-protein interaction (PPI) network was constructed using STRING and the central hub genes were identified depending on the degree, maximum neighborhood component (MNC), maximal clique centrality (MCC), centralities of closeness, and betweenness criteria using CytoHubba. Simultaneously, significant modules were selected using MCODE scores and identified their associated genes from the PPI networks. Moreover, metadata were created by listing all hub genes from previous studies and identified significant meta-hub genes whose occurrence frequency was greater than 3 among previous studies. Finally, six key candidate genes (TOP2A, CDC20, ASPM, PRC1, NUSAP1, and UBE2C) were determined by intersecting shared genes among central hub genes, hub module genes, and significant meta-hub genes. Two independent test datasets (GSE76427 and TCGA-LIHC) were utilized to validate these key candidate genes using the area under the curve. Moreover, the prognostic potential of these six key candidate genes was also evaluated on the TCGA-LIHC cohort using survival analysis.

Hepatocellular carcinoma (HCC) is the 3rd leading cause of cancer deaths globally[1]. Globally, more than of 80% liver cancers are responsible for HCC[2] and its prevalence is high in males compared to females[3]. It usually occurs in people aged 30–50 years[3]. Different factors such as hepatitis B or hepatitis C[4,5], alcohol abuse, smoking, obesity, and type 2 diabetes (T2D) were significantly associated with HCC[6]. Among them, Hepatitis B is one of the prominent risk factors for the development of HCC, responsible for 50% of cases[7]. Despite various treatment approaches, namely radiotherapy, chemotherapy, and target therapy have been commonly used to improve the prognosis and recurrence of HCC. Nevertheless, the survival rate of HCC patients is still low[8]. As a result, the risks of cancer death are still increased due to the lack of early detection and diagnosis of genes and

[1]School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Fukushima 965-8580, Japan. [2]Department of Computer Science and Engineering, Rajshahi University of Engineering & Technology, Rajshahi 6204, Bangladesh. [3]Statistics Discipline, Khulna University, Khulna 9208, Bangladesh. ✉email: jpshin@u-aizu.ac.jp

limited treatment facilities. Therefore, it is essential to develop a system for identifying the key or core genes for early detection and better prognosis of HCC.

Recently, bioinformatics analysis has been widely utilized to determine the key prognostic genes or biomarkers as well as their associated molecular pathways for multiple cancers, including HCC[8–58]. Zhou et al.[35] identified 15 prognostic biomarkers as well as their associated gene ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway using bioinformatics analysis. Chen et al.[39,59,60] also identified 11 potential biomarkers that can play crucial roles in the development and progression of HCC patients. Qiang et al.[40] proposed five core genes which were significantly associated with early diagnosis and poor prognosis of HBV-HCC. Wang et al.[41] identified 36 hub DEGs and illustrated that 10 candidate genes out of the 36 have significant effect on the tumorigenesis and progression of HCC. Among them, eight candidate genes were inversely related to the survival rate of HCC patients. Dai et al.[61] proposed a prognostic model for predicting the prognosis of HCC patients. They identified 17 genes that were potentially associated with the prognosis of HCC patients. These 17 genes were used to make a prognostic model using the Cox hazard regression model and validated its performance using the TCGA and GSE14520 datasets. They showed that six genes were involved in the prognosis of HCC patients. Most researchers simply used hub genes derived from the PPI network to identify the key or core genes. One of the major challenges in studying genetic data was the identification of relevant biomarkers or genes. Recently, machine learning (ML)-based techniques have gained more attraction to address this problem[59,60,62–66]. Despite the fact that several studies have been carried out for the identification and development of potential candidate genes for HCC[8–58,67], it remains a challenging issue and still has some scope for more research for the identification of potential genes as well as understanding molecular mechanisms for the development, pathogenies, and progression of HCC.

In this work, we used three microarray gene expression (MGE) datasets as training sets to determine the key or core candidate genes for HCC. First, we selected individual DEGs for three datasets. Secondly, support vector machine (SVM) with radial basis function (RBF) was implemented on the identified DEGs from each of the three datasets and calculated the classification accuracy of each DEG. We selected the DEGs from each of the three datasets that provided a classification accuracy of more than 80.0%. At the same time, the overlapping or shared DEGs were identified from three datasets. These overlapping or shared DEGs were called differentially expressed discriminative genes (DEDGs). Thirdly, DAVID was used to perform enrichment analysis on common DEDGs. Fourthly, PPI networks were constructed using STRING and visualized using Cytoscape. Then the hub genes were identified using degree, maximum neighborhood component (MNC), maximal clique centrality (MCC), closeness, and betweenness on the basis of cytoHubba. After that, the central hub genes were determined by overlapping or shared hub genes from the degree, MNC, MCC, centralities of closeness, and betweenness. Molecular Complex Detection (MCODE) was performed for cluster or module analysis and determined the important or significant modules as well as their associated genes. Moreover, the significant meta-hub genes were determined from meta-hub genes, which were extracted from existing studies. The key or core candidate genes were determined among the central hub genes, potential module hub genes, and significant meta-hub genes, which can be easily discriminated against in HCC patients compared to healthy controls. Furthermore, we used another two independent test datasets for the validation as well as to show the discriminative power of the key candidate genes. We also performed a survival analysis of the identified key candidate genes for HCC patients. Therefore, the overall flowchart of our proposed system to determine key candidate genes for HCC is presented in Fig. 1.

Results

Identification of DEGs from each dataset.

We implemented limma for identifying DEGs from each of the three GEO datasets (GSE36376, GSE39791, and GSE57957). Using the threshold of $|log_2FC|> 1$, and adj.p-value < 0.01, we identified 699 (up-regulated: 431 vs. down-regulated: 268), 428 (up-regulated: 88 vs. down-regulated: 340 DEGs), and 413 DEGs (up-regulated: 107; down-regulated: 306) DEGs between HCC and healthy controls from GSE36376, GSE39791, and GSE57957 datasets and their volcano plots and heatmap were presented in Fig. 2.

Identification of common DEDGs using SVM.

SVM with RBF kernel was applied on the identified DEGs (699 DEGs for GSE36376; 428 DEGs for GSE39791; and 413 DEGs for GSE57957) of each dataset in order to identify the DEDGs of HCC patients. Then, the classification accuracy was computed per gene for DEGs from each dataset. The calculation procedure is clearly discussed in the methodology section. The classification accuracies of all DEGs for individual datasets were ordered in descending order of magnitude, which is presented in Fig. 3. As shown in Fig. 3, we observed that a total of 502 from GSE36376, 169 from GSE39791, and 242 from GSE57957 DEGs were selected as DEDGs because their classification accuracy was more than or equal to 80.0%. Furthermore, 75 common DEDGs were determined among the identified DEDGS from GSE36376, GSE39791, and GSE57957 datasets, which is shown in Fig. 4.

Enrichment analysis of common DEDGS.

Enrichment analysis was conducted on 75 shared or overlapping DEDGs clearly grasp the mechanism and development of HCC. The functional characteristics of DEDGs were explored using GO and KEGG pathway analysis. The GO analysis was partitioned into three groups: biological process (BP), cellular component (CC), and morphological component. Using p-values (< 0.05), we identified the significant GO and KEGG pathways, and chose the top five prominent GO terms and KEGG pathway. The top five GO terms, including BP, CC, and MF, are presented in Table 1.

For BP-based GO terms, the common DEDGs were strongly enriched with retinol metabolic process, cellular response to cadmium ion, retinoid metabolic process cellular response to copper ion, and steroid catabolic process. Moreover, the extracellular region, extracellular exosome, extracellular space, high-density lipoprotein

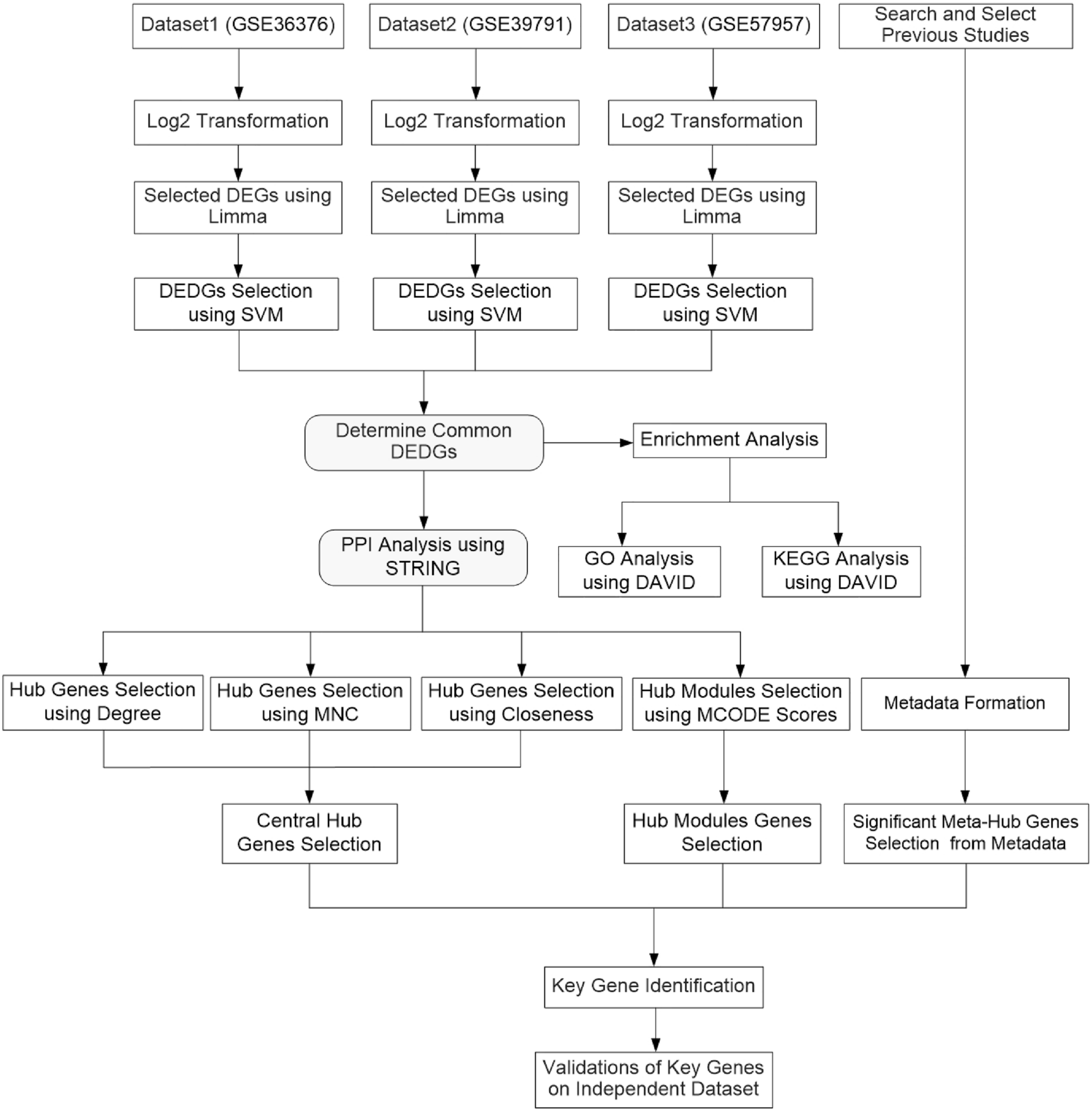
Dataset1 (GSE36376)
Dataset2 (GSE39791)
Dataset3 (GSE57957)
Search and Select Previous Studies
Log2 Transformation
Log2 Transformation
Log2 Transformation
Selected DEGs using Limma
Selected DEGs using Limma
Selected DEGs using Limma
DEDGs Selection using SVM
DEDGs Selection using SVM
DEDGs Selection using SVM
Determine Common DEDGs
Enrichment Analysis
PPI Analysis using STRING
GO Analysis using DAVID
KEGG Analysis using DAVID
Hub Genes Selection using Degree
Hub Genes Selection using MNC
Hub Genes Selection using Closeness
Hub Modules Selection using MCODE Scores
Metadata Formation
Central Hub Genes Selection
Hub Modules Genes Selection
Significant Meta-Hub Genes Selection from Metadata
Key Gene Identification
Validations of Key Genes on Independent Dataset

Figure 1. Flowchart of proposed system for the identification of key candidate genes for HCC.

particle, and apical plasma membrane were found to be top CC, which were significantly enriched with common DEDGs. As shown in Table 1, MF group GO terms, including retinol dehydrogenase activity; oxidoreductase activity; androsterone dehydrogenase activity; androstan-3-alpha,17-beta-diol dehydrogenase activity; and steroid dehydrogenase activity, were mainly enriched with common DEDGs.

The study of the KEGG pathway for common DEDGs is displayed in Table 2. As shown in Table 2, the common DEDGs were significantly associated with multiple pathways such as retinol metabolism, metabolic pathways, tryptophan metabolism, steroid hormone biosynthesis, and drug metabolism-cytochrome P450.

PPI network construction and central hub genes identification. STRING was utilized to build a PPI network to show the significant connections between proteins encoded by common DEDGs. Cytoscape was used to show the PPI network, which had 51 nodes and 144 edges (see Fig. 5a). Five hub gene-based identification algorithms, including the degree of connectivity, MNC, MCC, closeness, and betweenness in the Cytoscape plug-in cytoHubba, were implemented to determine the hub genes from PPI networks. Then we chose the top 30 hub genes from each algorithm. We made a Venn diagram among the five algorithms, which is shown in Fig. 5b. As shown in Fig. 5b, eight overlapping central hub genes were identified among these algorithms. These eight

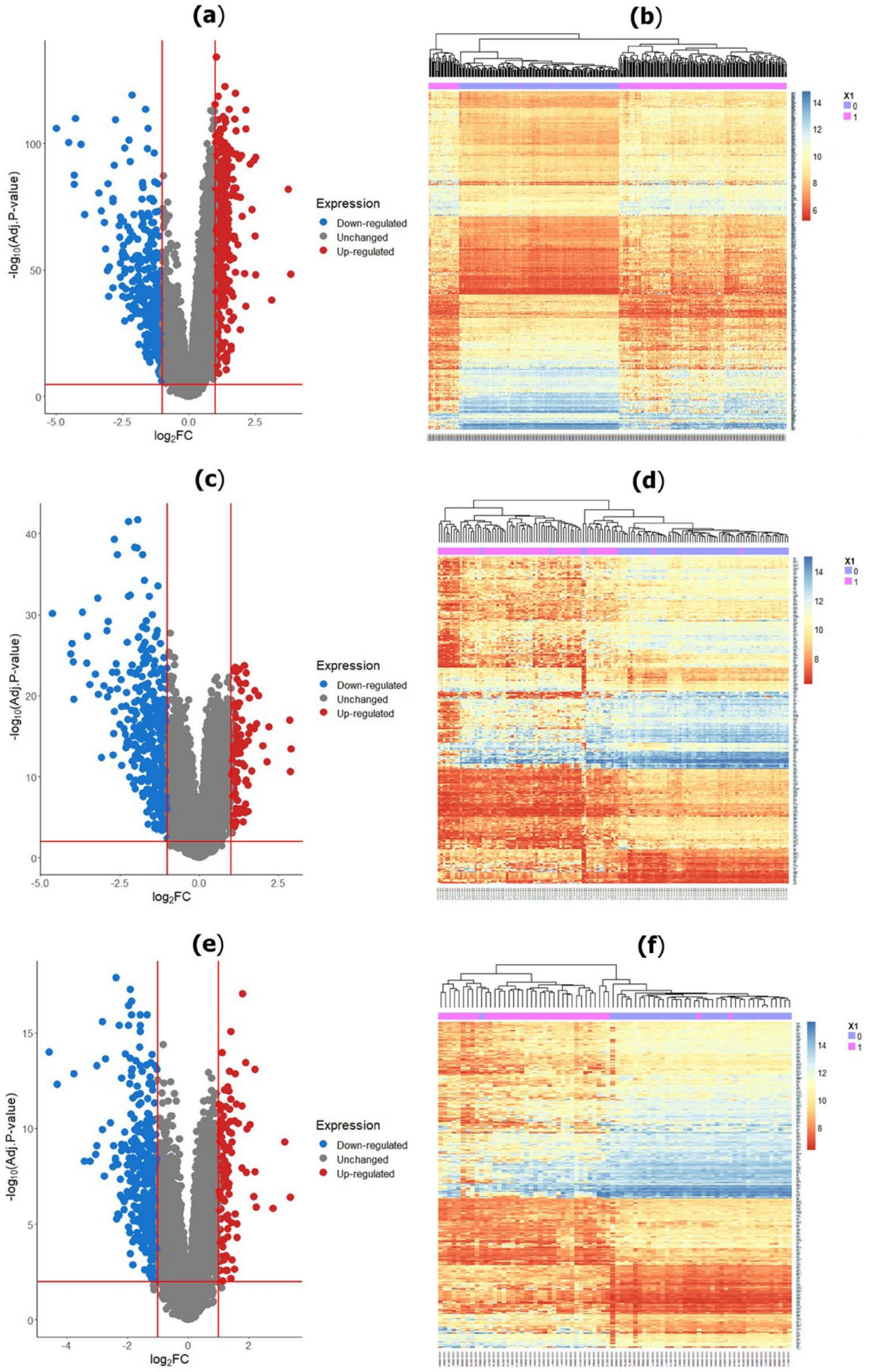

Figure 2. Volcano plot and heatmap of DEGs for each GEO dataset were generated using "ggplot2" version 3.3.6 package[110] (https://cran.r-project.org/package=ggplot2) and "NMF" version 0.24.0 package[111] (https://cran.r-project.org/package=NMF) in R . (**a**) Volcano plot and (**b**) heatmap of GSE36376 dataset; (**c**) Volcano plot and (**d**) heatmap of GSE39791 dataset; (**c**) Volcano plot and (**d**) heatmap of GSE57957. Dodger blue represents down-regulated, gray represents no significant genes, and fire brick represents up-regulated DEGs.

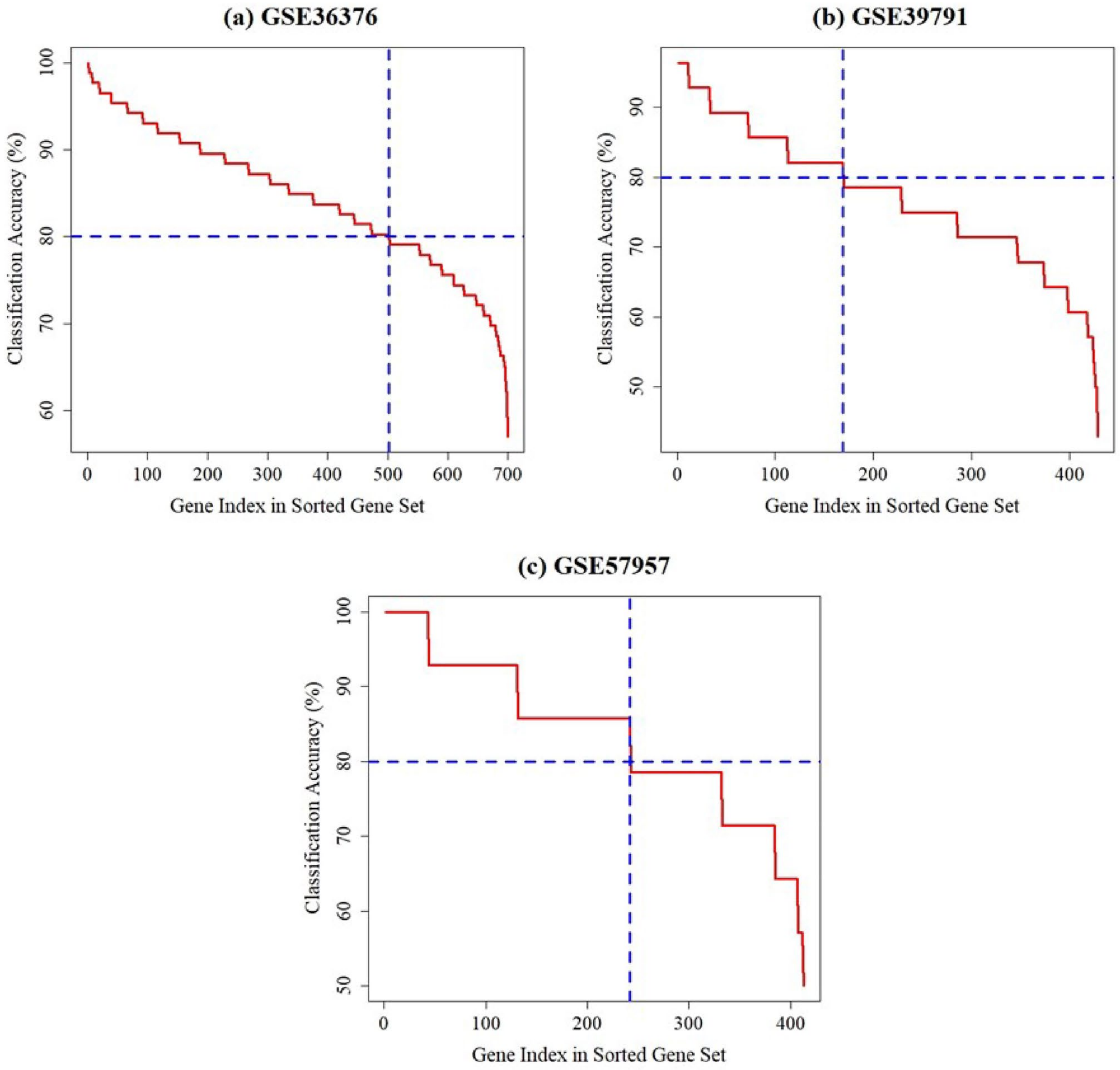

Figure 3. Classification accuracy of individual genes using SVM for three GEO datasets: (**a**) GSE36376; (**b**) GSE39791, and (**c**) GSE57957.

central hub genes were NUSAP1, TOP2A, CDC20, PRC1, UBE2C, ASPM, PNPLA7, and MT1E, which were utilized to determine the key or core genes for HCC.

Hub modules and its associated genes identification. Module or cluster analysis was performed using MCODE to determine the prominent modules. Three clusters or modules were generated using MCODE and provided 3–6 MCODE scores. We chose the prominent modules that provided the MCODE scores of ≥ 5 and the number of nodes ≥ 5. Finally, we chose module 1 as a prominent hub module that contained 6 nodes and 30 edges with the highest MCODE scores of 6 and their PPI networks were displayed in Fig. 6. The correspondence six genes were treated as hub module genes.

Identification of significant meta-hub genes from metadata. We reviewed 52 existing studies related to gene identification of HCC patients[8–58]. We listed their hub genes in order to make metadata which were presented in Table 3. To make metadata, we extracted 10 hub genes from Maddah et al.[9], 5 hub genes from Yan et al.[10], 20 from Zhao et al.[11], 7 from Zhao et al.[12], 10 from Liu et al.[13], 11 from Meng et al.[14], 42 from Rosli et al.[15], 5 from Zhang et al.[8], 5 from Li et al.[16], 8 from Li et al.[17], 5 from Tian et al.[18], 12 from Wan et al.[19], 10 from Zhu et al.[20], 10 from Wang et al.[21], 9 from Zhou et al.[22], 10 from Zhang et al.[23], 18 from Mou et al.[24], 8 from Wu et al.[25], 9 from Gui et al.[26], 10 from Wang et al.[27], 28 from Lu and Zhu[28], 6 from Bhatt et al.[29], 10 from Zhang

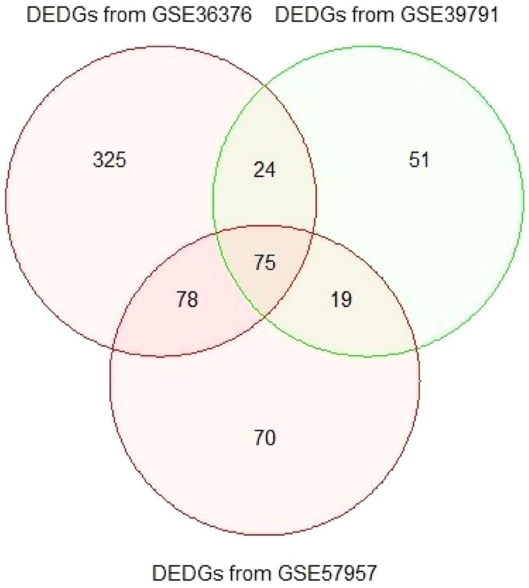

Figure 4. Identification of common or overlapping DEDGs among DEDGs from GSE36376, GSE39791, and GSE57957 datasets.

Category	GO ID	Descriptions	Count	p-value
BP	GO:0042572	Retinol metabolic process	7	3.41×10^{-8}
	GO:0071276	Cellular response to cadmium ion	5	1.38×10^{-5}
	GO:0001523	Retinoid metabolic process	4	1.35×10^{-4}
	GO:0071280	Cellular response to copper ion	4	1.51×10^{-4}
	GO:0006706	Steroid catabolic process	3	1.88×10^{-4}
CC	GO:0005576	Extracellular region	19	3.79×10^{-4}
	GO:0070062	Extracellular exosome	18	0.00173
	GO:0005615	Extracellular space	16	0.0031
	GO:0034364	High-density lipoprotein particle	3	0.004
	GO:0016324	Apical plasma membrane	6	0.011
MF	GO:0004745	Retinol dehydrogenase activity	5	5.81×10^{-7}
	GO:0016491	Oxidoreductase activity	9	2.46×10^{-6}
	GO:0047023	Androsterone dehydrogenase activity	3	3.56×10^{-4}
	GO:0047044	Androstan-3-alpha,17-beta-diol dehydrogenase activity	3	4.56×10^{-4}
	GO:0016229	Steroid dehydrogenase activity	3	0.001

Table 1. GO analysis of common DEDGs in terms of BP, CC, and MF. Top 5 items were selected.

Pathway ID	Descriptions	Count	p-value
hsa00830	Retinol metabolism	6	3.28×10^{-5}
hsa01100	Metabolic pathways	21	7.24×10^{-5}
hsa00380	Tryptophan metabolism	4	0.001
hsa00140	Steroid hormone biosynthesis	4	0.004
hsa00982	Drug metabolism-cytochrome P450	4	0.007

Table 2. KEGG pathway analysis of common DEDGs. Top five items were selected.

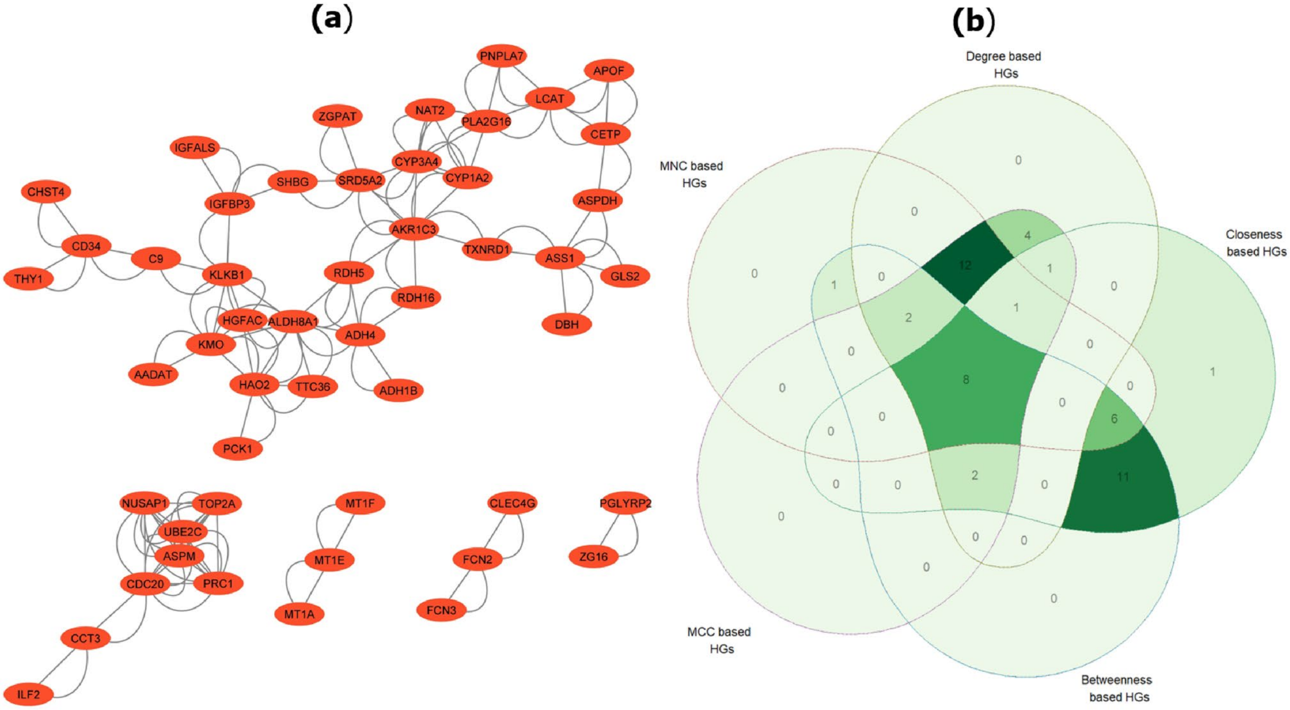

Figure 5. PPI network and Venn diagram for common DEDGs and central hub genes. (**a**) PPI network of common DEDGs with 51 nodes and 144 edges which was generated by Cytoscape 3.9.1[118] (www.cytoscape.org); (**b**) identification of central hub genes among five methods (Degree, MNC, MCC, Closeness, and Betweenness based HGs). Here, HGs represent the hub genes.

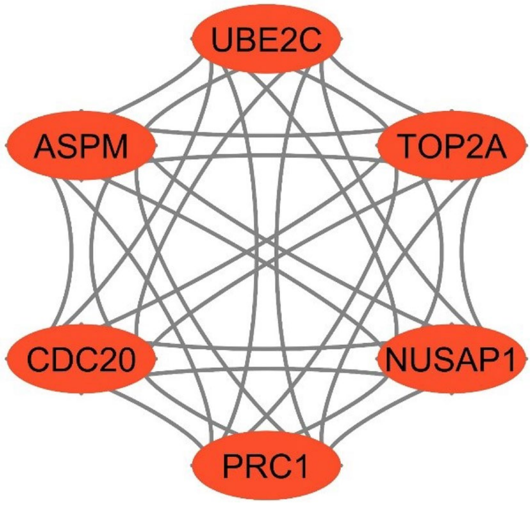

Figure 6. PPI network of module 1 with 6 nodes and 30 edges which was generated by Cytoscape 3.9.1[118] (www.cytoscape.org).

et al.[30], 13 from Jiang et al.[31], 20 from Zhang et al.[32], 12 from Wu et al.[33], 5 from Nguyen et al.[34], 15 from Zhou et al.[35], 6 from Yu et al.[36], 10 from Kakar et al.[37], 10 from Ji et al.[38], 11 from Chen et al.[39], 10 from Qiang et al.[40], 10 from Wang et al.[41], 10 from Zhang et al.[42], 14 from Kim et al.[43], 10 from Zhang et al.[44], 14 from Sha et al.[45], 10 from Chen et al.[46], 4 from He et al.[47], 10 from Zhang et al.[48], 4 from Hu et al.[49], 9 from Zhang et al.[50], 15 from Li et al.[51], 5 from Cao et al.[52], 7 from Yang et al.[53], 5 from Wang et al.[54], 9 from Jiang et al.[55], 16 from Li et al.[56], 15 from Xing et al.[57], 10 from Zhu W et al.[58], and 20 from Dai et al.[61]. Now, we took the union of extracted hub genes and got 214 hub genes as meta-hub genes. At the same time, we also computed the frequency of each meta-hub gene depending on how many studies got that gene as hub gene and selected 52 significant meta-hub genes because their frequency was more than 3. These selected 52 significant meta-hub genes were utilized for the determination of key genes.

SN	Authors	NHG	Associated hub genes	SN	Authors	NHG	Associated hub genes
1	Maddah et al.[9]	10	BUB1, CDCA8, DLGAP5, ASPM, POLQ,CENPE, WDHD1, HELLS, TRIP13, DEPDC1	27	Nguyen et al.[34]	5	TOP2A, RRM2, NEK2, CDK1, CCNB1
2	Yan et al.[10]	5	CCNA2, PLK1, CDC20, UBE2C, AURKA	28	Zhou et al.[35]	15	DTL, CDK1, CCNB1, RACGAP1, ECT2, NEK2, BUB1B, PBK, TOP2A, ASPM, HMMR, RRM2, CDKN3, PRC1, ANLN
3	Qian et al.[11]	16	ADNP, CASP2, CBX1, CPSF6, DHX9, HCFC1, ILF3, RCC2, KANSL1, NAA40, NCOA6, RALGAPB, SENP1, SMARCD1, YEATS2	29	Yu et al.[36]	6	TOP2A, MAD2L1, CDC6, CHEK1, UBE2C, CCNB1
4	Zhao et al.[12]	7	CCNA2, CCNB1, CDK1, MAD2L1, TOP2A, RRM2, NDC80	30	Kakar et al.[37]	10	CDK1,CCNA2, CCNB1, CCNB2, BUB1, NDC80, BUB1B, NCAPG, MAD2L1, CDC20
5	Liu et al.[13]	10	CYP3A4, UGT1A6, AOX1, UGT1A4, UGT2B15, CDK1, CCNB1, MAD2L1, CCNB2, CDC20	31	Ji et al.[38]	10	CDK1, CCNB1, CCNB2, PBK, ASPM,NDC80, AURKA, TPX2, KIF2C, CENPF
6	Meng et al.[14]	11	CDK1, CCNB2, CDC20, CCNB1, TOP2A, CCNA2,PBK, MELK, TPX2, KIF20A, AURKA	32	Chen et al.[39]	11	RRM2, NDC80, ECT2, CCNB1, ASPM, CDK1,PRC1, KIF20A, DTL, TOP2A, PBK
7	Rosli et al.[15]	42	CDK1, PPAP2B, CCNA2, SQLE, CCNB1,SULTIA3, NUSAP1, MAD2L1, LCAT, TOP2A, CETP, CCNB2, CFP, KIF11,FOS, NCAPG, CDK1, CDC20, TOP2A, TTK, C7, AURKA, C6, RRM2, NDC80, ACLY, MSH2, ESR1, CENPA, NDC80, MELK, CXCL12, PBK, DTL, NR1I2, IGF1, BUB1B, HBA1, PRC1, SPTBN2, KIF2C, CYP1A2	33	Qiang et al.[40]	10	CDK1, CCNB2, CDC20, BUB1, BUB1B, CCNB1, NDC80, CENPF, MAD2L1, NUF2
8	Zhang et al.[8]	10	GMPS, ACACA, ALB, TGFB1, KRAS, ERBB2, BCL2, EGFR, STAT3, CD8A	34	Wang et al.[41]	10	CDKN3, TOP2A, UBE2C, CDC20, PBK, ASPM, KIF20A, NCAPG, CCNB2, CYP3A4
9	Li et al.[16]	5	SPP1, COL1A2, IGF1, LGALS3, LPA	35	Zhang et al.[42]	10	CCNB1, AURKA, TOP2A, NEK2, CENPF, ASPM, KIF20A, NCAPG, CCNB2, CYP3A4
10	Li et al.[17]	8	BUB1, BUB1B, CCNA2, CCNB1, CDC20, CDK1, MAD2L1, CCNB2	36	Kim et al.[43]	14	ANLN, ASPM, BUB1B, CCNB1, CDK1, CDKN3, ECT2,HMMR, NEK2, PBK, PRC1, RACGAP1, RRM2, TOP2A
11	Tian et al.[18]	5	CDC20, TOP2A, RRM2, UBE2C, AOX1	37	Zhang et al.[44]	10	CCNB1, CDC20, CCNB2, CDK1, SPC24, CENPW, ZWINT, PTTG1, AURKA, UBE2C
12	Wan et al.[19]	12	GF1, IGF2, NDC80, CDK1, CENPF, CDCA8, CCNB1, BIRC5, NCAPG, SPC25, CDCA5, CENPU	38	Sha et al.[45]	14	TOP2A, HMMR, DTL, CCNB1, NEK2, PBK, RACGAP1, PRC1, CDK1, RRM2, ECT2, BUB1B, ANLN, ASPM
13	Zhu et al.[20]	10	CDK1, TOP2A, CCNB1, CDC20, PLK1, BIRC5, CCNB2, FOS, AURKA, AURKB	39	Chen et al.[46]	10	TOP2A, CCNB2, PRC1, RACGAP1, AURKA, CDKN3, NUSAP1, ASPM, CDCA5, NCAPG
14	WANG et al.[21]	10	TOP2A, CDK1, ITGA2, PLK1, ESR1, CCNB2, AURKA, BUB1, CCNA2, BUB1B	40	He et al.[47]	4	CDK1, PBK, RRM2, and ASPM
15	Zhou et al.[22]	9	ASPM, AURKA, CCNB2, CDKN3, MELK, NCAPG, NUSAP1, PRC1, TOP2A	41	Zhang et al.[48]	10	NEK2, ANLN, TOP2A, CENPF, ASPM, CDC20, CDK1, CCNB1, ECT2, CCNB2
16	Zhang et al.[23]	10	CDK1, CCNB1, AURKA, CCNA2, KIF11, BUB1B, TOP2A, TPX2, HMMR, CDC45	42	Hu et al.[49]	4	JUN, EGR1, MYC, CDKN1A
17	Mou et al.[24]	18	TOP2A, FOS, TK1, CDC20, ESR1, CCNB2, CXCL12, FOXO1, HMMR, VWF, ACSM3, COL4A1, ZIC2, RFC4, TXNRD1, GNAO1, CYP3A4, RAP2A	43	ZHANG et al.[50]	9	ALDH2, PPTG1, CYP2C8, ADH4, ADH1B, CYP2C8, CDC20, TOP2A, CCNB2
18	Wu et al.[25]	8	CDKN3, CDK1, CCNB1, TOP2A, CCNA2, CENPE, KCCNB2, PRC1, RRM2	44	Li et al.[51]	15	TOP2A, CDK1, CCNB1, BUB1, CENPF, CCNB2, TTK, KIF2C, HMMR, MELK, CENPE, KIF20A, KIF4A, PBK, DLGAP5
19	Gui et al.[26]	4	MT1X, BMI1, CAP2, TACSTD2	45	Cao et al.[52]	5	MCM3, CHEK1, KIF11, PBK, S100A9
20	Wang et al.[27]	10	TOP2A, CDK1, NDC80, CCNB1, HMMR, CENPF, AURKA, CDKN3, FOXM1, PTTG1	46	Yang et al.[53]	7	PITX2, PNCK, GLIS1, SCNN1G, MMP1, ZNF488, SHISA9
21	Lu[28]	28	NDUFC2, NDUFS7, NDUFB1, NDUFB9, NDUFA2, NDUFB7, NDUFA11, NDUFAF6, NDUFS6, NDUFB8, MRPS28, MRPS18A, MRPL14, MRPL12, MRPL54, MRPL55, MRPL52, MRPL13, MRPL27, MRPL24, NUF2, DSN1, GADD45GIP1, CHCHD1, STAG2, PPP1CC, CKAP5, ZWINT	47	WANG et al.[54]	5	CDK1, CCNB1, CCNB2, MAD2L1, TOP2A
22	Bhatt et al.[29]	6	MSH3, DMC1, ALPP, IL10, ZNF223, HSD17B7	48	Jiang et al.[55]	9	ANLN, BIRC5, BUB1B, CDC20, CDCA5, CDK1, NCAPG, NEK2, TOP2A
23	Zhang et al.[30]	10	CDC20, CCNB1, EIF4A3, H2AFX, NOP56, RFC4,NOP58, AURKA, PCNA, FEN1	49	Li et al.[56]	16	BIRC5, BUB1, CCNB2, CDC20, CDC25C, CDK1, CEP55, CXCL12, FOS, PRC1, KIF20A, NUSAP1, KIF2C, RACGAP, SPC24, TOP2A
24	Jiang et al.[31]	13	TLR1, TLR4, TLR7, TLR8, RIPK2, YWHAZ, FOS, FOSL2, HIF1A, FASLG, CCL4, CDK1A, DDIT3	50	Xing et al.[57]	15	TOP2A, PCNA, CCNB2, AURKA, CDKN3, BUB1, RFC4, CEP55, DLGAP5, MCM2, PRC1, RACGAP1, TPX2, CDC20, MCM4
25	Zhang et al.[32]	20	CDK1, CCNB1, CCNB2, CDC20, CCNA2, AURKA, MAD2L1, TOP2A, BUB1B, BUB1, ESR1, IGF1, FTCD, CYP3A4, SPP2, C8A, CYP2E1, TAT, F9, CYP2C9	51	Zhu et al.[58]	10	UBE2C, CDK1, TK, NCAPG, TOP2A, AURKA, MAD2L1, TOP2A, BUB1B, BUB1, RAD51AP1, ASPM, PBK, DLGAP5, NUSAP1
26	Wu et al.[33]	12	TTK, NCAPG, TOP2A, CCNB1, CDK1, PRC1, RRM2, UBE2C, ZWINT, CDKN3, AURKA, RACGAP1	52	Dai et al.[61]	20	ANLN, DLGAP5, NDC80, NUSAP1, RACGAP1, PBK, ZWINT, BUB1B, TOP2A, NUF2, CCNB1, RRM2, DTL, KIF20A, CDKN3, HMMR, PRC1, CCL20, NPY1R, CXL12

Table 3. Formation of metadata by listing hub genes from existing studies.

Key candidate genes identification. Eight central hub genes were identified from five methods (degree of connectivity, MNC, MCC, closeness, and betweenness), 6 hub module genes from potential hub modules, and 52 significant meta-hub genes from meta-hub genes. Six overlapping genes were identified using the Venn diagram from these three gene identification methods, which is presented in Fig. 7. These six genes (TOP2A, CDC20, ASPM, PRC1, UBE2C, and NUSAP1) were considered as key genes, which can be easily classified into the subjects as HCC and healthy.

Validation of key candidate genes. *Discriminative power analysis using ROC curve.* Six key or core genes (TOP2A, CDC20, ASPM, PRC1, UBE2C, and NUSAP1) were validated using AUC, computed from ROC curves. We compared the performance of two independent test datasets (GSE76427 and TCGA-LIHC) with one of our train datasets (GSE57957) in order to show the precision of the selected key candidate genes. The ROC curves of six key genes as well as their heatmap for both training and independent test datasets were illustrated in Fig. 8.

The ROC curve of six key candidate genes with their AUC values for the training dataset (GSE57957) was displayed in Fig. 8a: TOP2A (AUC: 0.936, 95% CI 0.871–1.000), CDC20 (AUC: 0.917, 95% CI 0.838–0.996), ASPM (AUC: 0.919, 95% CI 0.851–0.987), PRC1 (AUC: 0.938, 95% CI 0.871–1.000), UBE2C (AUC: 0.803, 95% CI 0.703–0.904), and NUSAP1 (AUC: 0.930, 95% CI 0.895–1.000). As displayed in Fig. 8c, the AUC values of six key or core genes were more than almost 0.780. The AUC values of six key or core genes for the GSE76427 dataset were: TOP2A (AUC: 0.900, 95% CI 0.851–0.949), CDC20 (AUC: 0.887, 95% CI 0.883–0.941), ASPM (AUC: 0.893, 95% CI 0.844–0.942), PRC1 (AUC: 0.931, 95% CI 0.889–0.975), UBE2C (AUC: 0.792, 95% CI 0.723–0.863), and NUSAP1 (AUC: 0.881, 95% CI 0.831–0.933).

Similarly, the ROC curves of six key candidate genes with their AUC values for the TCGA-LIHC-independent test dataset were presented in Fig. 8e. As presented in Fig. 8e, it was observed that six key candidate genes were provided the AUC values of more than 0.900 and their individual AUC values were as follows: TOP2A (AUC: 0.961, 95% CI 0.939–0.984), CDC20 (AUC: 0.968, 95% CI 0.949–0.986), ASPM (AUC: 0.960, 95% CI 0.938–0.983), PRC1 (AUC: 0.967, 95% CI 0.948–0.987), UBE2C (AUC: 0.965, 95% CI 0.946–0.985), and NUSAP1 (AUC: 0.919, 95% CI 0.889–0.949). Therefore, these six key genes (TOP2A, CDC20, ASPM, PRC1, UBE2C, and NUSAP1) showed strong discriminative power to classify HCC patients from healthy controls. These validations would be supported our findings and provided them more robust.

Survival analysis. In this work, we adopted survival analysis of six key candidate genes (TOP2A, CDC20, ASPM, PRC1, NUSAP1, and UBE2C) using univariate Cox regression in R and its results are presented in Fig. 9. As shown in Fig. 9, we observed that our identified six key candidate genes for HCC patinets such as TOP2A, CDC20, ASPM, PRC1, NUSAP1, and UBE2C were strongly associated with the survival status of HCC patients ($p < 0.05$). So, the over-expression levels of TOP2A, CDC20, ASPM, PRC1, NUSAP1, and UBE2C had poor survival periods compared to lower expression levels of that key candidate genes.

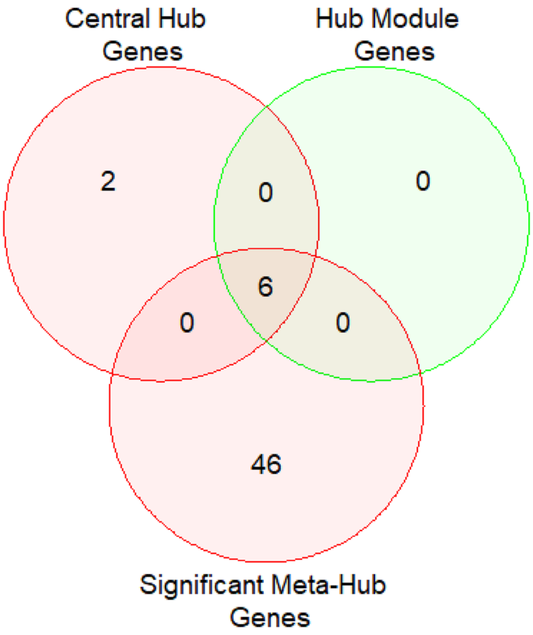

Figure 7. Identification of key candidate genes of HCC from central hub genes, hub module genes, and significant meta-hub genes.

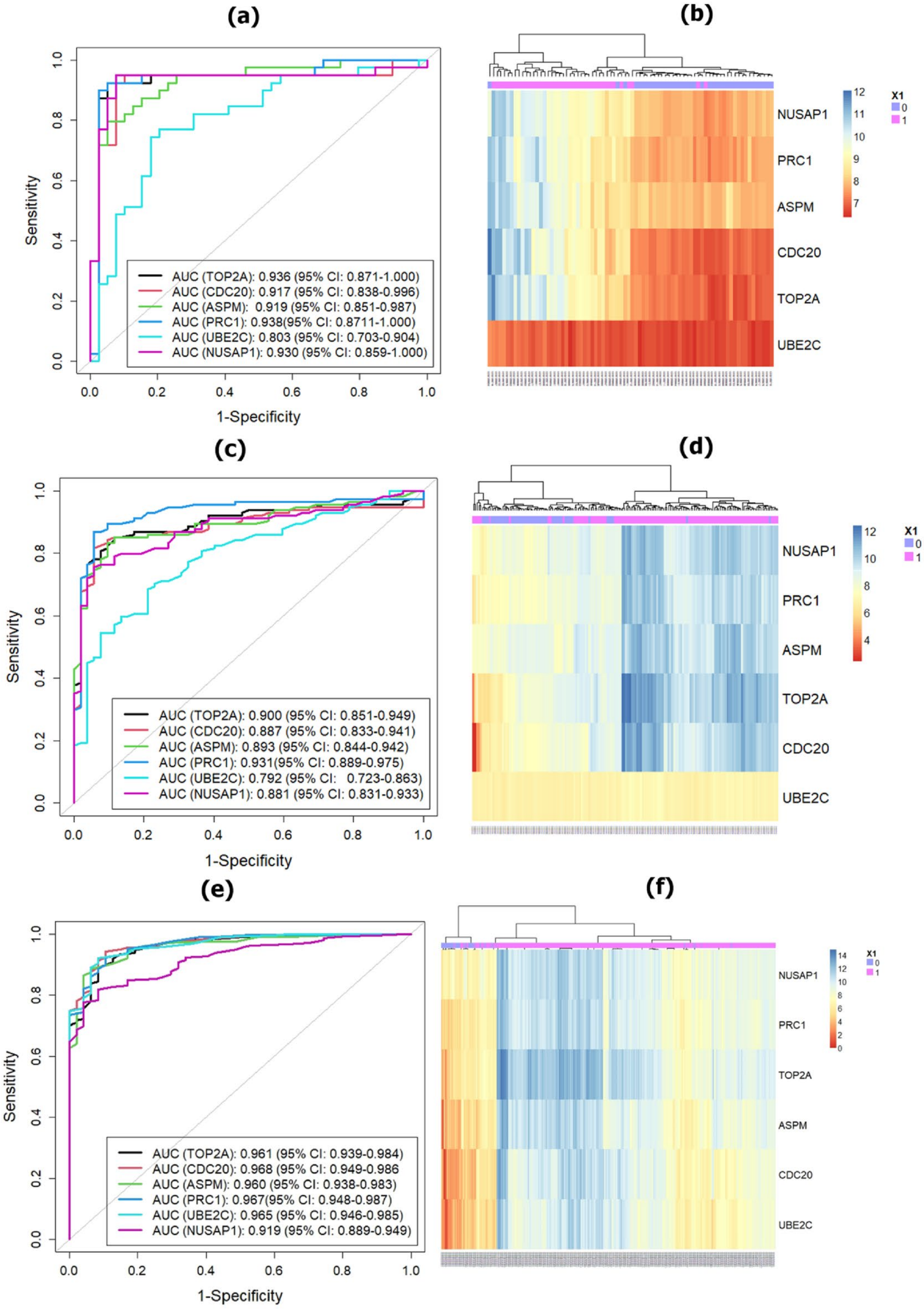

Figure 8. Validation of the six key candidate genes using AUC and heatmap: (**a**), (**b**) GSE57957-based training dataset; (**c**), (**d**) GSE76427-based independent test dataset; and (**e**), (**f**) TCGA-LIHC based independent test dataset. Whereas, ROC curves were generated using pROC version 1.18.0 package[121] and heatmap was generated using "NMF" version 0.24.0 package in R[111].

Discussion

In this work, we assessed three datasets, namely GSE36376, GSE39791, and GSE57957, to detect the DEGs for HCC patients. We determined 699, 428, and 413 DEGs using "limma" from the GSE36376, GSE39791, and

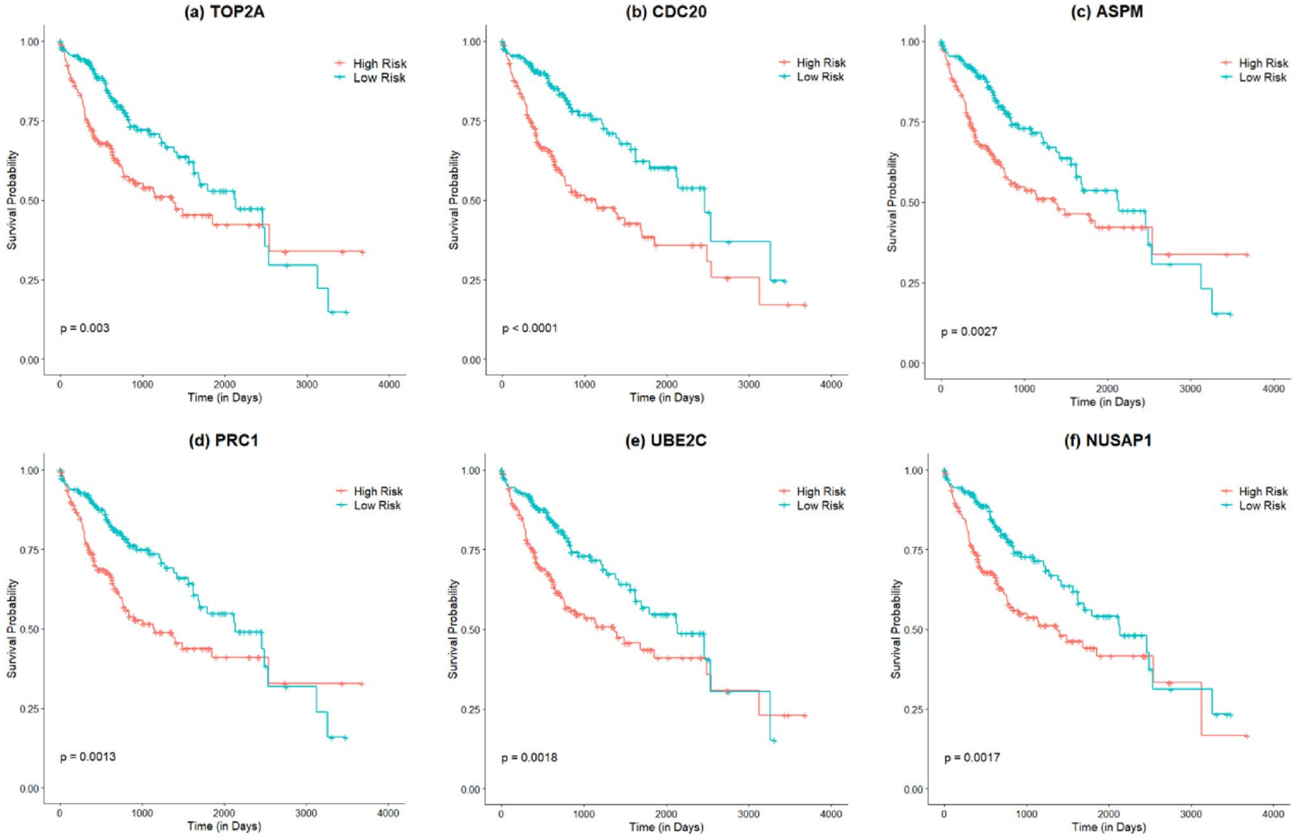

Figure 9. Survival analysis of six key candidate genes for HCC: (**a**) TOP2A; (**b**) CDC20; (**c**) ASPM; (**d**) PRC1; (**e**) NUSAP1; and (**f**) UBE2C. The horizontal axis (x-axis) represents the time to event (in days) and the vertical axis (y-axis) represents survival probability. The HCC patients were divided into two groups: high-risk and low-risk and assigned a color. The red line designates the samples with high risk, and the green line represents the samples with low risk. $p < 0.05$ indicates a statistically significant difference in mortality between groups. The survival plots were generated using the "Survfit" package in R[122].

GSE57957 datasets, which were illustrated in Fig. 2. Moreover, we implemented SVM to determine the DEDGs from individual datasets (see in Fig. 3) and selected overlapping or shared 75 DEDGs among the identified DEDGS from GSE36376, GSE39791, and GSE57957 datasets, which were clearly shown in Fig. 4. At the same time, enrichment analysis was executed on overlapping or shared DEDGs to clear understand their better exploration and molecular mechanism (see in Table 1). We found that the potential BP functional categories were strongly related to the development and progression of HCC patients. Retinol and retinoid metabolic processes have been linked to a variety of liver diseases, including fatty liver disease, which leads to HCC[68,69]. The rest of the BP categories were also enriched with common DEDGs, which also coincided with existing studies, like cellular response to cadmium ion[42,57,70], cellular response to copper ion[36,70], and steroid catabolic process[42].

The top 5 GO terms were significantly enriched with common DEDGS, which were also consistent with previous results, such as extra cellular region[35,37,38,57], extracellular exosome[37,38], extracellular space[37,38,57], high-density lipoprotein particle[57], and apical plasma membrane[53]. In the case of MFs, common DEDGs were also enriched with top five GO terms. Existing studies supported these enrichment factional categories, including retinol dehydrogenase activity[14], and oxidoreductase activity[37,38,42]. We also analyzed KEGG pathways and chose five pathways that were closely related to our overlapping DEDGs (see in Table 2). Different existing studies supported our findings, such as retinol metabolism[35,37,38,40,43,70], metabolic pathways[37,38], tryptophan metabolism[38,42,70], steroid hormone biosynthesis[42,70], and drug metabolism-cytochrome P450[35,42,70].

A PPI network was built with shared DEDGs using Cystoscape (see in Fig. 5a and then eight central hub genes (NUSAP1, TOP2A, CDC20, PRC1, UBE2C, ASPM, PNPLA7, and MT1E) were identified from five hub gene selection methods, which were presented in Fig. 5b. The potential modules were identified using MCODE scores and module 1 was identified due to having the highest MCODE scores. We selected six hub module genes from module 1 as well as constructed their PPI network (see in Fig. 6). In addition, we examined 52 papers and took the hub genes from earlier studies[8–58] in order to make metadata. At the same time, we listed 214 meta-hub genes by taking the union of extracted hub genes, which were presented in Table 3. We selected 52 significant meta-hub genes from the list of meta-hub genes whose frequency was greater than 3. Finally, we identified the six shared genes (TOP2A,CDC20, ASPM, PRC1, UBE2C, and NUSAP1) by intersecting central hub genes, hub module genes, and significant meta-hub genes, extracted from the earlier studies, known as key relevant or candidate genes, which were clearly depicted in Fig. 7. We validated these key relevant or candidate genes using

AUC for one training and two independent test datasets (see Fig. 8). We observed that these six key relevant or candidate genes had high discriminative power for the differentiation of HCC patients.

TOP2A is a cell cycle-related gene that encoded a DNA topoisomerase which controls and alters the topologic states of DNA during transcription. TOP2A overexpression has been identified as a core or potential biomarker for ovarian cancers[71], glioma[72], and lung cancers[73]. A study showed that TOP2A overexpression in HCC patients was significantly correlated with progression and poor prognosis[74,75]. In the case of our study, TOP2A was also considered as a key or core gene for the progression and development of HCC. This finding was coincided with previous studies[12,14,15,18,20–25,27,32–36,39,41–43,45,46,48,50,51,54–58,61].

CDC20 is a vital regulator of cell division in humans[76,77]. Overexpression or high expression of CDC20 has also been linked to lung cancer[78], colorectal cancer[79], breast cancer[80,81], and other cancers. Moreover, CDC20 was strongly correlated with poor prognosis in gastric cancer[82], bladder cancer[83], and breast cancer[84]. A study revealed that CDC20 over-expression was significantly associated with HCC[85]. Another recent study demonstrated that there existed a strong relationship between CDC20 overexpression and the prognosis of HCC[86]. Our findings also showed that CDC20 was a potential key biomarker that played an crucial or essential role for the development and progression of HCC. Different existing studies also supported our findings[10,13,14,17,18,20,24,30,32,37,40,41,44,48,50,55–57].

ASPM is a protein that have a major influence in the development of HCC. ASPM is located on chromosome 1 and band 1q31 and consists of 28 exons and 3477 amino-acid proteins[87]. Lots of studies have identified ASPM as a hub gene or key biomarker for multiple cancers[88–90]. Zhang et al.[90] reported that ASPM can be a promising therapeutic target for liver. Moreover, ASPM overexpression was strongly correlated with bladder cancer and consiered as promising predictor[91]. Our findings also illustrated that ASPM was a novel key biomarker for HCC, which was supported by the existing studies[9,22,35,38,39,41–43,45–48,58].

PRC1 is an essential protein that is the regulator of cytokinesis[92]. The higher expression level of PRC1 was found among HCC patients than healthy controls. The overexpression of PRC1 was associated with a poor prognosis for HCC patients[93]. Our work also indicated that PRC1 was a promising or key biomarker for the development of HCC, which coincided with previous studies[15,22,25,33,35,39,42,43,45,46,56,57,61].

Similarly, we proposed UBE2C as a key or core predictor for development of HCC, which was supported by various existing studies[10,18,33,36,41,44,58]. Xiong et al.[94] suggested UBE2C as a potential biomarker or gene for HCC. High expression of UBE2C was also found in HCC than healthy subjects[95]. UBE2C is not play a crucial role HCC but also in variety of cancers: lung cancer, gastric cancer[96,97].

NUSAP1 is a protein associated with the nucleolar-spindle that have a vital role in spindle microtubule organization[98]. overexpression of NUSAP1 was found in a variety of malignancies, including HCC[58,99], colon cancer[100,101], prostate cancer[102,103], and cervical carcinoma[104]. Moreover, overexpression of NUSAP1 was strongly linked with poor prognosis of prostate cancer[103] and colon cancer[101]. Another study revealed that NUSAP1 is related to HCC[105]. Roy et al.[105] illustrated that NUSAP1 expression might rise in HCC samples with low expression levels of miRNA 193a-5p, and that this overexpression was strongly associated with a shorter patient survival time. Our findings also illustrated that NUSAP1 was one of the key candidate genes that the highest expression levels were found in HCC subjects compared to healthy subjects. These findings were consistent with existing studies[15,22,46,56,58,61].

Moreover, two independent test datasets were also used to validate these six key candidate genes using AUC. A survival analysis was also performed of these six candidate genes for HCC patients. In both cases, our identified six key candidate genes (TOP2A, CDC20, ASPM, PRC1, UBE2C, and NUSAP1) showed significant association with the development and progression of HCC. This finding will provide evidence and new insight to physicians and readers in determining the diagnosis of HCC as well as the correlated pathway of HCC.

Materials and methods

Data acquisition and preprocessing.

In this work, three publicly available microarray gene expression datasets with GEO accession: GSE36376[66], GSE39791[106], and GSE57957[107] with GPL10558 [Illumina HumanHT-12 V4.0 expression bead chip] were used to determine the key candidate genes. Another two independent test datasets were used to validate key candidate genes. One independent dataset was taken from the GEO database with accession number: GSE76427 with GPL10558 platform[102] and another independent test dataset was taken from the Cancer Genome Atlas (TCGA) database. Microarray gene expression datasets were downloaded from the GEO database (www.ncbi.nlm.nih.gov/geo/) and TCGA-liver hepatocellular carcinoma (TCGA-LIHC) dataset was downloaded from the TCGA database (https://portal.gdc.cancer.gov/). The datasets underwent a log2 transformation and quintile normalization. Although these datasets were taken from the publicly available GEO repository, being human data, all methods were performed in accordance with the relevant guidelines and regulations. Table 4 presents a summary of the utilized datasets.

Identification of DEGs from each dataset.

To identify the DEGs between HCC and healthy controls, each of the selected datasets was analyzed using the "limma" package[108] in R-software with version 4.1.2. We computed the $|log_2FC|$ and adj. p-value of each gene from the selected dataset. "Bioconductor annotation"[109] package was used to convert microarray data probes into gene symbols. If multiple probes were matched with a gene symbol, take the gene with their associated expression values that provided the lowest or minimum adjusted p-value. The DEGs between HCC and healthy controls were identified with a cutoff of point: $|log_2FC| > 1$ and $adj.p - value < 0.01$ (false discovery rate). The volcano plot of DEGs was generated using the "ggplot2 version 3.3.6" package in R[110]. Moreover, a heat map of the expression of DEGs was generated with the "NMF" version 0.24.0 package in R[111].

Datasets	Platform	Total samples	HCC	Control
GSE36376[66]	GPL10558	433	240	193
GSE39791[106]	GPL10558	144	72	72
GSE57957[107]	GPL10558	78	39	39
GSE76427[102]	GPL10558	167	115	52
TCGA-LIHC	–	424	374	50

Table 4. Summary of utilized HCC datasets.

SVM-based identification of DEDGs from DEGs for each dataset. The main purpose of SVM is to identify a hyperplane in a high dimensional space[112,113] that can easily discriminate HCC patients from healthy control patients using the following discriminate function:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x_j) + b \tag{1}$$

where, b is the bias term.

In this study, we have used radial basis kernel, which is defined as follows:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{2}$$

We set the different values of cost (C) and gamma (γ) and tuned these values using a grid search method and select the optimal value of C and (γ) to improve classification accuracy. In this current study, we adopted SVM as a gene selection method, and its identification procedure is described as follows:

Step 1 Select one gene from a list of identified DEGs.
Step 2 Trained SVM-based model with five-fold cross-validation (CV) protocols.
Step 3 Calculate the classification accuracy for this selected gene.
Step 4 Repeat Step 2 to Step 3 for all identified DEGs.
Step 5 Sort the classification accuracy of all DEGs in descending order of magnitude.
Step 6 Choose the genes that will produce a classification accuracy of more than 80.0.

Identification of common DEDGs. After selecting differentially expressed discrimination genes (DEDGs) using SVM, we identified the shared or overlapping or common DEDGs among three datasets using the following formula:

$$\text{Common DEDGs} = \bigcap_{i=1}^{r} \text{Identified DEDGs from GEO Datasets}_i \tag{3}$$

where, r is the number of utilizing GEO dataset (here, r = 3).

Enrichment analysis of common DEDGs. To better understand the mechanism and progression of HCC patients, we obtained enrichment analysis, including GO and KEEG analysis[114,115] on DEDGs using DAVID version 6.8 tools[116] (david.ncifcrf.gov). A p-value < 0.05 was considered for significant.

PPI network analysis and central hub gene identification. The STRING version 11.5 software (www.string-db.org) was utilized to obtain the potential interactions among common DEDGs[117]. A protein-protein interaction (PPI) with a confidence score of $>$ 0.70 and a maximum number of interactors of 0 was preserved and loaded into Cystoscape version 3.9.1[118] to build a PPI network. The degree of connectivity, maximum neighborhood component (MNC), maximal clique centrality (MCC), centralities of closeness, and betweenness were computed using cytoHubba[119]. Then, we sorted the values of degree of connectivity, MNC, MCC, centralities of closeness, and betweenness in descending order of magnitude and chose the top 30 DEDGs, known as hub genes. The central hub genes were selected by overlapping hub genes, which were computed from the degree of connectivity, MNC, MCC, centralities of closeness, and betweenness. Mathematically, it is defined as follows:

$$\text{Central Hub Genes} = \bigcap_{i=1}^{hg} \text{Hub Genes from Identification Methods}_i \tag{4}$$

where, hg is the number of hub gene identification methods (Here, hg=5).

Hub modules and its associated genes identification. MCODE was used to determine the most closely connected modules from the PPI network[120]. We analyzed the modules with the following cutoff points: degree =2, cluster finding =haircut, nodes score =0.2, K-score =2, and max depth =100, respectively. We determined the potential modules that provided the MCODE with scores of $\geq$ 6 and the number of nodes of $\geq$ 6. Then, the hub module genes were identified using the following formula:

$$\text{Hub Module Genes} = \bigcup_{i=1}^{h_m} \text{Genes from Module}_i \tag{5}$$

where, h_m is the number of significant modules.

Significant meta-hub genes identification from metadata. We reviewed some existing studies related to HCC-based gene identification. To make metadata, we listed their identified hub genes for HCC, called "meta-hub genes," which can be written as follows:

$$\text{Meta-Hub Genes} = \bigcup_{i=1}^{m} \text{Hub Genes from Previous Study}_i \tag{6}$$

where, m is the number of studies obtained from obtaining hub genes (here, m = 52).

We also counted the frequency of each meta-hub gene depending on how many studies identified that gene as a hub gene. Finally, we identified significant meta-hub genes from meta-hub genes whose frequency was greater than or equal to 3, which can be written as follows:

$$\text{Significant Meta-Hub Genes} = \{g_i\};\ i = 1, 2, \ldots, n \tag{7}$$

where, $g_i \in$ meta-hub gene and n is the number of meta-hub genes whose frequency is $\geq$ 3

Key candidate genes identification. To identify the key candidate genes, we selected the central hub genes from the PPI network, hub module genes from significant modules, and significant meta-hub genes from existing studies. Therefore, we identified the key candidate genes for HCC using the following formula:

$$\text{Key Candidate Genes} = \bigcap_{i=1}^{k} \text{Important Genes from Identification Methods}_i \tag{8}$$

where, k is the number of significant gene identification methods (Here, k = 3). In this work, central hub genes, hub module genes, and significant gene selection methods will be considered "Important Gene Identification Methods".

Validation of key candidate genes. *Discriminative power analysis using ROC curve.* In this work, we used two independent test datasets in order to validate the key candidate genes. One independent test dataset (GSE76427) was taken from the GEO database, and another independent dataset was taken from the TCGA database. The description of these independent test datasets is more clearly explained in Table 4. We validated the selected key candidate genes using the area under the curve (AUC), computed from the receiver operating characteristic curve (ROC). In ROC analysis, first, we selected one gene and class label, and then we adopted logistic regression with the leave-one-out CV protocol. We computed AUC values using the "pROC" R-package[121]. Moreover, we also compared the performances of independent test datasets with one of our training datasets (GSE57957) in order to show the precision of the selected key candidate genes.

Survival analysis. In this work, we used TCGA-LIHC dataset for survival analysis in order to show prognostic status of key candidate genes. We classified HCC patients into high-risk and low-risk groups on the basis of median expression level of each key candidate gene. We performed survival analysis of our identified key candidate genes using the "Survfit" package in R language[122]. A p-value < 0.05 was considered statistically significant ("Supplementary information").

Data availability

The datasets generated and/or analyzed during the current study are available in the Gene Expression Omnibus (GEO) repository with accession numbers: GSE36376, GSE39791, GSE57957, and GSE76427 with GPL10558 platforms. One can easily download these datasets from the link: www.ncbi.nlm.nih.gov/geo/.

Received: 9 September 2022; Accepted: 2 March 2023
Published online: 07 March 2023

References

1. Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P. Global cancer statistics, 2002. *CA Cancer J. Clin.* **55**, 74–108. https://doi.org/10.3322/canjclin.55.2.74 (2005).
2. Yang, J. D. *et al.* A global view of hepatocellular carcinoma: Trends, risk, prevention and management. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 589–604. https://doi.org/10.1038/s41575-019-0186-y (2019).
3. Kumar, V., Abbas, A. K., Fausto, N. & Aster, J. C. *Robbins and Cotran Pathologic Basis of Disease* 9th edn. (Elsevier Health Sciences, 2015).
4. Huang, T. *et al.* The role of hepatitis c virus in the dynamic protein interaction networks of hepatocellular cirrhosis and carcinoma. *Int. J. Comput. Biol. Drug Design* **4**, 5–18. https://doi.org/10.1504/IJCBDD.2011.038654 (2011).

5. Yuan, W. *et al.* Comparative analysis of viral protein interaction networks in hepatitis b virus and hepatitis c virus infected hcc. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1844**, 271–279. https://doi.org/10.1016/j.bbapap.2013.06.002 (2014).
6. Llovet, J. M. *et al.* Hepatocellular carcinoma. *Nat. Rev. Dis. Primers* **7**, 6–34. https://doi.org/10.1038/s41572-020-00240-3 (2021).
7. Akinyemiju, T. *et al.* The burden of primary liver cancer and underlying etiologies from 1990 to 2015 at the global, regional, and national level: results from the global burden of disease study 2015. *JAMA Oncol.* **3**, 1683–1691. https://doi.org/10.1001/jamaoncol.2017.3055 (2017).
8. Zhang, C. *et al.* The identification of key genes and pathways in hepatocellular carcinoma by bioinformatics analysis of high-throughput data. *Med. Oncol.* **34**, 1–13. https://doi.org/10.1007/s12032-017-0963-9 (2017).
9. Maddah, R. *et al.* Identification of critical genes and pathways associated with hepatocellular carcinoma and type 2 diabetes mellitus using integrated bioinformatics analysis. *Inform. Med. Unlocked* **30**, 100956–100963. https://doi.org/10.1016/j.imu.2022.100956 (2022).
10. Yan, G. & Liu, Z. Identification of differentially expressed genes in hepatocellular carcinoma by integrated bioinformatic analysis. *bioRxiv.*https://doi.org/10.1101/570846 *(2019).*
11. Qian, Z., Yan, Z. & Zhengkui, L. Mining of gene modules and identification of key genes in hepatocellular carcinoma based on gene co-expression network analysis. in *Proceedings of the 2020 12th International Conference on Bioinformatics and Biomedical Technology*, 18–24. https://doi.org/10.1145/3405758.3405762 (2020).
12. Zhao, Y. & Xie, Y. Study on differential expression genes in hcc based on geo database. in *Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing*, 63–69. https://doi.org/10.1145/3448748.3448759 (2021).
13. Liu, J. *et al.* Identification of multiple hub genes and pathways in hepatocellular carcinoma: A bioinformatics analysis. *BioMed Res. Int.* **2021**, 1–11. https://doi.org/10.1155/2021/8849415 (2021).
14. Meng, Z. *et al.* Identification of potential hub genes associated with the pathogenesis and prognosis of hepatocellular carcinoma via integrated bioinformatics analysis. *J. Int. Med. Res.* **48**, 1–23. https://doi.org/10.1177/0300060520910019 (2020).
15. Rosli, A. F. C., Razak, S. R. A. & Zulkifle, N. Bioinformatics analysis of differentially expressed genes in liver cancer for identification of key genes and pathways. *Malaysian J. Med. Health Sci.* **15**, 18–24 (2019).
16. Li, Y. *et al.* Integrated bioinformatics analysis reveals key candidate genes and pathways associated with clinical outcome in hepatocellular carcinoma. *Front. Genet.* **11**, 814–819. https://doi.org/10.3389/fgene.2020.00814 (2020).
17. Li, Z., Lin, Y., Cheng, B., Zhang, Q. & Cai, Y. Identification and analysis of potential key genes associated with hepatocellular carcinoma based on integrated bioinformatics methods. *Front. Genet.* **12**, 571231–571245. https://doi.org/10.3389/fgene.2021.571231 (2021).
18. Tian, D., Yu, Y., Zhang, L., Sun, J. & Jiang, W. A five-gene-based prognostic signature for hepatocellular carcinoma. *Front. Med.* **8**, 1–24. https://doi.org/10.3389/fmed.2021.681388 (2021).
19. Wan, Z., Zhang, X., Luo, Y. & Zhao, B. Identification of hepatocellular carcinoma-related potential genes and pathways through bioinformatic-based analyses. *Genet. Testing Mole. Biomarkers* **23**, 766–777. https://doi.org/10.1089/gtmb.2019.0063 (2019).
20. Zhu, Q., Sun, Y., Zhou, Q., He, Q. & Qian, H. Identification of key genes and pathways by bioinformatics analysis with tcga rna sequencing data in hepatocellular carcinoma. *Mol. Clin. Oncol.* **9**, 597–606. https://doi.org/10.3892/mco.2018.1728 (2018).
21. Wang, J., Tian, Y., Chen, H., Li, H. & Zheng, S. Key signaling pathways, genes and transcription factors associated with hepatocellular carcinoma. *Mol. Med. Rep.* **17**, 8153–8160. https://doi.org/10.3892/mmr.2018.8871 (2018).
22. Zhou, L., Du, Y., Kong, L., Zhang, X. & Chen, Q. Identification of molecular target genes and key pathways in hepatocellular carcinoma by bioinformatics analysis. *OncoTargets Therapy* **11**, 1861. https://doi.org/10.2147/OTT.S156737 (2018).
23. Zhang, P. *et al.* Bioinformatics analysis of candidate genes and pathways related to hepatocellular carcinoma in china: A study based on public databases. *Pathol. Oncol. Res.* **27**, 588532–588546. https://doi.org/10.3389/pore.2021.588532 (2021).
24. Mou, T. *et al.* Identification and interaction analysis of key genes and micrornas in hepatocellular carcinoma by bioinformatics analysis. *World J. Surg. Oncol.* **15**, 1–9. https://doi.org/10.1186/s12957-017-1127-2 (2017).
25. Wu, M. *et al.* Analysis of potential key genes in very early hepatocellular carcinoma. *World J. Surg. Oncol.* **17**, 1–8. https://doi.org/10.1186/s12957-019-1616-6 (2019).
26. Gui, T., Dong, X., Li, R., Li, Y. & Wang, Z. Identification of hepatocellular carcinoma-related genes with a machine learning and network analysis. *J. Comput. Biol.* **22**, 63–71. https://doi.org/10.1089/cmb.2014.0122 (2015).
27. Wang, J. *et al.* Identification and validation of key genes in hepatocellular carcinoma by bioinformatics analysis. *Biomed Res. Int.* **2021**, 6662114–6662127. https://doi.org/10.1155/2021/6662114 (2021).
28. Lu, H. & Zhu, Q. Identification of key biological processes, pathways, networks, and genes with potential prognostic values in hepatocellular carcinoma using a bioinformatics approach. *Cancer Biother. Radiopharm.* **36**, 837–849. https://doi.org/10.1089/cbr.2019.3327 (2021).
29. Bhatt, S. *et al.* Deciphering key genes and mirnas associated with hepatocellular carcinoma via network-based approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **36**, 837–849. https://doi.org/10.1109/TCBB.2020.3016781 (2020).
30. Zhang, Y. *et al.* A gene module identification algorithm and its applications to identify gene modules and key genes of hepatocellular carcinoma. *Sci. Rep.* **11**, 1–14. https://doi.org/10.1038/s41598-021-84837-y (2021).
31. Jiang, X. & Hao, Y. Analysis of expression profile data identifies key genes and pathways in hepatocellular carcinoma. *Oncol. Lett.* **15**, 2625–2630. https://doi.org/10.3892/ol.2017.7534 (2018).
32. Zhang, X. *et al.* Identification of hub genes associated with hepatocellular carcinoma prognosis by bioinformatics analysis. *J. Cancer Therapy* **12**, 186–207. https://doi.org/10.4236/jct.2021.124019 (2021).
33. Wu, M., Liu, Z., Zhang, A. & Li, N. Identification of key genes and pathways in hepatocellular carcinoma: A preliminary bioinformatics analysis. *Medicine* **98**, 1–7. https://doi.org/10.1097/MD.0000000000014287 (2019).
34. Nguyen, T. B., Do, D. N., Nguyen-Thanh, T., Tatipamula, V. B. & Nguyen, H. T. Identification of five hub genes as key prognostic biomarkers in liver cancer via integrated bioinformatics analysis. *Biology* **10**, 957–970. https://doi.org/10.3390/biology10100957 (2021).
35. Zhou, Z. *et al.* Screening hub genes as prognostic biomarkers of hepatocellular carcinoma by bioinformatics analysis. *Cell Transplant.* **28**, 76S-86S. https://doi.org/10.1177/0963689719893950 (2019).
36. Yu, C., Chen, F., Jiang, J., Zhang, H. & Zhou, M. Screening key genes and signaling pathways in colorectal cancer by integrated bioinformatics analysis. *Mol. Med. Rep.* **20**, 1259–1269. https://doi.org/10.3892/mmr.2019.10336 (2019).
37. Kakar, M. *et al.* Identification of novel potential biomarkers in hepatocarcinoma cancer; a transcriptome analysis. *Preprint (Version 3) available at Research Square (02 March 2021)* 1–21. https://doi.org/10.21203/rs.3.rs-154350/v2 (2021).
38. Ji, Y., Yin, Y. & Zhang, W. Integrated bioinformatic analysis identifies networks and promising biomarkers for hepatitis b virus-related hepatocellular carcinoma. *Int. J. Genom.* **2020**, 1–18. https://doi.org/10.1155/2020/2061024 (2020).
39. Chen, D. *et al.* Bioinformatic evidence reveals that cell cycle correlated genes drive the communication between tumor cells and the tumor microenvironment and impact the outcomes of hepatocellular carcinoma. *BioMed Res. Int.* **2021**, 4092635–4092660. https://doi.org/10.1155/2021/4092635 (2021).
40. Qiang, R. *et al.* Identification of 5 hub genes related to the early diagnosis, tumour stage, and poor outcomes of hepatitis b virus-related hepatocellular carcinoma by bioinformatics analysis. *Comput. Math. Methods Med.* **2021**, 1–20. https://doi.org/10.1155/2021/9991255 (2021).

41. Wang, J. *et al.* Global analysis of gene expression signature and diagnostic/prognostic biomarker identification of hepatocellular carcinoma. *Sci. Progress* **104**, 1–7. https://doi.org/10.1177/00368504211029429 (2021).
42. Zhang, Y., Tang, Y., Guo, C. & Li, G. Integrative analysis identifies key mrna biomarkers for diagnosis, prognosis, and therapeutic targets of hcv-associated hepatocellular carcinoma. *Aging (Albany NY)* **13**, 12865–12895. https://doi.org/10.18632/aging.202957 (2021).
43. Kim, S.-H. *et al.* Identification of key genes and carcinogenic pathways in hepatitis b virus-associated hepatocellular carcinoma through bioinformatics analysis. *Ann. Hepato-biliary-pancreatic Surg.* **26**, 58–68. https://doi.org/10.14701/ahbps.21-108 (2022).
44. Zhang, G., Kang, Z., Mei, H., Huang, Z. & Li, H. Promising diagnostic and prognostic value of six genes in human hepatocellular carcinoma. *Am. J. Transl. Res.* **12**, 1239–1254 (2020).
45. Sha, M. *et al.* Identification of genes predicting unfavorable prognosis in hepatitis b virus-associated hepatocellular carcinoma. *Ann. Transl. Med.* **9**, 975–985. https://doi.org/10.21037/atm-21-2085 (2021).
46. Chen, H. *et al.* Identification of hub genes associated with immune infiltration and predict prognosis in hepatocellular carcinoma via bioinformatics approaches. *Front. Genet.* **11**, 575762–575779. https://doi.org/10.3389/fgene.2020.575762 (2021).
47. He, B. *et al.* Bioinformatics analysis of key genes and pathways for hepatocellular carcinoma transformed from cirrhosis. *Medicine.* **96**, 6938–6946. https://doi.org/10.1097/MD.0000000000006938 (2017).
48. Zhang, S., Peng, R., Xin, R., Shen, X. & Zheng, J. Conjoint analysis for hepatic carcinoma with hub genes and multi-slice spiral ct. *Medicine* **99**, e23099–e23110. https://doi.org/10.1097/MD.0000000000023099 (2020).
49. Hu, W. Q. *et al.* Identification of biological targets of therapeutic intervention for hepatocellular carcinoma by integrated bioinformatical analysis. *Med. Sci. Monitor* **24**, 3450–3461. https://doi.org/10.12659/MSM.909290 (2018).
50. Zhang, Q. *et al.* Prediction and analysis of weighted genes in hepatocellular carcinoma using bioinformatics analysis. *Mol. Med. Rep.* **19**, 2479–2488. https://doi.org/10.3892/mmr.2019.9929 (2019).
51. Li, N., Li, L. & Chen, Y. The identification of core gene expression signature in hepatocellular carcinoma. *Oxidative Med. Cell. Longevity* **2018**, 1–15. https://doi.org/10.1155/2018/3478305 (2018).
52. Cao, J., Zhang, R., Zhang, Y. & Wang, Y. Combined screening analysis of aberrantly methylated-differentially expressed genes and pathways in hepatocellular carcinoma. *J. Gastrointestinal Oncol.* **13**, 311–325. https://doi.org/10.21037/jgo-21-866 (2022).
53. Yang, L., Zeng, L.-F., Hong, G.-Q., Luo, Q. & Lai, X. Construction of a novel clinical stage-related gene signature for predicting outcome and immune response in hepatocellular carcinoma. *J. Immunol. Res.* **2022**, 1–10. https://doi.org/10.1155/2022/6535009 (2022).
54. Wang, M., Wang, L., Wu, S., Zhou, D. & Wang, X. Identification of key genes and prognostic value analysis in hepatocellular carcinoma by integrated bioinformatics analysis. *Int. J. Genom.* **2019**, 1–22. https://doi.org/10.1155/2019/3518378 (2019).
55. Jiang, N. *et al.* Identification of core genes related to progression and prognosis of hepatocellular carcinoma and small-molecule drug predication. *Front. Genet.* **12**, 608017–608036. https://doi.org/10.3389/fgene.2021.608017 (2021).
56. Li, L., Lei, Q., Zhang, S., Kong, L. & Qin, B. Screening and identification of key biomarkers in hepatocellular carcinoma: Evidence from bioinformatic analysis. *Oncol. Rep.* **38**, 2607–2618. https://doi.org/10.3892/or.2017.5946 (2017).
57. Xing, T., Yan, T. & Zhou, Q. Identification of key candidate genes and pathways in hepatocellular carcinoma by integrated bioinformatical analysis. *Exp. Therap. Med.* **15**, 4932–4942. https://doi.org/10.3892/etm.2018.6075 (2018).
58. Zhu, W., Xu, J., Chen, Z. & Jiang, J. Analyzing roles of nusap1 from clinical, molecular mechanism and immune perspectives in hepatocellular carcinoma. *Front. Genet.* **12**, 689159–689181. https://doi.org/10.3389/fgene.2021.689159 (2021).
59. Jiang, M. *et al.* Identification of hepatocellular carcinoma related genes with k-th shortest paths in a protein-protein interaction network. *Mol. BioSyst.* **9**, 2720–2728. https://doi.org/10.1039/C3MB70089E (2013).
60. Huang, T., Wang, J., Cai, Y.-D., Yu, H. & Chou, K.-C. Hepatitis c virus network based classification of hepatocellular cirrhosis and carcinoma. *PloS One* **7**, e34460. https://doi.org/10.1371/journal.pone.0034460 (2012).
61. Dai, Q. *et al.* Six genes involved in prognosis of hepatocellular carcinoma identified by cox hazard regression. *BMC Bioinform.* **22**, 1–12. https://doi.org/10.1186/s12859-021-04095-7 (2021).
62. Qing, J.-B., Song, W.-Z., Li, C.-Q. & Li, Y.-F. The diagnostic and predictive significance of immune-related genes and immune characteristics in the occurrence and progression of iga nephropathy. *J. Immunol. Res.* **2022**, 1–20. https://doi.org/10.1155/2022/9284204 (2022).
63. Yu, S.-H. *et al.* Lasso and bioinformatics analysis in the identification of key genes for prognostic genes of gynecologic cancer. *J. Pers. Med.* **11**, 1177. https://doi.org/10.3390/jpm11111177 (2021).
64. Basith, S., Hasan, M. M., Lee, G., Wei, L. & Manavalan, B. Integrative machine learning framework for the identification of cell-specific enhancers from the human genome. *Brief. Bioinform.* **22**, bbab252. https://doi.org/10.1093/bib/bbab252 (2021).
65. Hasan, Al Mehedi, Maniruzzaman, M. & Shin, J. Identification of key candidate genes for iga nephropathy using machine learning and statistics based bioinformatics models. *Sci. Rep.* **12**, 1–14. https://doi.org/10.1038/s41598-022-18273-x (2022).
66. Lim, H.-Y. *et al.* Prediction of disease-free survival in hepatocellular carcinoma by gene expression profiling. *Ann. Surg. Oncol.* **20**, 3747–3753. https://doi.org/10.1245/s10434-013-3070-y (2013).
67. Zeng, L. *et al.* Differential combinatorial regulatory network analysis related to venous metastasis of hepatocellular carcinoma. *BMC Genom.* **13**, 1–14. https://doi.org/10.1186/1471-2164-13-s8-s14 (2012).
68. Shirakami, Y., Sakai, H. & Shimizu, M. Retinoid roles in blocking hepatocellular carcinoma. *Hepatobiliary Surg. Nutr.* **4**, 222–228. https://doi.org/10.3978/j.issn.2304-3881.2015.05.01 (2015).
69. Pettinelli, P. *et al.* Altered hepatic genes related to retinol metabolism and plasma retinol in patients with non-alcoholic fatty liver disease. *PLoS One* **13**, e0205747–e0205763. https://doi.org/10.1371/journal.pone.0205747 (2018).
70. Lai, X. *et al.* A novel gene signature based on cdc20 and fcn3 for prediction of prognosis and immune features in patients with hepatocellular carcinoma. *J. Immunol. Res.* **2022**, 1–22. https://doi.org/10.1155/2022/9117205 (2022).
71. Gao, Y. *et al.* Top2a promotes tumorigenesis of high-grade serous ovarian cancer by regulating the tgf-β/smad pathway. *J. Cancer* **11**, 4181–4192. https://doi.org/10.7150/jca.42736 (2020).
72. Zhou, T., Wang, Y., Qian, D., Liang, Q. & Wang, B. Over-expression of top2a as a prognostic biomarker in patients with glioma. *Int. J. Clin. Exp. Pathol.* **11**, 1228–1237 (2018).
73. Ma, W. *et al.* Prognostic significance of top2a in non-small cell lung cancer revealed by bioinformatic analysis. *Cancer Cell Int.* **19**, 1–17. https://doi.org/10.1186/s12935-019-0956-1 (2019).
74. Cai, H., Shao, B., Zhou, Y. & Chen, Z. High expression of top2a in hepatocellular carcinoma is associated with disease progression and poor prognosis. *Oncol. Lett.* **20**, 1–9. https://doi.org/10.3892/ol.2020.12095 (2020).
75. Meng, J., Wei, Y., Deng, Q., Li, L. & Li, X. Study on the expression of top2a in hepatocellular carcinoma and its relationship with patient prognosis. *Cancer Cell Int.* **22**, 1–18. https://doi.org/10.1186/s12935-021-02439-0 (2022).
76. Weinstein, J., Jacobsen, F. W., Hsu-Chen, J., Wu, T. & Baum, L. G. A novel mammalian protein, p55cdc, present in dividing cells is associated with protein kinase activity and has homology to the saccharomyces cerevisiae cell division cycle proteins cdc20 and cdc4. *Mol. Cell. Biol.* **14**, 3350–3363. https://doi.org/10.1128/mcb.14.5.3350-3363.1994 (1994).
77. Weinstein, J. Cell cycle-regulated expression, phosphorylation, and degradation of p55cdc: A mammalian homolog of cdc20/fizzy/slp1. *J. Biol. Chem.* **272**, 28501–28511. https://doi.org/10.1074/jbc.272.45.28501 (1997).
78. Kato, T. *et al.* Overexpression of cdc20 predicts poor prognosis in primary non-small cell lung cancer patients. *J. Surg. Oncol.* **106**, 423–430. https://doi.org/10.1002/jso.23109 (2012).

79. Wu, W.-J. *et al.* Cdc20 overexpression predicts a poor prognosis for patients with colorectal cancer. *J. Transl. Med.* **11**, 1–8. https://doi.org/10.1186/1479-5876-11-142 (2013).
80. Karra, H. *et al.* Cdc20 and securin overexpression predict short-term breast cancer survival. *Br. J. Cancer* **110**, 2905–2913. https://doi.org/10.1038/bjc.2014.252 (2014).
81. Tang, J. *et al.* Overexpression of aspm, cdc20, and ttk confer a poorer prognosis in breast cancer identified by gene co-expression network analysis. *Front. Oncol.* **9**, 310–324. https://doi.org/10.3389/fonc.2019.00310 (2019).
82. Ding, Z.-Y., Wu, H.-R., Zhang, J.-M., Huang, G.-R. & Ji, D.-D. Expression characteristics of cdc20 in gastric cancer and its correlation with poor prognosis. *Int. J. Clin. Exp. Pathol.* **7**, 722–727 (2014).
83. Choi, J.-W., Kim, Y., Lee, J.-H. & Kim, Y.-S. High expression of spindle assembly checkpoint proteins cdc20 and mad2 is associated with poor prognosis in urothelial bladder cancer. *Virchows Archiv* **463**, 681–687. https://doi.org/10.1007/s00428-013-1473-6 (2013).
84. Alfarsi, L. H. *et al.* Cdc20 expression in oestrogen receptor positive breast cancer predicts poor prognosis and lack of response to endocrine therapy. *Breast Cancer Res. Treatment* **178**, 535–544. https://doi.org/10.1007/s10549-019-05420-8 (2019).
85. Li, J., Gao, J.-Z., Du, J.-L., Huang, Z.-X. & Wei, L.-X. Increased cdc20 expression is associated with development and progression of hepatocellular carcinoma. *Int. J. Oncol.* **45**, 1547–1555. https://doi.org/10.3892/ijo.2014.2559 (2014).
86. Zhang, X. *et al.* Connection between cdc20 expression and hepatocellular carcinoma prognosis. *Med. Sci. Monitor* **27**, e926760–e926765. https://doi.org/10.12659/MSM.926760 (2021).
87. Bond, J. *et al.* Aspm is a major determinant of cerebral cortical size. *Nat. Genet.* **32**, 316–320. https://doi.org/10.1038/ng995 (2002).
88. Pai, V. C. *et al.* Aspm promotes prostate cancer stemness and progression by augmenting wnt- dvl-3- β-catenin signaling. *Oncogene* **38**, 1340–1353. https://doi.org/10.1038/s41388-018-0497-4 (2019).
89. Hsu, C.-C. *et al.* The differential distributions of aspm isoforms and their roles in wnt signaling, cell cycle progression, and pancreatic cancer prognosis. *J. Pathol.* **249**, 498–508. https://doi.org/10.1002/path.5341 (2019).
90. Zhang, H. *et al.* Aspm promotes hepatocellular carcinoma progression by activating wnt/β-catenin signaling through antagonizing autophagy-mediated dvl2 degradation. *FEBS Open Bio* **11**, 2784–2799. https://doi.org/10.1002/2211-5463.13278 (2021).
91. Xu, Z., Zhang, Q., Luh, F., Jin, B. & Liu, X. Overexpression of the aspm gene is associated with aggressiveness and poor outcome in bladder cancer. *Oncol. Lett.* **17**, 1865–1876. https://doi.org/10.3892/ol.2018.9762 (2019).
92. Jiang, W. *et al.* Prc1: A human mitotic spindle-associated cdk substrate protein required for cytokinesis. *Mol. Cell* **2**, 877–885. https://doi.org/10.1016/S1097-2765(00)80302-0 (1998).
93. Yang, Z. *et al.* Ccnb2, cdc20, aurka, top2a, melk, ncapg, kif20a, ube2c, prc1, and aspm may be potential therapeutic targets for hepatocellular carcinoma using integrated bioinformatic analysis. *Int. J. General Med.* **14**, 10185–10194. https://doi.org/10.2147/IJGM.S341379 (2021).
94. Xiong, Y. *et al.* Ube2c functions as a potential oncogene by enhancing cell proliferation, migration, invasion, and drug resistance in hepatocellular carcinoma cells. *Biosci. Rep.* **39**, 1–8. https://doi.org/10.1042/BSR20182384 (2019).
95. Ieta, K. *et al.* Identification of overexpressed genes in hepatocellular carcinoma, with special reference to ubiquitin-conjugating enzyme e2c gene expression. *Int. J. Cancer* **121**, 33–38. https://doi.org/10.1002/ijc.22605 (2007).
96. Dastsooz, H., Cereda, M., Donna, D. & Oliviero, S. A comprehensive bioinformatics analysis of ube2c in cancers. *Int. J. Mol. Sci.* **20**, 2228–22247. https://doi.org/10.3390/ijms20092228 (2019).
97. Zhang, H. *et al.* Overexpression of ube2c correlates with poor prognosis in gastric cancer patients. *Eur. Rev. Med. Pharmacol. Sci.* **22**, 1665–1671. https://doi.org/10.26355/eurrev_201803_14578 (2018).
98. Petry, S. Mechanisms of mitotic spindle assembly. *Ann. Rev. Biochem.* **85**, 659–683. https://doi.org/10.1146/annurev-biochem-060815-014528 (2016).
99. Hou, S., Hua, L., Wang, W., Li, M. & Xu, L. Nucleolar spindle associated protein 1 (nusap1) facilitates proliferation of hepatocellular carcinoma cells. *Transl. Cancer Res.* **8**, 2113–2120. https://doi.org/10.21037/tcr.2019.09.28 (2019).
100. Han, G. *et al.* Nusap1 gene silencing inhibits cell proliferation, migration and invasion through inhibiting dnmt1 gene expression in human colorectal cancer. *Exp. Cell Res.* **367**, 216–221. https://doi.org/10.1016/j.yexcr.2018.03.039 (2018).
101. Liu, Z. *et al.* High nusap1 expression predicts poor prognosis in colon cancer. *Pathol.-Res. Practice* **214**, 968–973. https://doi.org/10.1016/j.prp.2018.05.017 (2018).
102. Gulzar, Z. G., McKenney, J. K. & Brooks, J. D. Increased expression of nusap in recurrent prostate cancer is mediated by e2f1. *Oncogene* **32**, 70–77. https://doi.org/10.1038/onc.2012.27 (2013).
103. Gordon, C. A., Gong, X., Ganesh, D. & Brooks, J. D. Nusap1 promotes invasion and metastasis of prostate cancer. *Oncotarget* **8**, 29935–29950. https://doi.org/10.18632/oncotarget.15604 (2017).
104. Li, H. *et al.* Nucleolar and spindle associated protein 1 promotes metastasis of cervical carcinoma cells by activating wnt/β -catenin signaling. *J. Exp. Clin. Cancer Res.* **38**, 1–18. https://doi.org/10.1186/s13046-019-1037-y (2019).
105. Roy, S. *et al.* microrna 193a–5p regulates levels of nucleolar-and spindle-associated protein 1 to suppress hepatocarcinogenesis. *Gastroenterology* **155**, 1951–1966. https://doi.org/10.1053/j.gastro.2018.08.032 (2018).
106. Kim, J. H. *et al.* Genomic predictors for recurrence patterns of hepatocellular carcinoma: Model derivation and validation. *PLoS Med.* **11**, e1001770–e1001786. https://doi.org/10.1371/journal.pmed.1001770 (2014).
107. Mah, W.-C. *et al.* Methylation profiles reveal distinct subgroup of hepatocellular carcinoma patients with poor prognosis. *PloS One* **9**, e104158–e104168. https://doi.org/10.1371/journal.pone.0104158 (2014).
108. Ritchie, M. E. *et al.* limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res.* **43**, 1–13. https://doi.org/10.1093/nar/gkv007 (2015).
109. Carlson, M. R. *et al.* Genomic annotation resources in r/bioconductor. *Stat. Genom.* **67**, 90. https://doi.org/10.1007/978-1-4939-3578-9_4 (2016).
110. Wickham, H. *et al.* ggplot2: Create elegant data visualisations using the grammar of graphics (3.3. 6)[computer software]. https://cran.r-project.org/package=ggplot2. Accessed 25 June 2022 (2022).
111. Gaujoux, R. & Seoighe, C. Nmf: Algorithms and framework for nonnegative matrix factorization (nmf). *R Package Version 0.20* **6**, http://CRAN.R-project.org/package=NMF (2015).
112. Hasan, M. A. M., Nasser, M., Pal, B. & Ahmad, S. Support vector machine and random forest modeling for intrusion detection system (ids). *J. Intell. Learn. Syst. Appl.* **2014**, 1. https://doi.org/10.4236/jilsa.2014.61005 (2014).
113. Jan, S. U., Lee, Y.-D., Shin, J. & Koo, I. Sensor fault classification based on support vector machine and statistical time-domain features. *IEEE Access* **5**, 8682–8690. https://doi.org/10.1109/ACCESS.2017.2705644 (2017).
114. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951. https://doi.org/10.1002/pro.3715 (2019).
115. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.*https://doi.org/10.1093/nar/gkac963 *(2022).*
116. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protocols* **4**, 44–57. https://doi.org/10.1038/nprot.2008.211 (2009).
117. Szklarczyk, D. *et al.* The string database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368. https://doi.org/10.1093/nar/gkw937 (2016).

118. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504. https://doi.org/10.1101/gr.123930 (2003).

119. Chin, C.-H. *et al.* cytohubba: Identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* **8**, 1–7. https://doi.org/10.1186/1752-0509-8-S4-S11 (2014).

120. Bader, G. D. & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **4**, 1–27. https://doi.org/10.1186/1471-2105-4-2 (2003).

121. Robin, X. *et al.* Proc: An open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinform.* **12**, 1–8. https://doi.org/10.1186/1471-2105-12-77 (2011).

122. Therneau, T. & Lumley, T. R survival package. *R Core Team*. https://rweb.webapps.cla.umn.edu/R/library/survival/doc/survival.pdf. Accessed 30 June 2022 (2013).

Author contributions

All listed authors participated meaningfully in the study, and they have seen and approved the submission of this manuscript. Conceptualization, M.A.M.H.; Methodology, M.A.M.H., M.M.; Data collection and curation: M.A.M.H., M.M., J.S.; Interpreted and analyzed the data, M.A.M.H., M.M., J.S.; Writing-original draft preparation, M.A.M.H., M.M.; Writing-review and editing, M.A.M.H., M.M., J.S.; Supervision, J.S., M.A.M.H.; Project administration and funding, J.S.

Funding

This work was supported by the Graduate School Research Fund of The University of Aizu, Japan.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-30851-1.

Correspondence and requests for materials should be addressed to J.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.