<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>TCGA whole-transcriptome sequencing data reveals significantly dysregulated genes and signaling pathways in hepatocellular carcinoma</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Ho, DWH; Kai, AKL; Ng, IOL</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Frontiers of Medicine, 2015, v. 9, n. 3, p. 322-330</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2015</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/212083">http://hdl.handle.net/10722/212083</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.</td>
</tr>
</tbody>
</table>
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

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2015

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.
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 \( \log_2(FC) \geq 2, \log_2(CPM) \geq 1 \), 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) \( \log_2(FC) = 3.64, \text{FDR} = 5.32E^-78 \) and CLEC4G (C-type lectin domain family 4, member G) \( \log_2(FC) = 8.96, \text{FDR} = \text{1.19E}^-80 \) 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...
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 various 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-N 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
Successful qPCR validation of top-listing DE genes.

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

<table>
<thead>
<tr>
<th>Gene set</th>
<th># of genes</th>
<th># of DE genes</th>
<th>FDR</th>
<th>DE genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO biological process</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MITOTIC_CELL_CYCLE</td>
<td>153</td>
<td>27</td>
<td>6.91E–33 5.70E–30</td>
<td>PRC1, PKMYT1, AURKA, CDKN2A, CDKN2C, CCNA2, CDC5, MAD2L1, ZWINT, NEK2, ANLN, NDC80, PLK1, E2F1, KIF23, KIF2C, DLGAP5, CDC6, KIF11, UBE2C, BUB1B, NCAPH, BUB1, CENPF, BIRC5, CENPE, CDKN3</td>
</tr>
<tr>
<td>CELL_CYCLE_PROCESS</td>
<td>193</td>
<td>28</td>
<td>1.48E–31 6.09E–29</td>
<td>AURKA, CDC5, 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</td>
</tr>
<tr>
<td>CELL_CYCLE.GO_0007049</td>
<td>315</td>
<td>32</td>
<td>6.90E–31 1.90E–28</td>
<td>AURKA, CCNA2, CDC5, 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</td>
</tr>
<tr>
<td>CELL_CYCLE_PHASE</td>
<td>170</td>
<td>25</td>
<td>2.02E–28 4.17E–26</td>
<td>PRC1, PKMYT1, AURKA, CDKN2A, CDKN2C, CDC5, CCNA2, MAD2L1, ZWINT, NEK2, ANLN, NDC80, PLK1, E2F1, KIF23, KIF2C, DLGAP5, CDC6, KIF11, UBE2C, BUB1B, NCAPH, BUB1, CENPF, BIRC5, CENPE, CDKN3</td>
</tr>
<tr>
<td>M_PHASE_OF_MITOTIC_CELL_CYCLE</td>
<td>85</td>
<td>19</td>
<td>1.49E–25 2.45E–23</td>
<td>PRC1, PKMYT1, AURKA, KIF2C, DLGAP5, CDC5, CCNA2, KIF11, UBE2C, MAD2L1, ZWINT, BUB1B, NEK2, ANLN, NCAPH, BUB1, BIRC5, NDC80, CENPE, PLK1</td>
</tr>
<tr>
<td>Gene set</td>
<td># of genes</td>
<td># of DE genes</td>
<td>P value</td>
<td>FDR</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------------</td>
<td>---------------</td>
<td>---------------</td>
<td>-----------</td>
</tr>
<tr>
<td><strong>GO cellular component</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NON_MEMBRANE_BOUND_ORGANELLE</td>
<td>631</td>
<td>29</td>
<td>2.11E–18</td>
<td>2.46E–16</td>
</tr>
<tr>
<td>INTRACELLULAR_NON_MEMBRANE_BOUND_ORGANELLE</td>
<td>631</td>
<td>29</td>
<td>2.11E–18</td>
<td>2.46E–16</td>
</tr>
<tr>
<td>MICROTUBULE_CYTOSKELETON</td>
<td>152</td>
<td>17</td>
<td>1.36E–17</td>
<td>1.05E–15</td>
</tr>
<tr>
<td>SPINDLE</td>
<td>39</td>
<td>11</td>
<td>1.51E–16</td>
<td>8.81E–15</td>
</tr>
<tr>
<td>CYTOSKELETAL_PART</td>
<td>235</td>
<td>18</td>
<td>1.29E–15</td>
<td>6.03E–14</td>
</tr>
<tr>
<td><strong>GO molecular function</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MICROTUBULE_MOTOR_ACTIVITY</td>
<td>16</td>
<td>5</td>
<td>1.96E–08</td>
<td>7.77E–06</td>
</tr>
<tr>
<td>MOTOR_ACTIVITY</td>
<td>28</td>
<td>5</td>
<td>4.19E–07</td>
<td>8.29E–05</td>
</tr>
<tr>
<td>CYTOSKELETAL_PROTEIN_BINDING</td>
<td>159</td>
<td>7</td>
<td>2.97E–05</td>
<td>0.004</td>
</tr>
<tr>
<td>CARBOHYDRATE_BINDING</td>
<td>72</td>
<td>5</td>
<td>4.90E–05</td>
<td>0.005</td>
</tr>
<tr>
<td>PROTEIN_KINASE_REGULATOR_ACTIVITY</td>
<td>39</td>
<td>4</td>
<td>6.21E–05</td>
<td>0.005</td>
</tr>
<tr>
<td><strong>KEGG pathway</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEGG_CELL_CYCLE</td>
<td>128</td>
<td>19</td>
<td>6.25E–22</td>
<td>1.16E–19</td>
</tr>
<tr>
<td>KEGG_OOCYTE_MEIOSIS</td>
<td>114</td>
<td>11</td>
<td>4.06E–11</td>
<td>3.78E–09</td>
</tr>
<tr>
<td>KEGG_PROGESTERONE_MEDIATED_OOCYTE_MATURATION</td>
<td>86</td>
<td>8</td>
<td>2.62E–08</td>
<td>1.63E–06</td>
</tr>
<tr>
<td>KEGG_P53_SIGNALING_PATHWAY</td>
<td>69</td>
<td>7</td>
<td>1.09E–07</td>
<td>5.06E–06</td>
</tr>
<tr>
<td>KEGG_TIGHT_JUNCTION</td>
<td>134</td>
<td>4</td>
<td>0.006</td>
<td>0.226</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Gene set</th>
<th># of genes</th>
<th># of DE genes</th>
<th>P value</th>
<th>FDR</th>
<th>DE genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GO biological process</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIGNAL_TRANSDUCTION</td>
<td>1634</td>
<td>70</td>
<td>9.98E–19</td>
<td>8.24E–16</td>
<td>DIRAS3, IGF1, IGF2, HPGD, GNA14, MARCO, CCL3, ADRA2B, TRX2R, ADRA1B, ADRA1A, WNK2, IIL1RL1, NR1I2, CXCL6, GRIAR, FCGFB2, GABBR3, CAMK2B, BCL2L10, AVPR1A, TRPV4, CCL19, NPY1R, APOA1, NR4A3, ESR1, CHRNA4, LILRB5, PDGFRA, ECM1, TNFRSF11B, GADD45G, GADD45B, BESC1B, IL18R1, SOCS2, SOCS3, PT1HR, CTND2, PTPRD, CRBBP, LIFR, CEACAM6, FPR1, CXCL12, NR4A1, CXCL14, SKAP1, WISP2, MGC2, ROK2, TACSTD2, EPHRA2, NTRK2, TGFalpha, CHL1, LYY6, VIP1R1, CD79A, IGF1BP1, PRKAR2B, RET, RAC1, ANXA3, SFRP5, SFRP1, GCGR, IGEALS, DTX1</td>
</tr>
<tr>
<td>Gene set</td>
<td># of genes</td>
<td># of DE genes</td>
<td>P value</td>
<td>FDR</td>
<td>DE genes</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------------</td>
<td>---------------</td>
<td>---------</td>
<td>-------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>RESPONSE TO EXTERNAL STIMULUS</td>
<td>312</td>
<td>25</td>
<td>5.02E-13</td>
<td>2.07E-10</td>
<td>SI0648, SERPINE1, F9, SA41, IL1RAP, TRPV4, CXCL2, FPR1, CXCL6, GCGR, CCL21, LECT2, CCL19, CD1D, ORM1, SELE, CXCL12, PGLYRP2, CCL3, ALB, LYVE1, CXCL14, CXCL13, FOS, MBL2</td>
</tr>
<tr>
<td>CARBOXYLIC ACID METABOLIC_PROCESS</td>
<td>178</td>
<td>19</td>
<td>2.13E-12</td>
<td>5.38E-10</td>
<td>SLC7A8, BBOX1, ASPA, GGT5, CYP39A1, ACOT12, GST21, SLC3A1, HAO2, AKR1D1, SLC27A5, GLS2, FTD0, IGFI, CYP4A11, GLYAT, GCK, HPGD</td>
</tr>
<tr>
<td>ORGANIC ACID METABOLIC_PROCESS</td>
<td>180</td>
<td>19</td>
<td>2.61E-12</td>
<td>5.38E-10</td>
<td>SLC7A8, BBOX1, ASPA, GGT5, CYP39A1, ACOT12, GST21, SLC3A1, HAO2, AKR1D1, SLC27A5, GLS2, FTD0, IGFI, CYP4A11, GLYAT, GCK, HPGD</td>
</tr>
<tr>
<td>LIPID METABOLIC_PROCESS</td>
<td>325</td>
<td>24</td>
<td>8.36E-12</td>
<td>1.38E-09</td>
<td>CYP3A4, ALDH8A1, LCAT, APOE, PTTPNM3, NR112, CYP39A1, CETP, THRSR, HAO2, SLC27A5, HPGD, APOA4, APOAI, BC02, ACOT12, NPC11, IP6K3, AKR1D1, CYP4A11, GLYAT, RDH16, UGT2B7, SMPD3</td>
</tr>
<tr>
<td>GO cellular component</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEMBRANE</td>
<td>1994</td>
<td>90</td>
<td>1.23E-25</td>
<td>2.87E-23</td>
<td>IL1RAP, GHR, SLC22A1, CFT, SLC3A1, CRA, GNA14, MARCO, ADRA2B, LYVE1, TBX2A2, ADRAB1, ADRR1A, TREH, C9, CD4, GPR128, ABCB11, STEAP4, CD163, GRIA3, CD1D, SLC16A4, GABBR3, NAPS8, CNGA1, PLEKHB1, PTTPR, AVPR1A, UNC93A, TRPV4, SELP, NPY1R, SELE, SLC3A4, RNAM1, CNTRFR, MRC1, HS3ST3A1, MAN1C1, SLC01B3, CHRNA4, CLEC4M, PDGFRF, HSFST3B1, CLEC1B, IL18R1, PKHD1, PTHHR, RBG, CR1, PTTRD, LIFR, SRPX, CEACAM6, C7, C9, FPR1, STAB2, SLC13A2, CLDN2, PRSS8, GGT5, EPCAM, TACSTD2, MEME, EP2, NTRK2, CHL1, LVE6, PROMI, VPR1, FXDI1, ITGB8, ITGA9, CD79A, SLC2A5, NGFR, PTTPNM3, CDHR2, SLC7A8, KCND3, B3GAT1, SIGLEC7, VSIG2, CLDN10, GCGR, SLC6A2, BAP31, SLC10A1</td>
</tr>
<tr>
<td>PLASMA_MEMBRANE</td>
<td>1426</td>
<td>75</td>
<td>2.65E-25</td>
<td>3.09E-23</td>
<td>LIFR, IL1RAP, CEACAM6, SLC22A1, GHR, CFT, SLC3A1, C9, FPR1, STAB2, GNA14, MARCO, ADRA2B, LYVE1, ADRAB1, TBX2A2, SLC13A2, ADRA1A, TREH, CLDN2, PRSS8, CD4, ABCB11, EPCAM, TACSTD2, MEME, EP2A, NTRK2, STEAP4, CD163, LVE6, GRIA3, CD1D, PROMI, SLC16A4, GABBR3, VPR1, CNGA1, FYDI1, ITGB8, ITGA9, CD79A, SLC2A5, NGFR, PTTPR, AVPR1A, UNC93A, TRPV4, SELP, NPY1R, SELE, SLC7A8, KCND3, SLC3A4, NCAM1, MRC1, SIGLEC7, SLC01B3, CHRNA4, CLEC4M, VSIG2, PDGFRF, HSS1ST3B1, CLEC1B, CLDN10, GCGR, IL18R1, SLC6A2, BAP31, PKHD1, PTHHR, RBG, CR1, PTTRD, SLC10A1</td>
</tr>
<tr>
<td>INTRINSIC TO MEMBRANE</td>
<td>1348</td>
<td>72</td>
<td>1.09E-24</td>
<td>8.46E-23</td>
<td>LIFR, IL1RAP, CEACAM6, SLC22A1, GHR, SLC3A1, C9, FPR1, STAB2, MARCO, ADRAB2, LYVE1, TBX2A2, ADRA1B, ADRA1A, TREH, C9, CD4, GPR128, ABCB11, EPCAM, TACSTD2, MEME, EP2A, NTRK2, STEAP4, CD163, LVE6, GRIA3, CD1D, PROMI, SLC16A4, GABBR3, VPR1, CNGA1, FYDI1, ITGB8, ITGA9, CD79A, SLC2A5, NGFR, PTTPR, AVPR1A, UNC93A, TRPV4, SELP, NPY1R, SELE, SLC7A8, KCND3, SLC3A4, NCAM1, MRC1, SIGLEC7, SLC01B3, CHRNA4, CLEC4M, VSIG2, PDGFRF, HSS1ST3B1, CLEC1B, CLDN10, GCGR, IL18R1, SLC6A2, BAP31, PKHD1, PTHHR, RBG, CR1, PTTRD, SLC10A1</td>
</tr>
<tr>
<td>INTEGRAL TO MEMBRANE</td>
<td>1330</td>
<td>70</td>
<td>1.21E-23</td>
<td>7.03E-22</td>
<td>LIFR, IL1RAP, CEACAM6, GHR, SLC22A1, GHR, SLC3A1, C9, FPR1, STAB2, MARCO, ADRAB2, LYVE1, TBX2A2, ADRA1B, ADRA1A, TREH, C9, GPR128, ABCB11, GGT5, TACSTD2, MEME, EP2A, NTRK2, CD163, CHL1, LVE6, CD1D, PROMI, SLC16A4, GABBR3, FOREIGN1, CNGA1, FYDI1, ITGB8, ITGA9, SLC2A5, PLEKHB1, NGFR, PTTPR, AVPR1A, UNC93A, TRPV4, SELP, NPY1R, SELE, SLC7A8, KCND3, SLC3A4, NCAM1, MRC1, B3GAT1, SIGLEC7, HSS1ST3A1, MAN1C1, SLC01B3, CHRNA4, CLEC4M, VSIG2, PDGFRF, HSS1ST3B1, CLEC1B, GCGR, SLC6A2, PTHHR, RBG, CR1, PTTRD, SLC10A1</td>
</tr>
</tbody>
</table>
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,

### GO molecular function

<table>
<thead>
<tr>
<th>Gene set</th>
<th># of genes</th>
<th># of DE genes</th>
<th>P value</th>
<th>FDR</th>
<th>DE genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEMBRANE_PART</td>
<td>1670</td>
<td>78</td>
<td>4.27E–23</td>
<td>1.99E–21</td>
<td>IL1RAP, SLC22A1, GHR, CFTF, SLC3A1, C8A, GNA14, MARCO, ADR2B, LYVE1, TXA2R, ADR1B, ADR1A, TREH, C49, GPR128, ABCB11, CD163, CD1D, SLC16A4, GABRB3, CNGA1, PLEKH1B, PTPRS, AVPR1A, SELP, NPY1R, SLC34A2, NCAM1, CNTRF, MRC1, HSST3A1, MAN1C1, SLC1O1B3, CHRNA4, CLEC4M, PDGFRa, HSST3B1, CLEC1B, PKHD1, PTH1R, RBG1, CR1, CTPRBD, 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, VSG12, CLDN10, GCGR, SLC6A2, SLC10A1</td>
</tr>
<tr>
<td>OXYGEN_BINDING</td>
<td>22</td>
<td>12</td>
<td>9.98E–18</td>
<td>3.95E–15</td>
<td>CYP3A4, CYP3A7, CYP1A1, CYP2C9, CYP26A1, CYP1A2, CYP2E1, ALB, CYP2A6, CYP2A7, CYP8B1, HBB</td>
</tr>
<tr>
<td>RECEPTOR_ACTIVITY</td>
<td>583</td>
<td>34</td>
<td>3.28E–13</td>
<td>6.50E–11</td>
<td>LIFR, PTPRS, AVPR1A, IL13A2, GHR, FPR1, STAB2, HPGD, MARCO, PGLYRP2, RET, CNTRF, ADR2B, GABRP, MRC1, NRA3, ADR1B, TXA2R, ADR1A9, SIGLEC7, MCC, CC4, GP1R128, CHRNA4, LILRB5, TACSTD2, PDGFRa, TNFRSF1B, CLEC1B, GCGR, GABRB3, VIPR1, PTH1R, PTPRD</td>
</tr>
<tr>
<td>OXIDOREDUCTASE_ACTIVITY</td>
<td>289</td>
<td>24</td>
<td>6.83E–13</td>
<td>9.02E–11</td>
<td>CYP1A4, ALDH8A1, BBOX1, CYP26A1, CYP1A2, GPD1, CYP8B1, CYP39A1, GSTZ1, HAO2, PHGDH, HPGD, ADH1B, AKR7A3, KMO, BCO2, TDO2, SRD5A2, ACADL, CYP2A6, ADH6, ADH4, CYP4A11, RDH16</td>
</tr>
<tr>
<td>TRANSMEMBRANE_RECEPTOR_ACTIVITY</td>
<td>418</td>
<td>25</td>
<td>2.78E–10</td>
<td>2.75E–08</td>
<td>LIFR, PTPRS, AVPR1A, IL13A2, GHR, FPR1, STAB2, HPGD, CNTRF, ADR2B, GABRP, ADR1B, TXA2R, ADR1A9, CD4, GP1R128, CHRNA4, LILRB5, PDGFRa, CLEC1B, GCGR, GABRB3, VIPR1, PTH1R, PTPRD</td>
</tr>
<tr>
<td>RECEPTOR_BINDING</td>
<td>377</td>
<td>23</td>
<td>9.87E–10</td>
<td>7.82E–08</td>
<td>APOE, PITPM3, IL1RN, TNFRSF1B, SAA1, TGF6, CXCL2, CXCL6, CCL21, CCL19, IFI1, IFI2, CCL12, APOA1, SOC2, CCL3, ANPTL1, CXCL14, CXCL13, MBL2, ADAMTS13, TNXB, DXT1</td>
</tr>
<tr>
<td>KEGG pathway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEGG_RETINOL_METABOLISM</td>
<td>64</td>
<td>25</td>
<td>6.92E–31</td>
<td>1.29E–28</td>
<td>CYP4A11, CYP3A4, ADH1B, ADH1C, ADH4, ADH1A, CYP26A1, CYP2C9, CYP2C19, CYP2C8, CYP2B6, CYP2A13, UGT1A4, CYP3A7, LRAT, CYPA26, CYPA27, CYPA422, CYP1A1, CYP1A2, ADH6, UGT2A1, CYP3A43, RDH16, UGT2B7</td>
</tr>
<tr>
<td>KEGG_DRUG_METABOLISM_CYTOCHROME_P450</td>
<td>72</td>
<td>23</td>
<td>4.00E–26</td>
<td>3.72E–24</td>
<td>CYP2E1, CYP3A44, GSTZ1, ADH1B, ADH1C, ADH4, ADH1A, CYP2C9, CYP2C19, CYP2C8, CYP2B6, CYP2A13, UGT1A4, CYP3A7, GSTM5, GSTA2, CYP2A6, CYP2A7, CYP1A2, ADH6, UGT2A1, CYP3A43, UGT2B7</td>
</tr>
<tr>
<td>KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450</td>
<td>70</td>
<td>21</td>
<td>2.63E–23</td>
<td>1.63E–21</td>
<td>CYP2E1, CYP3A44, GSTZ1, ADH1B, ADH1C, ADH4, ADH1A, CYP2C9, CYP2C19, CYP2C8, CYP2B6, UGT1A4, CYP3A7, GSTM5, GSTA2, CYP1A1, CYP1A2, ADH6, UGT2A1, CYP3A43, UGT2B7</td>
</tr>
<tr>
<td>KEGG_DRUG_METABOLISM_OTHER_ENZYMES</td>
<td>51</td>
<td>12</td>
<td>1.75E–12</td>
<td>8.13E–11</td>
<td>CYP1A4, UPP2, CDA, CYP2A13, UGT1A4, CYP3A7, CYP2A6, CYP2A7, CYPA21A4, CYPA43, Nat2, UGT2B7</td>
</tr>
<tr>
<td>KEGG_LINOLEIC_ACID_METABOLISM</td>
<td>29</td>
<td>9</td>
<td>6.96E–11</td>
<td>2.59E–09</td>
<td>CYP2E1, CYP3A44, CYP1A2, PLA2G2A, CYP3A43, CYP2C9, CYP2C19, CYP2C8, CYP3A7</td>
</tr>
</tbody>
</table>

DE genes. A number of DE genes that were reported in previous studies [8,9], such as ALG11, 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,
FOXMI 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 FOXMI 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 FOXMI 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 FOXMI 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 cooperation 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, CDK2A, CDK2B, and CDK3) (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 is 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