Front. Med.
DOI 10.1007/s11684-015-0408-9

RESEARCH ARTICLE

TCGA whole-transcriptome sequencing data reveals significantly dysregulated genes and signaling pathways in hepatocellular carcinoma

Daniel Wai-Hung Ho, Alan Ka-Lun Kai, Irene Oi-Lin Ng (✉)

Department of Pathology and State Key Laboratory for Liver Research, The University of Hong Kong, Hong Kong SAR, China

Abstract **This study systematically evaluates the TCGA whole-transcriptome sequencing data of hepatocellular carcinoma (HCC) by comparing the global gene expression profiles between tumors and their corresponding non-tumorous liver tissue. Based on the differential gene expression analysis, we identified a number of novel dysregulated genes, in addition to those previously reported. Top-listing upregulated (*CENPF* and *FOXM1*) and downregulated (*CLEC4G*, *CRHBP*, and *CLEC1B*) genes were successfully validated using qPCR on our cohort of 65 pairs of human HCCs. Further examination for the mechanistic overview by subjecting significantly upregulated and downregulated genes to gene set enrichment analysis showed that different cellular pathways were involved. This study provides useful information on the transcriptomic landscape and molecular mechanism of hepatocarcinogenesis for development of new biomarkers and further in-depth characterization.**

Keywords TCGA; whole-transcriptome sequencing; HCC; liver cancer

Introduction

Hepatocellular carcinoma (HCC) is a common type of cancer and one of the leading causes of cancer-related mortality worldwide [1,2]. HCC is an aggressive malignancy and patients with HCC have a poor prognosis. Unfortunately, only a few effective treatment options are available. Despite much effort in studying the molecular mechanism of HCC carcinogenesis, current understanding on this lethal disease is still limited.

In the past, delineating the underlying genome-wide HCC regulatory and interaction networks primarily relied on microarray-based technology [3–7]. Recent advancement in next-generation sequencing facilitated the realization of whole-transcriptome sequencing (WTS). This new technological platform allows more comprehensive and accurate examination of global gene expression profile. Currently, only a few studies have utilized WTS strategies in delineating the transcriptomic landscape of HCC [8,9] or liver cancer stem cells [10]. However, all of them are limited by small sample size in providing a comprehensive and representative overview of HCC transcriptome. The Cancer Genome Atlas (http://cancergenome.nih.gov/) represents a global collaboration in cancer research. It has large collections of tissue samples, which were examined in multiple aspects (e.g., genomic, transcriptomic, and epigenetic). More importantly, the data are of open access and freely available to all researchers for use in their own studies. Therefore, the relatively large TCGA HCC WTS data set was utilized in the discovery of the current study.

In our study, we extracted WTS data from the collections of free-access repositories from all 50 HCC cases, in which tumorous (T) and their corresponding non-tumorous (NT) liver tissue was available and analyzed by TCGA. We compared global gene expression profiles between T and NT liver tissue and identified differentially expressed (DE) genes. Top-listing genes were validated by quantitative PCR (qPCR) by using an independent sample cohort (n = 65). DE genes were then subjected to gene set enrichment analysis, and we identified gene sets and signaling pathways that were significantly enriched with upregulated and downregulated genes. These genes are attractive molecular targets and are worthy of further investigation, and they may be used as HCC biomarkers.

Received January 14, 2015; accepted May 25, 2015
Correspondence: iolng@hku.hk

Materials and methods

TCGA WTS data of HCC

From the TCGA data portal (http://cancergenome.nih.gov/), we extracted all available WTS data of HCC (a total of 50 cases), which have both T and their corresponding NT samples, through bulk download mode [liver HCC (cancer type), RNASeqV2 (data type), level 3 (archive type) and 1.12.0 (data version)]. The data were generated based on Illumina HiSeq 2000 platform and annotated to reference transcript set of UCSC hg19 gene standard track. Gene expression data were available as upper quartile normalized RSEM count estimates. Extracted data were used without further transformation, except by rounding off values to integers.

Validation sample cohort of paired HCCs

A cohort of 65 surgically resected HCCs and their corresponding NT livers were randomly selected for validation. The specimens were collected from patients who underwent surgical resection for HCC at Queen Mary Hospital, Hong Kong. All of them were obtained immediately after surgical resection, snap-frozen in liquid nitrogen and kept at −80°C. Each case had both frozen tissue blocks and formalin-fixed paraffin-embedded tissue; frozen sections were cut from tumor blocks and stained for histological examination to ensure a homogenous cell population of tissue. The use of the tissue was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster. The demographic data of the patients are summarized in Supplementary Table 1.

Differential gene expression detection

Differential gene expression (DGE) analysis was performed using edgeR [11]. It uses negative binomial models to capture variance dispersion for WTS read count data, empirical Bayes estimation for gene-specific variation, and generalized linear models applicable to general experiments. As suggested by edgeR, genes with very low read counts are usually not of interest in DGE analysis; hence, average count-per-million (CPM) was used to determine whether a gene was reasonably expressed or not. Subsequently, log2(fold change), log2(CPM), statistical significance, and the corresponding false discovery rate (FDR) were reported by edgeR. DE genes were selected based on these parameters, with the T/NT expression fold change (FC) denoting upregulation or downregulation.

Gene set enrichment analysis on DE genes

To evaluate the mechanistic overview of DGE for HCC, the significantly upregulated and downregulated genes were tested for gene set or pathway enrichment by uGPA package [12]. Enrichment analyses of the upregulated and downregulated genes were performed separately as recommended previously [13]. Curated gene sets were obtained from MSigDB v4.0 (www.broadinstitute.org/gsea/msigdb) and classified into functional gene sets according to the domains of gene ontology (GO) [i.e., biological process (825 gene sets), cellular component (233 gene sets), and molecular function (396 gene sets)] or pathway gene sets according to canonical pathways as documented by KEGG (186 gene sets). uGPA takes DGE events as input and assesses them for enrichment events within gene sets or signaling pathways by cumulative hypergeometric test. An FDR of < 0.05 was treated as significant event.

Validation on top-listing gene candidates by qPCR in human HCCs

To confirm the WTS findings on DGE, the top-listing upregulated (*CENPF* and *FOXM1*) and downregulated (*CLEC4G*, *CRHBP*, and *CLEC1B*) genes were subjected to qPCR by TaqMan real-time qPCR assays (Hs01118845_m1, Hs01073586_m1, Hs00962163_g1, Hs00181810_m1, and Hs00212925_m1), following manufacturer's instructions. Total RNA was extracted by Trizol (Invitrogen) and cDNA was synthesized by reverse transcription kit (Life Technologies) on the validation sample cohort ($n = 65$).

Results

Comparison of global gene expression profiles of HCC T and NT tissue

By comparing the WTS read counts of the various genes between T and NT tissue and subsequently applying the selection criteria of log2(FC) $\geqslant 2$, log2(CPM) $\geqslant 1$, and FDR < 0.05, 734 genes were regarded as having DGE, among which 220 were upregulated and 514 were downregulated (Fig. 1). In terms of statistical significance, *CENPF* (centromere protein F, 350/400 kDa) (log2FC = 3.64, FDR = 5.32E−78) and *CLEC4G* (C-type lectin domain family 4, member G) (log2FC = −8.96, FDR = 1.19E−80) were the most significantly upregulated and downregulated genes, respectively (Supplementary Tables 2 and 3).

Successful validation of top-listing candidates by qPCR

Top-listing upregulated (*CENPF* and *FOXM1*) and downregulated (*CLEC4G*, *CRHBP*, and *CLEC1B*) genes were subjected to qPCR assays on our validation sample cohort of 65 HCC pairs. All of these genes were found to be

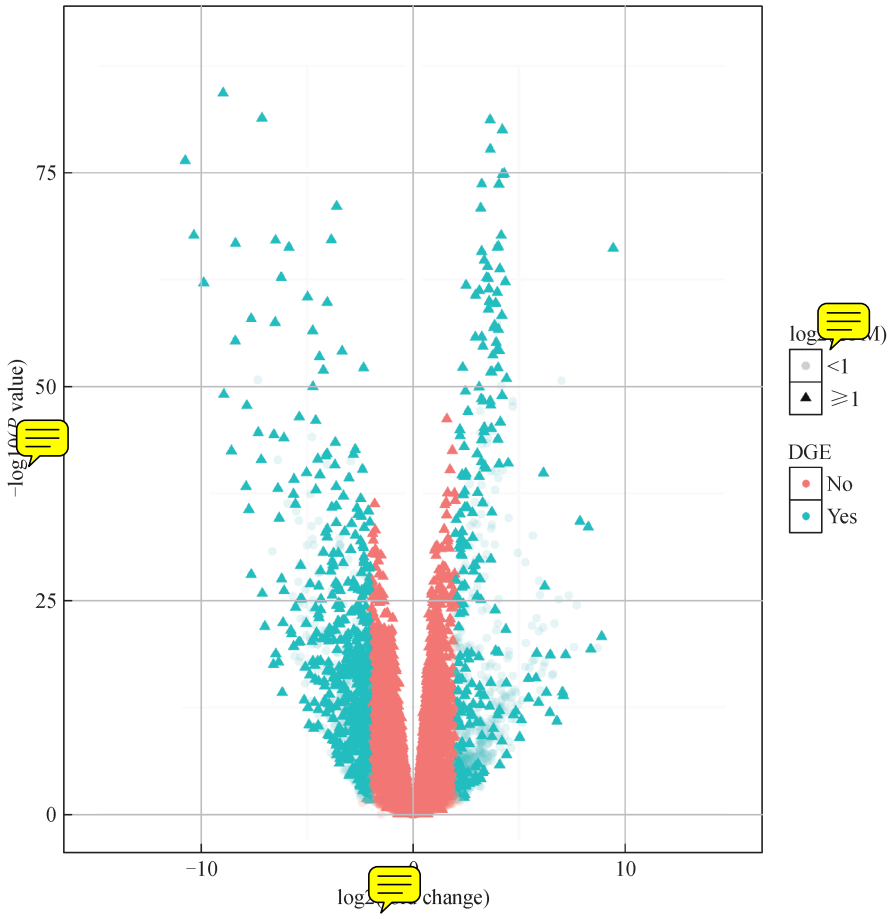

Fig. 1 Volcano plot of the WTS data. The color of the data points denotes the status of DGE and the intensity (light vs. dark) and shape (round dot vs. triangle) of the data points denote the average expression level of genes as defined by log2(CPM) (< 1 vs. $\geqslant 1$).

successfully validated ($P < 0.0001$, Mann-Whitney U test) and the dysregulation trend matched with those observed in the TCGA WTS data (Fig. 2).

Significantly enriched pathways for upregulated and downregulated genes

By subjecting the significantly upregulated genes to enrichment analysis on gene sets based on GO (i.e., biological process, cellular component, and molecular function) and KEGG canonical pathways, we observed that upregulated genes were significantly enriched in various domains (Table 1). For GO biological process, the genes were mainly enriched in cell cycle processes. For GO cellular component, non-membrane-bound organelles and cytoskeleton were involved. For GO molecular function, motor activity and various binding activities were implicated. Based on the canonical signaling pathways documented in KEGG, pathways on cell cycle and p53 signaling were significantly enriched.

Meanwhile, downregulated genes were also subjected to gene set enrichment analysis (Table 2). For GO biological process, the genes were mainly related to signal transduction, response to stimulus, and various metabolic processes. For GO cellular component, they were implicated in membrane and extracellular matrix (ECM). For GO molecular function, they were involved in versatile types of activities including oxygen binding, receptor activity, and oxidoreductase activity. They were also enriched in canonical signaling pathways that are related to metabolism of various substrates.

Discussion

In the current study, we made use of the T-NT TCGA WTS data extracted from 50 HCC pairs to provide useful transcriptomic landscape for HCC. We systematically compared the gene expression profiles of HCC T samples with their corresponding NT samples, and identified 734

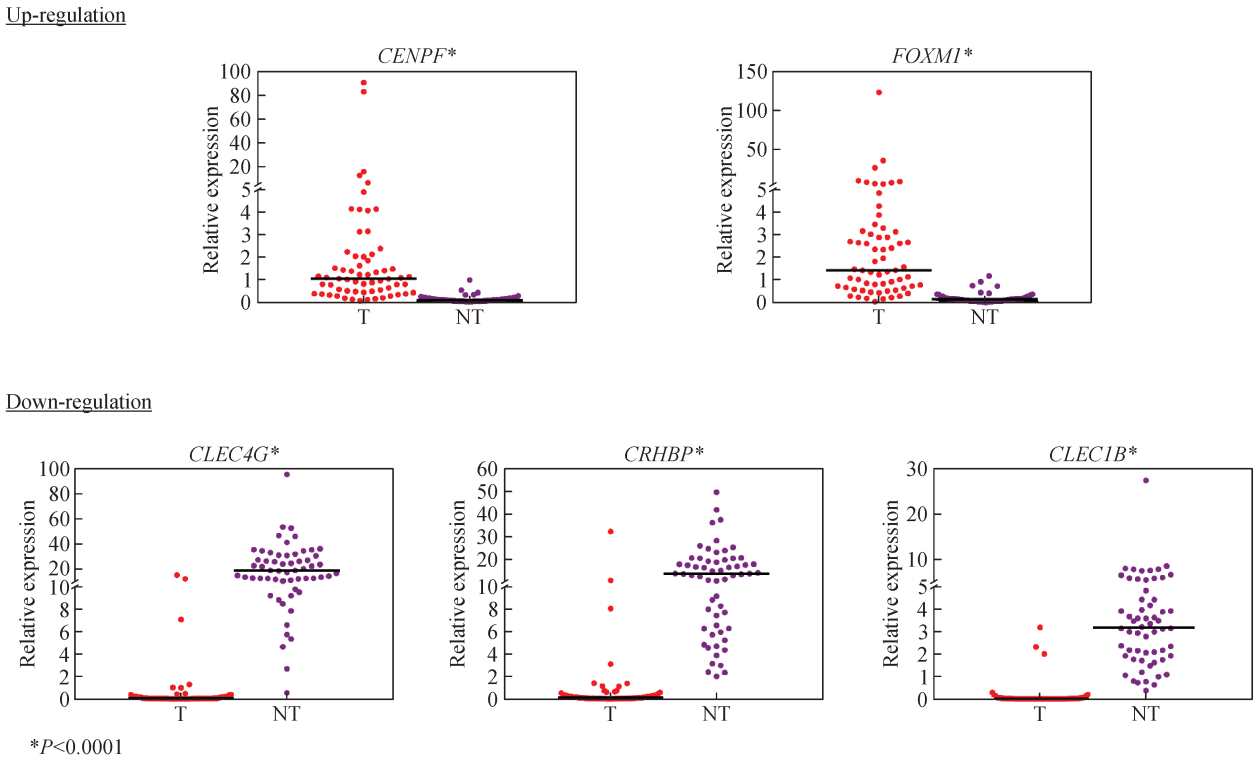

Fig. 2 Successful qPCR validation of top-listing DE genes.

Table 1 Summary of gene set enrichment analysis on significantly upregulated genes

Gene set	# of genes	# of DE genes	*P* value	FDR	DE genes
GO biological process					
MITOTIC_CELL_CYCLE	153	27	6.91E−33	5.70E−30	*PRC1, PKMYT1, AURKA, CDKN2A, CDKN2C, CCNA2, CDCA5, MAD2L1, ZWINT, NEK2, ANLN, NDC80, PLK1, E2F1, KIF23, KIF2C, DLGAP5, CDC6, KIF11, UBE2C, BUB1B, NCAPH, BUB1, CENPF, BIRC5, CENPE, CDKN3*
CELL_CYCLE_PROCESS	193	28	1.48E−31	6.09E−29	*AURKA, CDCA5, CCNA2, MAD2L1, ZWINT, NEK2, NDC80, CDC6, NCAPH, BUB1, BIRC5, CDKN3, PRC1, PKMYT1, CDKN2A, CDKN2C, ANLN, PLK1, E2F1, KIF23, KIF2C, DLGAP5, KIF11, UBE2C, BUB1B, CENPF, CENPE, RACGAP1*
CELL_CYCLE_GO_0007049	315	32	6.90E−31	1.90E−28	*AURKA, CCNA2, CDCA5, MAD2L1, ZWINT, NEK2, NDC80, CDC20, CDT1, CDC6, NCAPH, CDC45, BUB1, BIRC5, CDKN3, PRC1, PKMYT1, CDKN2A, CDKN2C, ANLN, PLK1, E2F1, KIF23, KIF2C, DLGAP5, KIF11, MCM2, UBE2C, BUB1B, CENPF, CENPE, RACGAP1*
CELL_CYCLE_PHASE	170	25	2.02E−28	4.17E−26	*PKMYT1, AURKA, CDKN2A, CDKN2C, CDCA5, CCNA2, MAD2L1, ZWINT, NEK2, ANLN, NDC80, PLK1, E2F1, KIF2C, DLGAP5, CDC6, KIF11, UBE2C, BUB1B, NCAPH, BUB1, CENPF, BIRC5, CENPE, CDKN3*
M_PHASE_OF_MITOTIC_CELL_CYCLE	85	19	1.49E−25	2.45E−23	*PKMYT1, AURKA, KIF2C, DLGAP5, CDCA5, CCNA2, KIF11, UBE2C, MAD2L1, ZWINT, BUB1B, NEK2, ANLN, NCAPH, BUB1, BIRC5, NDC80, CENPE, PLK1*

(*Continued*)

Gene set	# of genes	# of DE genes	*P* value	FDR	DE genes
GO cellular component					
NON_MEMBRANE_BOUND_ ORGANELLE	631	29	2.11E−18	2.46E−16	*CDCA5, ZWINT, KIF4A, NDC80, MAPT, CDK1, BUB1, CCNB2, PRC1, CDKN2A, ACTN2, ANLN, PLK1, KIF23, KIF2C, DLGAP5, KIF11, NEB, AURKA, MAD2L1, NEK2, CDC20, CDT1, TOP2A, BIRC5, MCM2, BUB1B, CENPF, CENPE*
INTRACELLULAR_NON_MEMBRANE_ BOUND_ORGANELLE	631	29	2.11E−18	2.46E−16	CDCA5, ZWINT, KIF4A, NDC80, MAPT, CDK1, BUB1, CCNB2, PRC1, CDKN2A, ACTN2, ANLN, PLK1, KIF23, KIF2C, DLGAP5, KIF11, NEB, AURKA, MAD2L1, NEK2, CDC20, CDT1, TOP2A, BIRC5, MCM2, BUB1B, CENPF, CENPE
MICROTUBULE_CYTOSKELETON	152	17	1.36E−17	1.05E−15	*PRC1, AURKA, KIF4A, NEK2, CDC20, PLK1, KIF23, KIF2C, DLGAP5, MAPT, TOP2A, KIF11, CDK1, BUB1, CENPF, BIRC5, CCNB2*
SPINDLE	39	11	1.51E−16	8.81E−15	*KIF23, KIF4A, PRC1, AURKA, DLGAP5, BUB1, KIF11, CDK1, CENPF, BIRC5, CDC20*
CYTOSKELETAL_PART	235	18	1.29E−15	6.03E−14	*AURKA, KIF4A, NEK2, CDC20, MAPT, TOP2A, CDK1, BUB1, BIRC5, PRC1, ACTN2, ANLN, PLK1, KIF23, KIF2C, DLGAP5, KIF11, CENPF*
GO molecular function					
MICROTUBULE_MOTOR_ACTIVITY	16	5	1.96E−08	7.77E−06	*KIF23, KIF4A, KIF11, CENPE, KIF2C*
MOTOR_ACTIVITY	28	5	4.19E−07	8.29E−05	*KIF23, KIF4A, KIF2C, KIF11, CENPE*
CYTOSKELETAL_PROTEIN_BINDING	159	7	2.97E−05	0.004	*NRCAM, ACTN2, ANLN, MAPT, BIRC5, RACGAP1, MAPK8IP2*
CARBOHYDRATE_BINDING	72	5	4.90E−05	0.005	*REG3A, CD34, MDK, THBS4, LPL*
PROTEIN_KINASE_REGULATOR_ ACTIVITY	39	4	6.21E−05	0.005	*SFN,CDKN2A,CDKN2C,MAPK8IP2*
KEGG pathway					
KEGG_CELL_CYCLE	128	19	6.25E−22	1.16E−19	*CDC45, PLK1, CCNA2, BUB1, MCM2, PTTG1, CDC6, CDC20, CCNB1, CCNE1, SFN, E2F1, CDK1, BUB1B, MAD2L1, CDKN2A, CDKN2C, PKMYT1, CCNB2*
KEGG_OOCYTE_MEIOSIS	114	11	4.06E−11	3.78E−09	*PLK1, BUB1, PTTG1, CDC20, CCNB1, CCNE1, CDK1, MAD2L1, AURKA, PKMYT1, CCNB2*
KEGG_PROGESTERONE_MEDIATED_ OOCYTE_MATURATION	86	8	2.62E−08	1.63E−06	*CDK1, MAD2L1, PLK1, CCNA2, BUB1, PKMYT1, CCNB2, CCNB1*
KEGG_P53_SIGNALING_PATHWAY	69	7	1.09E−07	5.06E−06	*SFN, CDK1, CDKN2A, RRM2, CCNB2, CCNB1, CCNE1*
KEGG_TIGHT_JUNCTION	134	4	0.006	0.226	*MYH4, ACTN2, CTNNA2, PPP2R2C*

Table 2 Summary of gene set enrichment analysis on significantly downregulated genes

Gene set	# of genes	# of DE genes	*P* value	FDR	DE genes
GO biological process					
SIGNAL_TRANSDUCTION	1634	70	9.98E−19	8.24E−16	*DIRAS3, IGF1, IGF2, HPGD, GNA14, MARCO, CCL3, ADRA2B, TBXA2R, ADRA1B, ADRA1A, WNK2, IL1RL1, NR1I2, CXCL6, GRIA3, FCGR2B, GABRB3, CAMK2B, BCL2L10, AVPR1A, TRPV4, CCL19, NPY1R, APOA1, NR4A3, ESR1, CHRNA4, LILRB5, PDGFRA, ECM1, TNFRSF11B, GADD45G, GADD45B, CLEC1B, IL18R1, SOCS2, SOCS3, PTH1R, CTNND2, PTPRD, CRHBP, LIFR, CEACAM6, FPR1, CXCL12, NR4A1, CXCL14, SKAP1, WISP2, MCC, RND2, TACSTD2, EPHA2, NTRK2, TGFA, CHL1, LY6E, VIPR1, CD79A, IGFBP1, PRKAR2B, RET, RCAN1, ANXA3, SFRP5, SFRP1, GCGR, IGFALS, DTX1*

(*Continued*)

Gene set	# of genes	# of DE genes	*P* value	FDR	DE genes
RESPONSE_TO_EXTERNAL_STIMULUS	312	25	5.02E−13	2.07E−10	*S100A8, SERPINE1, F9, SAA1, IL1RAP, TRPV4, CXCL2, FPR1, CXCL6, GCGR, CCL21, LECT2, CCL19, CD1D, ORM1, SELE, CXCL12, PGLYRP2, CCL3, ALB, LYVE1, CXCL14, CXCL13, FOS, MBL2*
CARBOXYLIC_ACID_METABOLIC_PROCESS	178	19	2.13E−12	5.38E−10	*SLC7A8, BBOX1, ASPA, GGT5, CYP39A1, ACOT12, GSTZ1, SLC3A1, SDS, HAO2, AKR1D1, SLC27A5, GLS2, FTCD, IGF1, CYP4A11, GLYAT, GCK, HPGD*
ORGANIC_ACID_METABOLIC_PROCESS	180	19	2.61E−12	5.38E−10	*SLC7A8, BBOX1, ASPA, GGT5, CYP39A1, ACOT12, GSTZ1, SLC3A1, SDS, HAO2, AKR1D1, SLC27A5, GLS2, FTCD, IGF1, CYP4A11, GLYAT, GCK, HPGD*
LIPID_METABOLIC_PROCESS	325	24	8.36E−12	1.38E−09	*CYP3A4, ALDH8A1, LCAT, APOF, PITPNM3, NR1I2, CYP39A1, CETP, THRSP, HAO2, SLC27A5, HPGD, APOA4, APOA1, BCO2, ACOT12, NPC1L1, IP6K3, AKR1D1, CYP4A11, GLYAT, RDH16, UGT2B7, SMPD3*
GO cellular component					
MEMBRANE	1994	90	1.23E−25	2.87E−23	*IL1RAP, GHR, SLC22A1, CFTR, SLC3A1, C8A, GNA14, MARCO, ADRA2B, LYVE1, TBXA2R, ADRA1B, ADRA1A, TREH, CA9, CD4, GPR128, ABCB11, STEAP4, CD163, GRIA3, CD1D, SLC16A4, GABRB3, NAPSB, CNGA1, PLEKHB1, PTPRS, AVPR1A, UNC93A, TRPV4, SELP, NPY1R, SELE, SLC34A2, NCAM1, CNTFR, MRC1, HS3ST3A1, MAN1C1, SLCO1B3, CHRNA4, CLEC4M, PDGFRA, HS3ST3B1, CLEC1B, IL18R1, PKHD1, PTH1R, RHBG, CR1, PTPRD, LIFR, SRPX, CEACAM6, C7, C9, FPR1, STAB2, SLC13A2, CLDN2, PRSS8, GGT5, EPCAM, TACSTD2, MME, EPHA2, NTRK2, CHL1, LY6E, PROM1, VIPR1, FXYD1, ITGB8, ITGA9, CD79A, SLC5A1, NGFR, PITPNM3, CDHR2, SLC7A8, KCND3, B3GAT1, SIGLEC7, VSIG2,CLDN10, GCGR, SLC6A2, BASP1, SLC10A1*
PLASMA_MEMBRANE	1426	75	2.65E−25	3.09E−23	*LIFR, IL1RAP, CEACAM6, SLC22A1, GHR, CFTR, SLC3A1, C9, FPR1, STAB2, GNA14, MARCO, ADRA2B, LYVE1, ADRA1B, TBXA2R, SLC13A2, ADRA1A, TREH, CLDN2, PRSS8, CD4, ABCB11, EPCAM, TACSTD2, MME, EPHA2, NTRK2, STEAP4, CD163, LY6E, GRIA3, CD1D, PROM1, SLC16A4, GABRB3, VIPR1, CNGA1, FXYD1, ITGB8, ITGA9, CD79A, SLC5A1, NGFR, PTPRS, AVPR1A, UNC93A, TRPV4, SELP, NPY1R, SELE, SLC7A8, KCND3, SLC34A2, NCAM1, MRC1, SIGLEC7, SLCO1B3, CHRNA4, CLEC4M, VSIG2, PDGFRA, HS3ST3B1, CLEC1B, CLDN10, GCGR, IL18R1, SLC6A2, BASP1, PKHD1, PTH1R, RHBG, CR1, PTPRD, SLC10A1*
INTRINSIC_TO_MEMBRANE	1348	72	1.09E−24	8.46E−23	*LIFR, IL1RAP, CEACAM6, SLC22A1, GHR, SLC3A1, C8A, C7, C9, FPR1, STAB2, MARCO, ADRA2B, LYVE1, TBXA2R, ADRA1B, ADRA1A, SLC13A2, TREH, CA9, GPR128, ABCB11, GGT5, TACSTD2, MME, EPHA2, NTRK2, CD163, CHL1, LY6E, CD1D, PROM1, SLC16A4, GABRB3, VIPR1, CNGA1, FXYD1, ITGB8, ITGA9, SLC5A1, PLEKHB1, NGFR, PTPRS, PITPNM3, CDHR2, AVPR1A, SELP, NPY1R, SLC7A8, KCND3, SLC34A2, NCAM1, MRC1, B3GAT1, SIGLEC7, HS3ST3A1, MAN1C1, SLCO1B3, CHRNA4, CLEC4M, VSIG2, PDGFRA, HS3ST3B1, CLEC1B, GCGR, SLC6A2, PKHD1, PTH1R, RHBG, CR1, PTPRD, SLC10A1*
INTEGRAL_TO_MEMBRANE	1330	70	1.21E−23	7.03E−22	*LIFR, IL1RAP, CEACAM6, GHR, SLC22A1, SLC3A1, C8A, C7, C9, FPR1, STAB2, MARCO, ADRA2B, LYVE1, ADRA1B, TBXA2R, ADRA1A, SLC13A2, CA9, GPR128, ABCB11, GGT5, TACSTD2, MME, EPHA2, NTRK2, CD163, CHL1, LY6E, CD1D, PROM1, SLC16A4, GABRB3, VIPR1, CNGA1, FXYD1, ITGB8, ITGA9, SLC5A1, PLEKHB1, NGFR, PTPRS, PITPNM3, CDHR2, AVPR1A, SELP, NPY1R, SLC7A8, KCND3, SLC34A2, NCAM1, MRC1, B3GAT1, SIGLEC7, HS3ST3A1, MAN1C1, SLCO1B3, CHRNA4, CLEC4M, VSIG2, PDGFRA, HS3ST3B1, CLEC1B, GCGR, SLC6A2, PTH1R, RHBG, CR1, PTPRD, SLC10A1*

(*Continued*)

Gene set	# of genes	# of DE genes	*P* value	FDR	DE genes
MEMBRANE_PART	1670	78	4.27E−23	1.99E−21	*IL1RAP, SLC22A1, GHR, CFTR, SLC3A1, C8A, GNA14, MARCO, ADRA2B, LYVE1, TBXA2R, ADRA1B, ADRA1A, TREH, CA9, GPR128, ABCB11, CD163, CD1D, SLC16A4, GABRB3, CNGA1, PLEKHB1, PTPRS, AVPR1A, SELP, NPY1R, SLC34A2, NCAM1, CNTFR, MRC1, HS3ST3A1, MAN1C1, SLCO1B3, CHRNA4, CLEC4M, PDGFRA, HS3ST3B1, CLEC1B, PKHD1, PTH1R, RHBG, CR1, PTPRD, LIFR, CEACAM6, C7, C9, FPR1, STAB2, SLC13A2, CLDN2, GGT5, TACSTD2, MME, EPHA2, NTRK2, CHL1, LY6E, PROM1, VIPR1, FXYD1, ITGB8, ITGA9, CD79A, SLC5A1, NGFR, PITPNM3, CDHR2, SLC7A8, KCND3, B3GAT1, SIGLEC7, VSIG2, CLDN10, GCGR, SLC6A2, SLC10A1*
GO molecular function					
OXYGEN_BINDING	22	12	9.98E−18	3.95E−15	*CYP3A4, CYP3A7, CYP1A1, CYP2C19, CYP26A1, CYP1A2, CYP2E1, ALB, CYP2A6, CYP2A7, CYP8B1, HBB*
RECEPTOR_ACTIVITY	583	34	3.28E−13	6.50E−11	*LIFR, PTPRS, AVPR1A, IL13RA2, GHR, FPR1, STAB2, HPGD, MARCO, PGLYRP2, RET, CNTFR, ADRA2B, GABRP, MRC1, NR4A3, ADRA1B, TBXA2R, ADRA1A, SIGLEC7, MCC, CD4, GPR128, CHRNA4, LILRB5, TACSTD2, PDGFRA, TNFRSF11B, CLEC1B, GCGR, GABRB3, VIPR1, PTH1R, PTPRD*
OXIDOREDUCTASE_ACTIVITY	289	24	6.83E−13	9.02E−11	*CYP3A4, ALDH8A1, BBOX1, CYP26A1, CYP1A2, GPD1, CYP8B1, CYP39A1, GSTZ1, HAO2, PHGDH, HPGD, ADH1B, AKR7A3, KMO, BCO2, TDO2, SRD5A2, ACADL, CYP2A6, ADH6, ADH4, CYP4A11, RDH16*
TRANSMEMBRANE_ RECEPTOR_ACTIVITY	418	25	2.78E−10	2.75E−08	*LIFR, PTPRS, AVPR1A, IL13RA2, GHR, FPR1, STAB2, HPGD, CNTFR, ADRA2B, GABRP, ADRA1B, TBXA2R, ADRA1A, CD4, GPR128, CHRNA4, LILRB5, PDGFRA, CLEC1B, GCGR, GABRB3, VIPR1, PTH1R, PTPRD*
RECEPTOR_BINDING	377	23	9.87E−10	7.82E−08	*APOF, PITPNM3 ,IL1RN, TNFRSF11B, SAA1, TGFA, CXCL2, CXCL6, CCL21, CCL19, IGF1, IGF2, CXCL12, APOA1, SOCS2, CCL3, ANGPTL1, CXCL14, CXCL13, MBL2, ADAMTS13, TNXB, DTX1*
KEGG pathway					
KEGG_RETINOL_METABOLISM	64	25	6.92E−31	1.29E−28	*CYP4A11, CYP3A4, ADH1B, ADH1C, ADH4, ADH1A, CYP26A1, CYP2C9, CYP2C19, CYP2C8, CYP2B6, CYP2A13, UGT1A4, CYP3A7, LRAT, CYP2A6, CYP2A7, CYP4A22, CYP1A1, CYP1A2, ADH6, UGT2A1, CYP3A43, RDH16, UGT2B7*
KEGG_DRUG_METABOLISM_ CYTOCHROME_P450	72	23	4.00E−26	3.72E−24	*CYP2E1, CYP3A4,GSTZ1, ADH1B, ADH1C, ADH4, ADH1A, CYP2C9, CYP2C19, CYP2C8, CYP2B6, CYP2A13, UGT1A4, CYP3A7, GSTM5, GSTA2, CYP2A6, CYP2A7, CYP1A2, ADH6, UGT2A1, CYP3A43, UGT2B7*
KEGG_METABOLISM_OF_ XENOBIOTICS_BY_ CYTOCHROME_P450	70	21	2.63E−23	1.63E−21	*CYP2E1, CYP3A4, GSTZ1, ADH1B, ADH1C, ADH4, ADH1A, CYP2C9, CYP2C19, CYP2C8, CYP2B6, UGT1A4, CYP3A7, GSTM5, GSTA2, CYP1A1, CYP1A2, ADH6, UGT2A1, CYP3A43, UGT2B7*
KEGG_DRUG_METABOLISM_ OTHER_ENZYMES	51	12	1.75E−12	8.13E−11	*CYP3A4, UPP2, CDA, CYP2A13, UGT1A4, CYP3A7, CYP2A6, CYP2A7, UGT2A1, CYP3A43, NAT2, UGT2B7*
KEGG_LINOLEIC_ACID_ METABOLISM	29	9	6.96E−11	2.59E−09	*CYP2E1, CYP3A4, CYP1A2, PLA2G2A, CYP3A43, CYP2C9, CYP2C19, CYP2C8, CYP3A7*

DE genes. A number of DE genes that were reported in previous studies [8,9], such as *ALG1L*, *SERPINA11*, *TMEM82*, *GPC3*, *SPINK1*, and *ESM1*, were also detected in the current study. In addition, many other novel genes were found to be significantly upregulated (Supplementary Table 2) and downregulated (Supplementary Table 3). *CENPF* (centromere protein F) and *FOXM1* (forkhead box M1) were among the top-listing significantly upregulated genes. CENPF is required for kinetochore function and chromosome segregation in mitosis. On the other hand,

FOXM1 is a transcription factor that regulates the expression of cell cycle genes for DNA replication and mitosis. It may also have roles in controlling cell proliferation and DNA-break repair of DNA damage checkpoint response. Intriguingly, through an integrative computational approach in which the interactomes of human and mice were compared, *CENPF* and *FOXM1* were predicted to be the master regulators for prostate cancer malignancy [14]. Moreover, they were also shown to act synergistically in driving aggressive prostate cancer. Knockdown of *CENPF* and *FOXM1* synergistically reduced the proliferation of prostate cancer cells and tumor growth in cell-line-derived xenografts. It was further shown that knockdown of *CENPF* expression reduced the binding of FOXM1 to its targets. These two proteins were also demonstrated to co-localize in nucleus and their co-expression was a robust prognostic indicator of poor survival and metastasis. Thus, the concurrent upregulation of them in HCC likely suggests a similar synergistic co-operation in hepatocarcinogenesis.

Among the most significantly downregulated genes, we noted multiple members of the C-type lectin family (*CLEC4G*, *CLEC1B*, and *CLEC4M*) and *CRHBP* [corticotropin-releasing factor (CRF) binding protein]. C-type lectins are calcium-dependent glycan binding proteins and function as adhesion and signaling receptors in various immune functions, including inflammation and immunity to tumor and virally infected cells [15]. According to the Human Protein Atlas [16], *CLEC4G*, *CLEC1B*, and *CLEC4M* are predominantly expressed in liver; however, *CLEC4G* and *CLEC4M* are expressed at very low levels or are undetectable in liver cancer tissue (data not available for *CLEC1B* on liver cancer tissue). This finding suggests that disruption of expression of these C-type lectin proteins may have a role in the pathogenesis of HCC. CRHBP is a member of the CRF system. Activation of CRF receptors, particularly CRFR2 was shown to inhibit tumor progression, modulate proliferation and apoptosis, and interfere with angiogenesis through reduction of VEGF expression *in vivo* in various cancers [17–20]. A recent study also indicated that reduced expression of *CRHBP* was associated with a more aggressive behavior of human kidney cancer, suggesting depletion of CRHBP may be involved in renal carcinogenesis [21].

Gene set enrichment analysis further provides a mechanistic overview of HCC. First, proteins of various cell cycle processes were frequently upregulated, particularly for multiple cyclins and cyclin-dependent kinases (*CCNA2*, *CCNB1*, *CCNB2*, *CCNE1*, *CDK1*, *CDKN2A*, *CDKN2C,* and *CDKN3*) (Supplementary Tables 2 and 4). Given that cell cycle is controlled at various checkpoints by regulating cyclins, cyclin-dependent kinases and other cell cycle proteins [22,23], upregulation of these genes may lead to disruption in cell cycle control and result in abnormal cell proliferation. Second, the expression of many genes for various metabolic processes was preferentially downregulated in HCC, including metabolism of retinol, fatty acids, amino acids and carbohydrates, steroid hormone biosynthesis, and glycolysis and gluconeogenesis. In particular, multiple components of cytochrome P450 were significantly downregulated in HCC (Supplementary Tables 3 and 5) and they play critical roles in biosynthesis and metabolism [24]. Besides, they are also involved in the removal of toxic substances from the body [25,26]. Meanwhile, numerous cytokines (CCLs and CXCLs) were also downregulated in HCC (Supplementary Tables 3 and 5). Cytokines and its receptors are important for triggering immune responses through the action of various immune cells [27]. These immune responses are critical in responses against infection [28] and cancer [29]. Overall, these findings suggest altered metabolic and immune systems of HCC compared with non-tumorous hepatocytes.

In the initial global analyses of the TCGA WTS data of HCC and subsequent validation by an independent sample cohort, we discovered several promising gene candidates and pathways that are significantly dysregulated in HCC. These findings shed light on some novel targets that may potentially drive hepatocarcinogenesis. However, further functional characterization and *in vivo* validation using animal model are needed to substantiate our findings.

In conclusion, this study explored the molecular mechanism of hepatocarcinogenesis through assessment of TCGA WTS data of HCC and validation of some of the top-listing DE genes in an independent cohort. It provides useful information on the transcriptomic landscape as well as a mechanistic overview of HCC. Our findings offer novel insights and useful support in biomarker development and suggest new potential targets in HCC characterization.

Acknowledgements

The study was supported in part by the SK Yee Medical Research Fund 2011 and the Small Project Funding (201309176065). IOL Ng is Loke Yew Professor in Pathology.

Compliance with ethics guidelines

Daniel WH Ho, Alan KL Kai, and Irene OL Ng declare that they have no conflict of interest. The use of human tissue in this study was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (UW 09-185).

Electronic Supplementary Material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s11684-015-0408-9 and is accessible for authorized users.

References

1. El-Serag HB. Hepatocellular carcinoma. N Engl J Med 2011; 365 (12): 1118–1127
2. Villanueva A, Llovet JM. Liver cancer in 2013: Mutational landscape of HCC—the end of the beginning. Nat Rev Clin Oncol 2014; 11(2): 73–74
3. Jia HL, Ye QH, Qin LX, Budhu A, Forgues M, Chen Y, Liu YK, Sun HC, Wang L, Lu HZ, Shen F, Tang ZY, Wang XW. Gene expression profiling reveals potential biomarkers of human hepatocellular carcinoma. Clin Cancer Res 2007; 13(4): 1133–1139
4. Lee JS, Thorgeirsson SS. Comparative and integrative functional genomics of HCC. Oncogene 2006; 25(27): 3801–3809
5. Marshall A, Lukk M, Kutter C, Davies S, Alexander G, Odom DT. Global gene expression profiling reveals SPINK1 as a potential hepatocellular carcinoma marker. PLoS ONE 2013; 8(3): e59459
6. Patil MA, Chua MS, Pan KH, Lin R, Lih CJ, Cheung ST, Ho C, Li R, Fan ST, Cohen SN, Chen X, So S. An integrated data analysis approach to characterize genes highly expressed in hepatocellular carcinoma. Oncogene 2005; 24(23): 3737–3747
7. Skawran B, Steinemann D, Weigmann A, Flemming P, Becker T, Flik J, Kreipe H, Schlegelberger B, Wilkens L. Gene expression profiling in hepatocellular carcinoma: upregulation of genes in amplified chromosome regions. Mod Pathol 2008; 21(5): 505–516
8. Huang Q, Lin B, Liu H, Ma X, Mo F, Yu W, Li L, Li H, Tian T, Wu D, Shen F, Xing J, Chen ZN. RNA-Seq analyses generate comprehensive transcriptomic landscape and reveal complex transcript patterns in hepatocellular carcinoma. PLoS ONE 2011; 6(10): e26168
9. Lin KT, Shann YJ, Chau GY, Hsu CN, Huang CY. Identification of latent biomarkers in hepatocellular carcinoma by ultra-deep whole-transcriptome sequencing. Oncogene 2014; 33(39): 4786–4794
10. Ho DW, Yang ZF, Yi K, Lam CT, Ng MN, Yu WC, Lau J, Wan T, Wang X, Yan Z, Liu H, Zhang Y, Fan ST. Gene expression profiling of liver cancer stem cells by RNA-sequencing. PLoS ONE 2012; 7 (5): e37159
11. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010; 26(1): 139–140
12. Ho DW, Ng IO. uGPA: unified Gene Pathway Analyzer package for high-throughput genome-wide screening data provides mechanistic overview on human diseases. Clin Chim Acta 2015; 441: 105–108
13. Hong G, Zhang W, Li H, Shen X, Guo Z. Separate enrichment analysis of pathways for up- and downregulated genes. J R Soc Interface 2014; 11(92): 20130950
14. Aytes A, Mitrofanova A, Lefebvre C, Alvarez MJ, Castillo-Martin M, Zheng T, Eastham JA, Gopalan A, Pienta KJ, Shen MM, Califano A, Abate-Shen C. Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. Cancer Cell 2014; 25(5): 638–651
15. Cummings RD, McEver RP. C-type lectins. In: Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Bertozzi CR, Hart GW, Eltzler ME. Essentials of Glycobiology. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 2009
16. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Björling L, Ponten F. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol 2010; 28(12): 1248–1250
17. Bale TL, Giordano FJ, Hickey RP, Huang Y, Nath AK, Peterson KL, Vale WW, Lee KF. Corticotropin-releasing factor receptor 2 is a tonic suppressor of vascularization. Proc Natl Acad Sci USA 2002; 99(11): 7734–7739
18. Graziani G, Tentori L, Portarena I, Barbarino M, Tringali G, Pozzoli G, Navarra P. CRH inhibits cell growth of human endometrial adenocarcinoma cells via CRH-receptor 1-mediated activation of cAMP-PKA pathway. Endocrinology 2002; 143(3): 807–813
19. Hao Z, Huang Y, Cleman J, Jovin IS, Vale WW, Bale TL, Giordano FJ. Urocortin2 inhibits tumor growth via effects on vascularization and cell proliferation. Proc Natl Acad Sci USA 2008; 105(10): 3939–3944
20. Wang J, Xu Y, Xu Y, Zhu H, Zhang R, Zhang G, Li S. Urocortin's inhibition of tumor growth and angiogenesis in hepatocellular carcinoma via corticotrophin-releasing factor receptor 2. Cancer Invest 2008; 26(4): 359–368
21. Tezval H, Atschekzei F, Peters I, Waalkes S, Hennenlotter J, Stenzl A, Becker JU, Merseburger AS, Kuczyk MA, Serth J. Reduced mRNA expression level of corticotropin-releasing hormone-binding protein is associated with aggressive human kidney cancer. BMC Cancer 2013; 13(1): 199
22. Graña X, Reddy EP. Cell cycle control in mammalian cells: role of cyclins, cyclin dependent kinases (CDKs), growth suppressor genes and cyclin-dependent kinase inhibitors (CKIs). Oncogene 1995; 11 (2): 211–219
23. Lew DJ, Kornbluth S. Regulatory roles of cyclin dependent kinase phosphorylation in cell cycle control. Curr Opin Cell Biol 1996; 8 (6): 795–804
24. Nebert DW, Russell DW. Clinical importance of the cytochromes P450. Lancet 2002; 360(9340): 1155–1162
25. Denison MS, Whitlock JP Jr. Xenobiotic-inducible transcription of cytochrome P450 genes. J Biol Chem 1995; 270(31): 18175–18178
26. Guengerich FP. Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. Chem Res Toxicol 2001; 14(6): 611–650
27. Burkholder B, Huang RY, Burgess R, Luo S, Jones VS, Zhang W, Lv ZQ, Gao CY, Wang BL, Zhang YM, Huang RP. Tumor-induced perturbations of cytokines and immune cell networks. Biochim Biophys Acta 2014; 1845(2): 182–201
28. Lacy P, Stow JL. Cytokine release from innate immune cells: association with diverse membrane trafficking pathways. Blood 2011; 118(1): 9–18
29. Lippitz BE. Cytokine patterns in patients with cancer: a systematic review. Lancet Oncol 2013; 14(6): e218–e228