Genome-wide Transcriptional Sequencing Identifies Novel Mutations in Metabolic Genes in Human Hepatocellular Carcinoma

DAOUD M. MEERZAMAN[1,2], CHUNHUA YAN[1], QING-RONG CHEN[1], MICHAEL N. EDMONSON[1], CARL F. SCHAEFER[1], ROBERT J. CLIFFORD[2], BARBARA K. DUNN[3], LI DONG[2], RICHARD P. FINNEY[1], CONSTANCE M. CULTRARO[2], YING HU1, ZHIHUI YANG[2], CU V. NGUYEN[1], JENNY M. KELLEY[2], SHUANG CAI[2], HONGEN ZHANG[2], JINGHUI ZHANG[1,4], REBECCA WILSON[2], LAUREN MESSMER[2], YOUNG-HWA CHUNG[5], JEONG A. KIM[5], NEUNG HWA PARK[6], MYUNG-SOO LYU[6], IL HAN SONG[7], GEORGE KOMATSOULIS[1] and KENNETH H. BUETOW[1,2]

[1]*Center for Bioinformatics and Information Technology, National Cancer Institute, Rockville, MD, U.S.A.;*
[2]*Laboratory of Population Genetics, National Cancer Institute, National Cancer Institute, Bethesda, MD, U.S.A.;*
[3]*Basic Prevention Science Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, U.S.A;*
[4]*Department of Biotechnology/Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, U.S.A.;*
[5]*Department of Internal Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea;*
[6]*Department of Internal Medicine, University of Ulsan College of Medicine, Ulsan University Hospital, Ulsan, Korea;*
[7]*Department of Internal Medicine, College of Medicine, Dankook University, Cheon-An, Korea*

Abstract. *We report on next-generation transcriptome sequencing results of three human hepatocellular carcinoma tumor/tumor-adjacent pairs. This analysis robustly examined ~12,000 genes for both expression differences and molecular alterations. We observed 4,513 and 1,182 genes demonstrating 2-fold or greater increase or decrease in expression relative to their normal, respectively. Network analysis of expression data identified the Aurora B signaling, FOXM1 transcription factor network and Wnt signaling pathways pairs being altered in HCC. We validated as differential gene expression findings in a large data set containing of 434 liver normal/tumor sample pairs. In addition to known driver mutations in TP53 and CTNNB1, our mutation analysis identified non-synonymous mutations in genes implicated in metabolic diseases, i.e. diabetes and obesity: IRS1, HMGCS1, ATP8B1, PRMT6 and CLU, suggesting a common molecular etiology for HCC of alternative pathogenic origin.*

Correspondence to: Daoud M. Meerzaman, Center for Biomedical Informatics & Information Technology, 9609 Medical Center Drive, 1W466, Rockville, MD 20850, U.S.A. Tel: +1 2402765205, Fax: +1 2402767886, e-mail: meerzamd@mail.nih.gov

Key Words: Hepatocellular carcinoma (HCC), RNA-seq, gene expression, mutation.

Worldwide, liver cancer is the fifth most common cancer and the third most common cause of cancer-related mortality (1). Over 75% of liver cancers are hepatocellular carcinomas (HCCs), which are adenocarcinomas that occur in the context of cirrhosis in 60% to 85% of cases (2). Although over 80% of HCC cases occur in developing countries in Asia and Africa, incidence is on the rise in developed countries as well (3-4). This increase has been equally attributed to a rise in HCV infection and to the worldwide increase in the number of people with diabetes and obesity (4).

Although the association of HCC with chronic liver disease is well-established, typically due to liver cirrhosis resulting from HBV/HCV infection and/or other liver disease, the exact developmental etiology of HCC remains undefined. In developed countries chronic liver disease is most commonly due to non-alcoholic fatty liver disease (NAFLD), which encompasses a spectrum of liver disorders, ranging histopathologically from the milder hepatic steatosis/isolated fatty liver to the more aggressive non-alcoholic steatohepatitis (NASH). It is NASH that may progress to liver cirrhosis, ultimately leading to further complications such as hepatic failure and HCC (3, 5-7).

Next-generation sequencing (NGS) of the complete RNA transcriptome (RNA-seq) offers a novel approach for systematically characterizing the underlying molecular etiology of HCC. RNA-seq not only permits for the accurate

Table I. *Summary of mapping data generated by RNA-seq.*

Sample	Total reads	Mapped reads		RefSeq exon junction reads		Non-RefSeq exon junction reads		EST exon junction reads		Genes (reads ≥1)	Genes (rpkm ≥1)
PHC98023	35 M	28 M	81%	7 M	19%	28 K	0.08%	561 K	1.59%	17,471	11,087
PHC98024	38 M	30 M	80%	9 M	24%	47 K	0.12%	520 K	1.37%	17,383	11,421
PHC98025	35 M	31 M	88%	9 M	26%	32 K	0.09%	439 K	1.26%	16,981	10,415
PHC98026	39 M	35 M	89%	9 M	24%	48 K	0.12%	416 K	1.07%	17,335	10,819
PHC98027	41 M	36 M	86%	10 M	23%	48 K	0.12%	866 K	2.09%	17,188	10,987
PHC98028	41 M	36 M	89%	10 M	24%	50 K	0.12%	648 K	1.59%	18,420	12,481

Column 1: sample identification number; column 2: total sequencing reads for each sample; column 3: number of mapped reads and percentage of total reads that were mapped; columns 4, 5, 6: number of reads and percentage of total reads that cross an exon junction and map to RefSeq exons (column 4), non-RefSeq exons (column 5), and EST exon junctions (column 6); column 7: number of genes that are represented by ≥1 read; column 8: number of genes normalized by RPKM ≥1. RPKM, reads per kilobase of exon model per million mapped reads.

measurement of transcript expression levels, but also reliably identifies and assesses rare RNA species, alternative splicing, base substitutions, insertions and deletions (indels), allele-specific expression and RNA editing (8-10). In the present study, we report on a survey of the complete transcriptome landscape and the mutation profile of hepatocellular carcinoma using RNA-seq in three pairs of HCC tumor and tumor-adjacent tissue samples. Key findings observed in the present survey are examined in larger collections of tumor and normal samples in order to establish robust evidence of prevalence and validity.

Materials and Methods

Tissue samples. All tissue samples in the present study were collected anonymously at the Asan Medical Center in South Korea during 1998-2001. The study design adhered to all NIH standards for human subject research and was approved by the NIH office of Human Subject Research. RNA was isolated from three HCC tumor (PHC9824, PHC9826 and PHC9828) and adjacent normal samples (PHC9823, PHC9825 and PHC9827). Two of the normal/tumor pairs (PHC9823/24 and PHC9827/28) were HCV-infected; the third pair (PHC9825/26) was HCV/HBV-negative.

RNA Sequencing Method. Total RNA was isolated from three HCC tumor and normal adjacent tissue pairs using the RNeasy kit (Qiagen) according to the manufacturer's instructions. Double-stranded cDNA was generated using random hexamer primers and reverse transcriptase. Sequencing adaptors were ligated using the Illumina Genomic DNA sample prep kit. Fragments (approximately 340-bp long) were isolated by gel electrophoresis and amplified by limited cycles of PCR. Large-scale deep sequencing was carried out on the Illumina Solexa Genome Analyzer, as described in the Illumina mRNA expression analysis protocol (http://www.illumina.com).

Mapping of RNA-seq reads to the genome. Mapping of RNA deep sequencing reads from three pairs of liver tumor and matched normal tissues was carried out using Burrows-Wheeler Aligner (BWA: http://bio-bwa.sourceforge.net/). Reads of 75mer were mapped to four reference databases: hg18 (NCBI build 36), refFlat, alternative-splicing exons and ESTs. Our alternative-splicing database consists of all sequential combinations of non-adjacent exons that are not found in the RefFlat database. EST exon sequences were extracted from the hg18 sequence using exon coordinates in the UCSC intronEst table.

Differentially expressed genes in RNA-seq. We have already reported on the normalized expression values as RPKM (reads per kilobase of exon model per million mapped reads) (8). For a given gene, reads were normalized against exon size and the number of reads sequenced to generate the RPKM value. Log2 ratios of RPKM values were calculated for each gene in paired tumor and adjacent normal samples.

To validate the differentially expressed genes identified from RNA-seq analysis, we downloaded a large HCC gene expression dataset from GEO repository with the accession number of GSE14520 GPL3921 (http://www.ncbi.nlm.nih.gov/geo) which contains 212 liver normal tissues and 222 liver tumor tissues (11). Profiling was performed on an Affymetrix HT HG-U133A platform and the data were normalized using Robust Multi-array Average (RMA) method and global median centering (11). We extracted the data for the genes overlapping with the ones shown on the differential expressed gene list. In the case of genes with multiple probe sets, the maximum gene expression was calculated. The t-statistics were used to compare the gene expression between liver normal and tumor tissues.

Single-nucleotide variant (mutation) calls. Somatic non-silent mutations (missense, non-sense, frameshift) were identified by our in-house software, Bambino (12) and by VarScan (version 2.2, Wash U) (13) from BAM files for each paired tumor and adjacent normal samples. Two criteria were applied: 1) Mutations from Bambino with a minimum of 4× coverage, greater than 10% alternative allele frequency in tumor samples and less than 1% alternative allele frequency in normal samples; and 2) mutations from Bambino which did not meet the previous criteria but showed a p-value <0.05 in VarScan. All somatic non-silent mutations were manually reviewed by examining the alignment in Bambino alignment view.

Mutation validation. Sanger sequencing was applied to PCR products generated using both cDNA and genomic DNA templates. We validated the candidate variants in cDNA and genomic DNA

with primers designed by the primer3 program (14). Sanger sequencing was performed on the PCR products. Each chromatogram was base-called with Phred using the option designed to generate a polymorphism file, which details secondary and alternate base call information (15).

Prediction of AA substitution using logE and SIFT. LogE (16) and SIFT (17) are used to predict the AA substitution. A logE score whose absolute value is greater than or equal to 1 indicates that the amino acid alteration is likely to affect protein (16). SIFT value ≤0.05 is predicted to be deleterious (17).

Pathway analysis of differentially expressed and validated mutated genes. We analyzed the differentially expressed genes and mutated genes by identifying over-represented pathways and by constructing gene-gene networks. Differentially expressed genes were mapped to pathways in the Pathway Interaction Database (PID) (18) and the significance of an individual pathway's being affected was computed using R's hyper-geometric distribution function, adjusted for multiple hypothesis testing using the Benjamini-Hochberg false discovery rate (19). Differentially expressed genes were also assembled into novel networks with direct interactions obtained from PID.

Analysis of copy number variation using Affymetrix SNP6.0 assay. The Affymetrix SNP6.0 assay was performed according to the manufacturer's instructions (Affymetrix, Santa Clara, CA, USA). Assay runs were performed in 96 well plates containing three tumor and normal adjacent samples, four Asian HapMap samples (NA18954, NA18971, NA18603 and NA18995), the Affymetrix103 control DNA and a negative control (H_2O). Data generated by SNP6.0 assays was analyzed with the Affymetrix Genotyping Console version 3.0 *birdseed* algorithm. Samples were analyzed for copy number variation using the Affymetrix Genotyping Console program with default parameters and the HapMap270 reference model. The resulting copy number log2ratio data served as input for the R *DNAcopy* package, which implements the circular binary segmentation (CBS) algorithm (20). We converted fractional CBS copy number values to discrete copy number states using the following thresholds: CBS copy number <=1.8 represents copy number loss; CBS copy number >=2.2 indicated copy number gain.

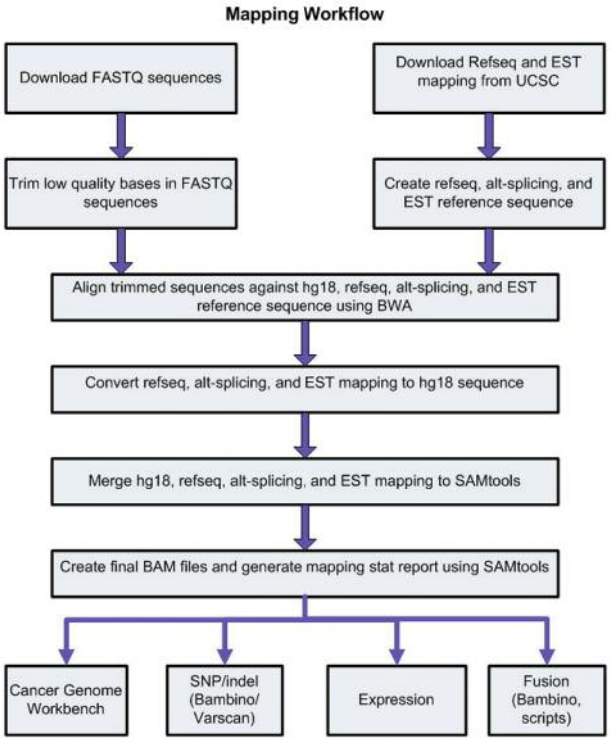

Figure 1. *Mapping Workflow. The above workflow diagram illustrates the sequence of steps involved in the alignment and mapping of reads to the human genome (hg18). Reads are trimmed and mapped to the RefSeq, EST, and hg18 databases. The best mapping for each read is selected to build the final BAM files in hg18 coordinates. UCSC: University of California Santa Cruz Genome Browser (http://genome.ucsc.edu); BWA: Burrows-Wheeler Alignment tool (http://bio-bwa.sourceforge.net/); SAMtools - utilities for manipulating alignments in the SAM/BAM format (http://samtools.sourceforge.net/); Bambino - an in-house single-nucleotide variation (snv) call program (https://cgwb.nci.nih.gov/); VarScan – a variant call program from Washington University (http://varscan.sourceforge.net/).*

Results

The complete workflow process for mapping the RNA-seq data from our HCC tumor samples (PHC98024, PHC98026 and PHC98028) is described in Figure 1.

Identification of differentially expressed genes by RNA-Seq. For expression analysis, the number of reads mapped to exon regions of each gene was calculated and normalized using the RPKM method (8). Our data revealed ~17,500 genes with at least one read and ~12,000 genes with RPKM ≥1 (Table I). These findings are consistent with previous reports that 10-12,000 genes are expressed in normal/cancerous liver tissue (21). Our initial analysis used RPKM >1 for tumor or normal and fold change between tumor and normal of >2 in at least 2 sample pairs. Applying these thresholds, we identified 4,513 up-regulated and 1,182 down-regulated genes in tumors compared to matched normal samples (Table II).

When the most stringent filter (RPKM ≥10 for tumor with RPKM ≤1 for normal and fold change ≥10) was applied to gene expression levels, a total of 58 genes were identified as up-regulated in at least two tumor-normal pairs (Table II and Figure 2). Out of these 58 genes, 26 have been previously identified as differentially expressed in HCC, while the remaining 32 have not been reported for HCC to date. Additionally, a total of 98 genes were down-regulated when similar strict criteria (RPKM ≤1 for tumor with RPKM ≥10 for normal and fold change ≥10) were applied (Table II and Figure 2). This sub-set of 58 up-regulated and 98 down-regulated genes were then subjected to additional analyses. To more robustly assess the prevalence of the differential expression of these 156 genes, we obtained a large HCC

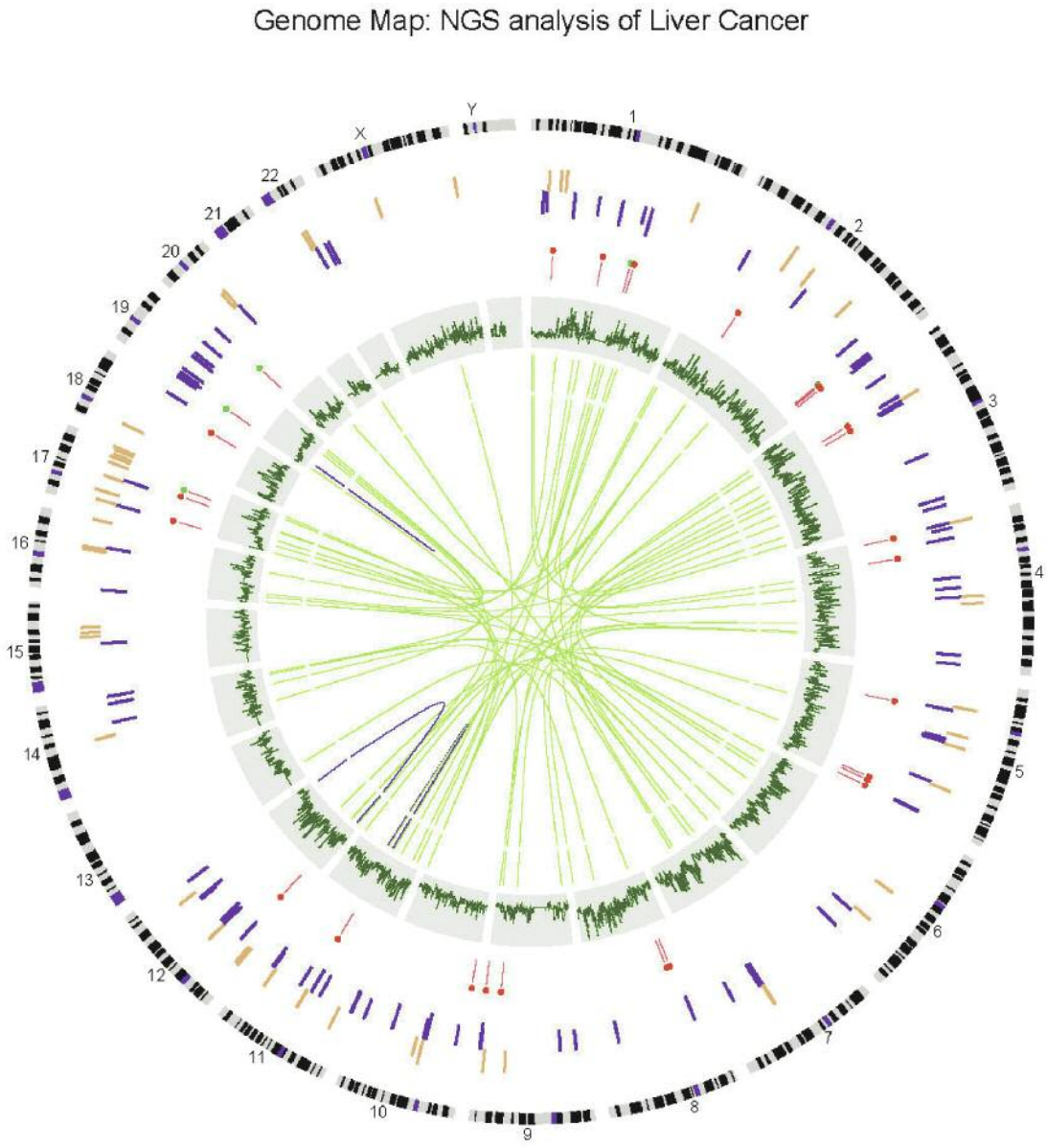

Figure 2. *Diagrammatic representation of expression levels, non-synonymous mutations, translocations, and copy number variations in the mRNA sequences of three HCC tissue samples. Chromosome ideograms are displayed along the outer ring (track 1); these are oriented pter-qter in a clockwise direction with centromeres indicated in blue. From outside to inside, the track just internal to the chromosome ideogram represents differential expression of genes at each locus (track 2). Orange bars indicate 58 up-regulated genes and blue bars indicate 98 down-regulated genes. Thirty non-synonymous mutations appear as red lines in the next track (track 3), copy number variation is indicated in the next track in dark green lines (track 4). The innermost track consists of light green lines depicting inter-chromosomal re-arrangements and blue lines depicting intra-chromosomal re-arrangements (track 5).*

gene expression dataset from the GEO repository (http://www.ncbi.nlm.nih.gov/geo) (11), containing 434 liver tissues with 212 normal and 222 tumor tissues. Only 118 of 156 differentially expressed genes were assayed in this dataset. Over 90% of the 118 genes had the same expression pattern as in the original RNA-seq data and showed statistically significant differential expression between normal liver and tumor tissues (FDR<0.01).

Identifying somatic mutations associated with HCC. The RNA-seq data revealed a total of 712,451 single-nucleotide variants (SNVs) present in tumor and/or normal tissues (Table III). We focused our subsequent analyses on the 149 non-synonymous somatic mutations. Among these, 124 have not been previously reported in HCC except for *CTNNB1* and *TP53*. Sixty-three of the mutations exhibited ≥30% frequency reads for the alternative allele in one

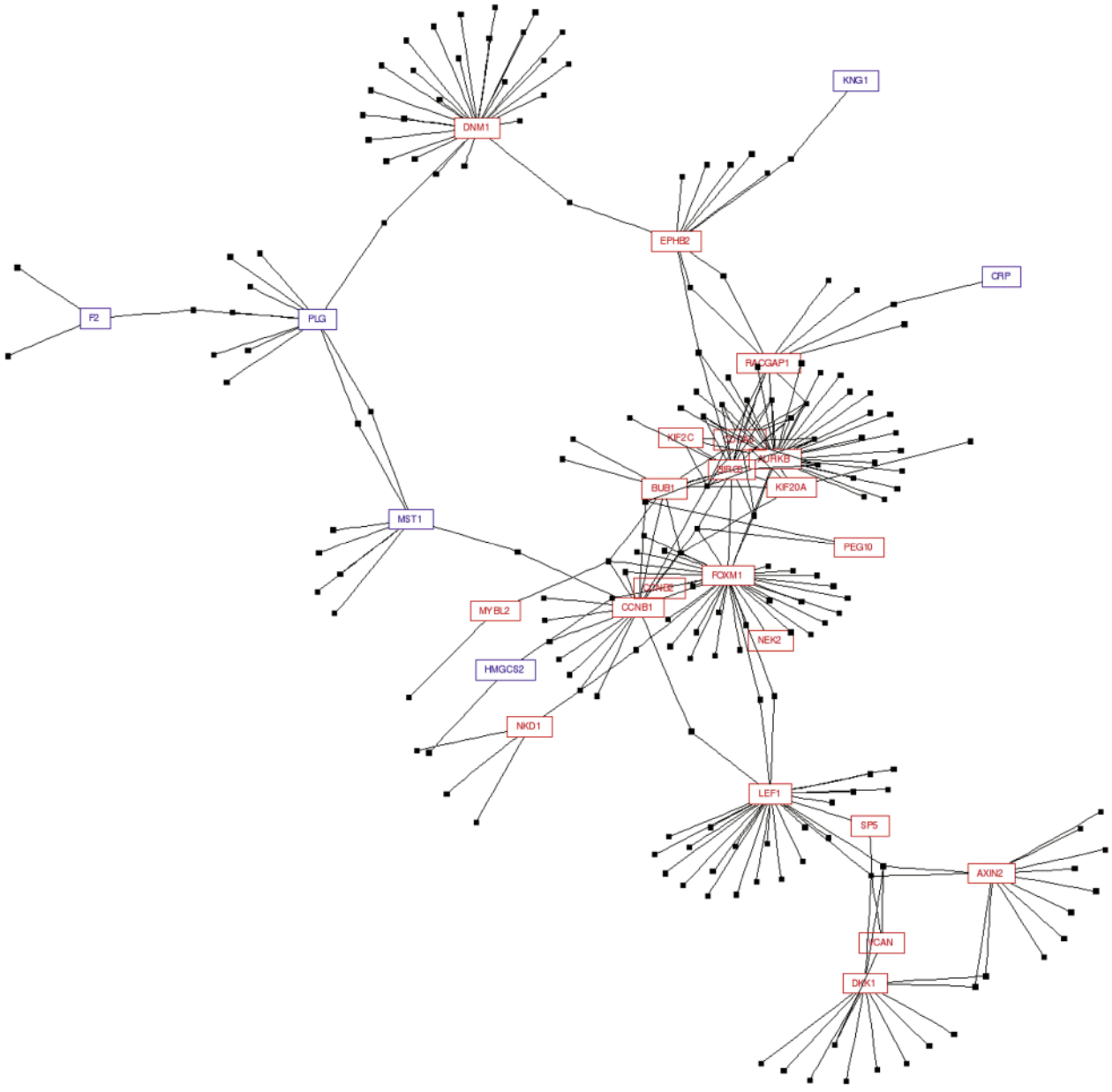

Figure 3. *Network diagram of the interaction of a subset of the 58 up-regulated and 98 down-regulated genes. Up-regulated genes are highlighted in red and the down-regulated genes are highlighted in blue. Twenty-one up-regulated and six down-regulated genes form a network. Nineteen of these 27 genes form a major gene-gene interaction network. Out of these 19 genes, eight genes (CCNB2, CDCA8, DNM1, KIF2C, NKD1, RACGAP1, VCAN, and MST1) have been previously reported in HCC; to date, the other 13 genes have not been documented in HCC. Each gene is connected by a solid line to each of the other genes with which it interacts. Theses interacting genes are displayed as small solid squares.*

tumor but not in the paired normal tissue. Among these 63 mutations, 30 mutations from 29 genes could be validated by Sanger sequencing as shown in Table III and Figure 2, track 4, red lines. Only *TP53* was mutated in two samples. To more robustly assess prevalence of mutations identified from this transcriptome-wide survey, we performed follow-up analysis of an additional 25 sets of paired HCC tumor samples and matched control tissues. This analysis revealed that approximately one-third of HCC samples were mutated in the *TP53* gene. Two of these *TP53*-mutated HCC samples were also mutated in the *CTNNB1* gene; one of these samples had 2 *TP53* mutations. No other genes were found to have recurrent mutations in the extended sample set.

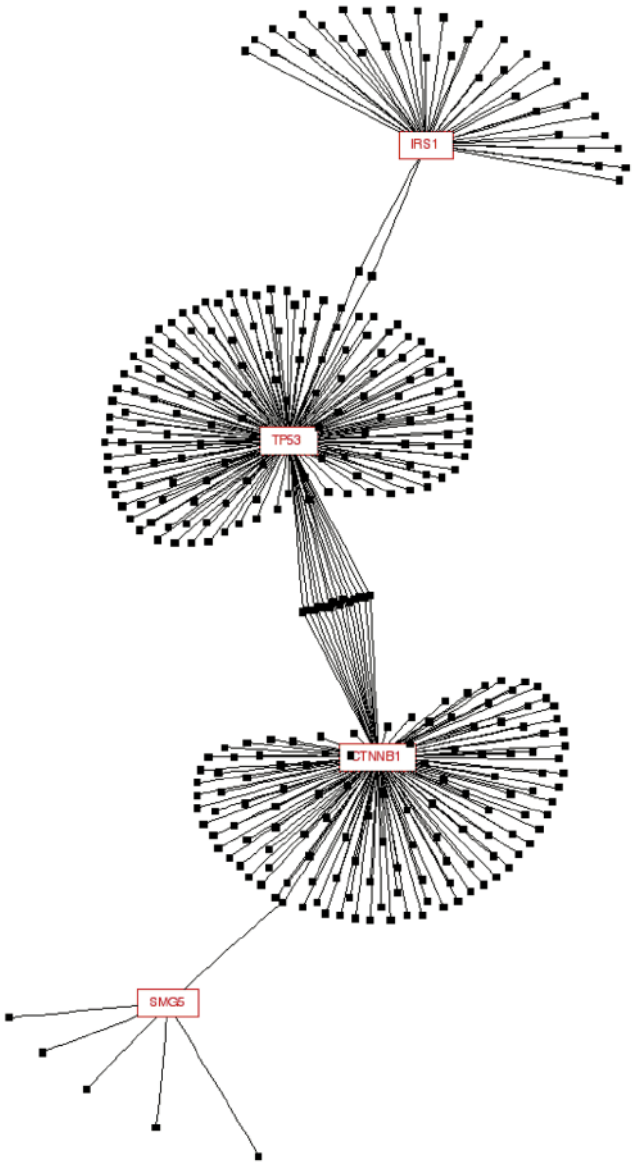

Figure 4. *Gene interaction network of a subset of the 29 mutated genes. Only the four mutated genes that exhibit multiple interactions (IRS1, TP53, CNNB1 and SMG5) are shown in the present figure. Mutated genes are highlighted in red. The genes with which they interact are shown as small solid squares, each of which is connected by a solid line to its interacting mutated gene. The IRS1 interacts with many other genes including BCAR1, IGF1, IGF1R, IL2RG, IL4, IL4R, INS, INSR, IRS2, JAK1, JAK3, PIK3CA, PIK3R1, SH2B2, and SOS1. SMG5 interacts with DKC1, HDAC8, HSP90AA1, PRKACA, TGES3 and TER.*

To evaluate whether the 30 validated mutations are likely to alter protein function, we performed two different statistical analyses, SIFT and LogE (16-17). Both algorithms identified three genes in which the observed mutation was predicted to confer deleterious effects on protein function (Table IV). SIFT identified 13 potentially deleterious mutations that were not detected by LogE; LogE identified one potential functionally significant mutation not detected by SIFT (Table IV). The two *TP53* mutations were predicted to be deleterious by both algorithms.

Pathway analysis of differentially expressed genes. Next, we performed pathway analysis using the up-regulated and down-regulated genes. For a pathway to be considered statistically significant, two or more of differentially expressed genes must "hit" that pathway, with an FDR adjusted p-value of <0.01. When the set of up- and down-regulated genes was analyzed, the Aurora B signaling, FOXM1 transcription factor network and Wnt signaling pathways were observed to be significantly associated with HCC (Table V).

Next we performed gene-gene network analysis to examine whether the 58 up-regulated genes and 98 down-regulated genes interact with each other and form a network. To perform this, we leveraged the gene-gene interaction data present in the NCI/Nature PID. As shown in Figure 3, 27 of these genes are highly connected hubs in a novel, complex network. Importantly, 19 of the 27 hub genes have not yet been reported in hepatocellular carcinoma.

Pathway analysis of mutated genes. We carried out further pathway analysis using the 29 genes with 30 validated mutations by applying the same analytical approach as the one used for the expression pathway analysis. Only six (*CTNNB1, DOCK10, HMGCS1, IRS1, SMG5, TP53*) of the 29 mutated genes are present in canonical pathways. These genes are distributed among 38 pathways in a manner such that no more than one gene "hit" any individual pathway.

In order to gain insight into whether these mutated genes interact to form a larger network, we used an approach that was similar to the pathway analysis for gene expression differences. We discovered that four of the mutated genes (*TP53, IRS1, CTNNB1* and *SMG5*) form a complex network (Figure 4).

As the remaining mutated genes are not contained in any of the pathways in the PID, we performed analyses of these genes using Gene Ontology and found that many were associated with liver metabolism (Table IV).

DNA copy number variations and their association with mutation and expression. When DNA copy number variations (CNVs) were determined for the three paired tumor-matched normal sample sets using the Affymetrix SNP6.0 platform, we identified 18,000 non-overlapping autosomal segments for which at least one HCC tissue sample showed CNV relative to its paired normal tissue sample. Sixteen segments demonstrated copy number gain in all 3 tumors. These copy number gain segments were located in six broad chromosomal regions: 1q21-44, 2q31, 3p24, 8q24, 17q11-12 and 20p12. Sixty-nine segments showed

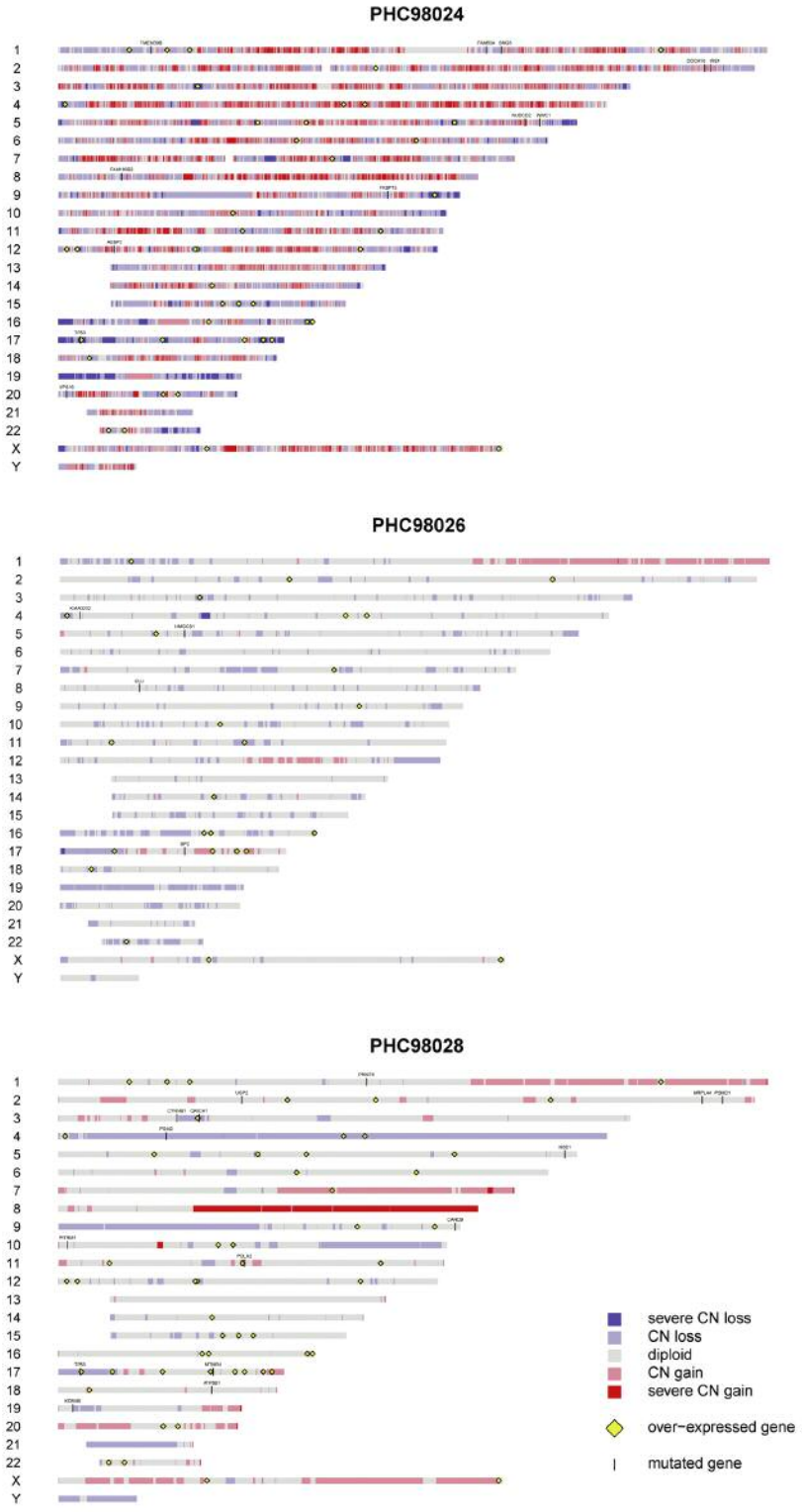

Figure 5. *Genome alterations observed in three HCC samples. Copy number alterations are indicated by shading: red and pink represent copy number gain (red greater than pink) relative to the normal adjacent sample; dark blue to light blue represent copy number loss (dark blue greater than light blue) relative to the matched normal adjacent sample; grey represents no copy number alteration. Yellow diamonds mark the locations of the subset of frequently overexpressed genes that are up-regulated in a particular sample. Black vertical hashes mark the locations of validated non-synonymous somatic mutations identified in a sample. Chromosome numbers are listed on the left side and sample identification numbers are listed at the top of each figure.*

Table II.

A. Assessment of differentially expressed genes.

Expression		Up-regulated (≥2 samples)	Down-regulated (≥2 samples)
RPKM	Fold change		
t≥1 or n ≥1	2	4,513	1,182
t≥10 and n ≤1	≥10	58	–
t≤1 and n ≥10	≥10	–	98

The expression level in reads per kilobase of exon model per million mapped reads (RPKM) and fold change of the three tumor/non-tumor pairs. Only genes up- or down-regulated in at least 2 out of 3 samples are counted. t: Expression level in tumor samples in RPKM; n: expression level in normal samples in RPKM; fold: the fold change of the expression level in tumor relative to non-tumor tissue in paired tumor – non-tumor samples.

B. Complete list of up/down regulated genes

Up-regulated

ADRA2C,ALDH3A1,APCDD1,AURKB,AXIN2,BIRC5,BMP4,BUB1,C12orf75,C1QTNF3,C6orf173,CCNB1,CCNB2,CDC2,CDC45L,CDCA3,CDCA5, CDCA8,CDKN3,CDR2L,CDT1,COL7A1,DBNDD1,DKK1,DNM1,DYNC1I1,EPHB2,FAM83D,FOXM1,IRX3,KIAA0101,KIF20A,KIF2C,LARP6,LEF1, LPPR1,MAGED4,MAGED4B,MAP2K6,MMP11,MYBL2,NCAM1,NEK2,NKD1,PEG10,RACGAP1,REG3A,RNF43,SLC22A11,SLC6A8,SP5,TOP2A,TP BG,TROAP,TSPAN5,TUBB3,USH1C,VCAN

Down-regulated

ACOT12,ADH1A,AFM,AGXT,AKR1D1,APOA4,APOF,AQP7,ASPG,BHMT,C3P1,C8A,C9,C9orf150,CDA,CLEC4G,CNDP1,COLEC10,CPS1,CRHBP, CRP,CXCL14,CYP2A7,CYP2C9,CYP2D6,CYP2E1,CYP3A7,CYP4F2,CYP8B1,DBH,DGAT2,ECHDC3,ECM1,F11,F12,F2,FAM99A,FAM99B,FCN2,F ETUB,FGL1,FLJ23569,G6PC,GBA3,GBP7,GLS2,GLYATL1,GPD1,HAL,HGFAC,HMGCS2,HPD,HPGD,HPN,HSD17B13,HYAL1,IGFALS,ITIH3,KL KB1,KNG1,KRT7,LOC100128675,LOC284422,LOC388503,MARCO,MBL2,MST1,MT1H,MUPCDH,OGDHL,OSTalpha,PCDH24,PLG,PON1,PON3, PRODH2,RAB17,SDCBP2,SERPINA3,SERPINA6,SHBG,SLC10A1,SLC27A2,SLC2A9,SLC38A3,SLC6A1,SLCO1B1,SLPI,SOCS2,SPP2,TMEM82,TM PRSS6,TTC36,UPB1,UROC1,VIPR1,VNN1,XDH

Table III. *Summary statistics of somatic non-synonymous mutation discovery.*

Filtering Tools	Samples			
	PHC98024	PHC98026	PHC98028	Total
Bambino	260,022	206,623	245,806	712,451
Bambino (dbSNP)	29,938	29,497	37,298	96,733
Putative somatic mutations (non-synonymous only)	139	113	93	345
Putative somatic mutations (non-synonymous, dbSNP)	14	2	12	28
After manual review	56	43	50	149
After manual review (dbSNP)	11	2	12	25
Validated (Sanger sequencing)	12	4	14	30
Validated (dbSNP)	2	0	2	4

The following steps were used as filters to identify the final set of somatic mutations: 1) Bambino software identified a total of 712,451 variants out of which 96,733 were present in the dbSNP database; 2) 345 non-synonymous mutations were identified, out of which 28 were in dbSNP; 3) the mutations were then filtered using manual review alignments, revealing 149 mutations out of which 25 were in dbSNP; and 4) as a final step, sixty-three mutations with ≥30% alternative allele frequency were selected for validation by Sanger sequencing. The validation confirmed 30 mutations, out of which 4 were in dbSNP.

copy number loss in all three tumors relative to their matched normal tissues. The largest of these recurrent regions of loss are located in 17p12-13, 19p13 and 10q21-26 (Figure 5).

We then mapped the 30 validated mutations onto CNV segments. Sixteen mutations lie in regions of copy number loss; 14 mutations lie in regions that do not show copy number alterations relative to adjacent normal tissue. Two genes, *NDS1* and *ATP8B1* genes are altered in all 3 tumor samples with mutation in sample PHC98028 and copy number reduction in samples PHC98024 and PHC98026.

Table IV. *Mutations and predicted effects of amino acid changes*.

Gene symbol	Description	Sample ID	Nucleotide change	Amino acid change	LogE/ SIFT*
AEBP2	AE binding protein 2	98024	A>G	M377V	-/t
*ATP8B1***	ATPase, class I, type 8B, member 1	98028	G>T	N899K	-/d
CARD9	Caspase recruitment domain family, member 9	98028	C>A	A331S	-/t
*CLU***	Clusterin	98026	G>A	H37Y	-/t
*CTNNB1**	Catenin (cadherin-associated protein), beta 1, 88kDa	98028	C>A	S37Y	-/d
DOCK10	Dedicator of cytokinesis 10	98024	T>G	E2129A	d/d
FAM160B2	Family with sequence similarity 160, member B2	98024	A>G	T315A	-/t
FAM63A	Family with sequence similarity 63, member A	98024	T>G	E168A	-/t
FKBP15	FK506 binding protein 15, 133kDa	98024	T>C	H515R	-/d
*HMGCS1***	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1	98026	T>C	T125A	t/d
*IRS1***	Insulin receptor substrate 1	98024	T>C	T793A	-/t
KDM4B	Lysine (K)-specific demethylase 4B	98028	T>A	F203I	t/d
KIAA0232	KIAA0232	98026	A>G	I1100V	-/d
MRPL44	Mitochondrial ribosomal protein L44	98028	G>T	A21S	-/t
MTMR4	Myotubularin related protein 4	98028	T>C	Y1121C	t/d
NSD1	Nuclear receptor binding SET domain protein 1	98028	C>A	R1661S	-/d
NUDCD2	NudC domain containing 2	98024	G>T	D116E	-/t
*PGM2***	Phosphoglucomutase 2	98028	G>T	A574S	-/t
PITRM1	Pitrilysin metallopeptidase 1	98028	T>A	Q8L	t/d
POLA2	Polymerase (DNA directed), alpha 2	98028	T>C	F337S	-/t
*PRMT6***	Protein arginine methyltransferase 6	98028	C>T	A135V	-/t
PSMD1	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 1	98028	A>G	Y147C	-/d
QRICH1	Glutamine-rich 1	98028	T>C	Y550C	-/d
SMG5	Smg5 homolog, nonsense mediated mRNA decay factor	98024	C>T	V794I	-/d
SP2	Sp2 transcription factor	98026	A>T	Q131L	-/d
TMEM39B	Transmembrane protein 39B	98024	T>C	F323S	-/t
*TP53**	Tumor protein p53	98028	G>A	P278S	d/d
*TP53**	Tumor protein p53	98024	T>C	Y236C	d/d
VPS16	Vacuolar protein sorting 16 homolog	98024	G>A	A235T	d/t
WWC1	WW and C2 domain containing 1	98024	A>T	E862V	-/t

*Genes previously identified as being involved in HCC. **Genes involved in metabolism.A logE score less than -1 or greater than 1 indicates the amino acid change is likely to affect protein function; a SIFT score less than or equal to 0.05 predicts that the amino acid variant is deleterious (see Methods). Results are indicated in the last column: *LogE and SIFT results: t, tolerant; d, deleterious; na, not analyzed; -, no.

Table V. *Pathway analysis of 58 up-regulated and 98 down-regulated genes*.

FDR	Pathway	Genes
2.14E-06	Aurora B signaling	AURKB,BIRC5,BUB1,CDCA8,KIF20A,KIF2C,RACGAP1
2.64E-05	FOXM1 transcription factor network	AURKB,BIRC5,CCNB1,CCNB2,FOXM1,NEK2
1.79E-03	Canonical Wnt signaling pathway	AXIN2,DKK1,LEF1,NKD1
6.57E-03	Signaling by Aurora kinases	AURKB

Pathways that are statistically significantly associated with the subset of genes that are up-regulated or down-regulated in HCC tumor versus normal tissue. These pathways are derived from the NCI/Nature PID and are considered to have statistically significant associations when correction for the FDR (false discovery rate) has a $p<0.01$.

In similar fashion, we investigated possible correlations of gene expression with copy number alterations. We observed that of the 58 up-regulated genes, only *PEG10, RNF43, AXIN2, MAP2K6, BIRC5* and *APCDD1* are located at positions exhibiting recurrent patterns of copy number gain. Mapping of the 98 down-regulated genes onto CNV regions revealed that 42 genes (43%) map to regions with frequent copy number loss. In contrast, six down-regulated genes are located in genomic regions showing recurrent copy number gain.

Discussion

Considerable heterogeneity exists in the somatic genetic etiology of HCC. Although the origin of this heterogeneity is still unknown, molecular studies of HCC suggest a connection not only with HBV/HCV infection, but also with abnormalities of metabolic balance, such as diabetes and obesity (22-23). To address this issue, we surveyed three paired HCC and adjacent normal tissue samples using RNA-seq. We identified differentially expressed genes as well as somatic mutations. We observed 58 up-regulated and 98 down-regulated genes in tumor samples when compared to adjacent normal liver tissue (Table II). Aurora B signaling, FOXM1 transcription factor network and Wnt signaling pathways are significantly enriched in the differentially expressed genes of HCC. Only Wnt the signaling pathway has been previously implicated in the development of HCC. Thus, all pathways, which play a key role in cell proliferation, cytokinesis and DNA repair, may be involved in the development and progression of HCC (11, 24-25).

Gene-gene network interaction analysis of the 156 differentially expressed genes revealed that 27 genes form a large complex network (Figure 3). This analysis allowed us to investigate new networks that have not previously been associated with HCC. The following up-regulated genes were identified by this analysis: *KIF2C* and *CDCA8*, key components of the Aurora B signaling pathway (24, 26); *NKD1*, a major component of the Wnt signaling pathway (27); *CCNB2*, a key member of the *FOXM1* pathway (28); *RACGAP1*, a key member of the cytokinesis pathway (29); and *DNM1*, Dynamin, a large GTPase involved in clathrin-mediated endocytosis and other vesicular trafficking processes (30). As discussed above, only genes belonging to the Wnt signaling pathway have previously been documented in relation to HCC carcinogenesis. With regard to the down-regulated genes, *HMGCS2* (3-hydroxy-3-methylglutaryl-coenzyme A synthase) and *PLG* (plasminogen) are newly-identified in HCC and have functions that are of interest from the perspective of carcinogenesis. *HMGCS2* expression is associated with differentiation and its down-regulation has been demonstrated in moderately- and poorly-differentiated colorectal adenocarcinomas (31). The protein encoded by the *PLG* gene, plasminogen, is a secreted blood zymogen and is converted to plasmin and angiostatin when activated by proteolysis. Angiostatin, in turn, is a potent inhibitor of angiogenesis, tumor growth and metastasis; consistent with the notion that angiogenesis contributes to carcinogenesis in tumors characterized by down-regulated *PLG*.

HCC typically develops in the setting of liver cirrhosis, which is associated with chronic liver disease (1). The current dominant view is that non-viral associated HCCs evolve in tissue that has been subjected to physiological insults. For example, disruption of normal metabolic balance results in an increase in the release of free fatty acids (FFA) from adipocytes, release of multiple pro-inflammatory cytokines including tumor necrosis factor-alpha (*TNF*), interleukin-6 (*IL6*), leptin, and resistin (7). Together, the elevated levels of such factors contribute to abnormal hepatic conditions, including NAFLD and NASH, which may progress to cirrhosis and ultimately to HCC (7). Among the 30 mutations identified in our tumor samples six are located in genes that are implicated in metabolic processes, five of which are involved in diseases such as diabetes and obesity: *IRS1, HMGCS1, ATP8B1, PRMT6* and *CLU* (Table IV).

IRS1 is known to be involved in diabetes, obesity and activation of cytokine signaling pathways (32-33). In addition, overexpression of *IRS1* has been documented in over 90% of HCC cases (34). Yet, mutations in *IRS1* have not previously been reported in HCC. The identification of a novel non-synonymous mutation in codon 793 (threonine to alanine) potentially alters the activity of the IRS1 protein. The position of this mutation supports such a functional effect since the mutation is adjacent to Ser794, which has previously been shown to play a critical role in insulin signal transduction (35-36). Changes in the phosphorylation state of Ser794 are associated with a variety of insulin-mediated activities (35, 37). Specifically, phosphorylation of this amino acid alters the efficiency of insulin signal transduction, eventually causing insulin resistance in diabetic animals (37). The multiple roles played by phosphorylated Ser794 are potentially influenced by the mutation that we observed in the neighboring Thre793. Thus, this discovery of an *IRS1* mutation in HCC suggests that dysregulated *IRS1* may play a role in the etiology of liver cancer.

Interaction analysis of canonical pathways contained in the PID (Figure 4) reinforces the potential importance of the role played by *IRS1* in HCC tumorigenesis. Thus, network analysis of all possible interactions among the 29 mutated genes in our study revealed that the network comprises four hub genes: *IRS1, TP53, CTNNB1* and *SMG5*. All four of these network-interacting mutated genes individually affect processes known to be dysregulated in cancer. Apart from *TP53*, the most commonly mutated gene in cancer, *CTNNB1*, encodeing beta-catenin, is frequently mutated in specific cancers, including colon cancer and HCC (38-39). The fourth gene in the network hub is *SMG5*, which is involved in protection of telomere integrity, and thus maintenance of chromosomal stability (40). Furthermore, degradation of *SMG5* results in a major reduction of total telomerase activity (41).

TP53 mutation has a well-established association with many cancers, including HCC. HCV-infected cells exhibit a tendency toward increased mutation rate, leading to a 5-10-fold increase in mutation frequency of genes such as the immunoglobulin heavy chain gene, *BCL-6, TP53*, and the *CTNNB1* gene (39). Our own initial observations concur, given that two of our three original liver cancer samples

infected with HCV are also the same two that exhibit mutations in the *TP53* gene. In addition, 50% of the additional 25 HCC tissues which we analyzed exhibited *TP53* mutations; these *TP53*-mutated samples were also positive for HBV infection, unlike the other 50% HCC samples, which were *TP53*-mutation negative. Furthermore, out of the three HCC samples (Figure 5), the two samples with HCV infection and *TP53* mutation (PHC98024 and PHC98028) exhibited the highest levels of copy number variations. This observation is consistent with the known role of the TP53 protein in maintaining genomic stability *via* the DNA repair mechanisms (42). In contrast to the common epithelial cancers such as breast, prostate, and colon, HCC is not characterized by a high rate of mutations in conventional oncogenes and tumor suppressor genes, other than *TP53* (43). The results of our study are consistent with a model in which HCC development is associated with metabolic changes due to HCV/HBV infection, diabetes, obesity or metabolic syndrome. Insulin resistance and the subsequent inflammatory cascade that characterize NAFLD, NASH, and cirrhosis appear to play an important role in carcinogenesis in the liver. Thus, the entire constellation of these metabolic changes culminates in a substantially increased risk of HCC development.

Acknowledgements

The Authors would like to acknowledge and thank Dr. Katherine McGlynn for insightful discussions and Dr. Gang Wu for technical assistance.

References

1 Gomaa AI, Khan SA, Toledano MB, Waked I and Taylor-Robinson SD: Hepatocellular carcinoma: epidemiology, risk factors and pathogenesis. World J Gastroenterol *14*: 4300-4308, 2008.

2 Lata J: Chronic liver diseases as liver tumor precursors. Dig Dis *28*: 596-599, 2010.

3 Davila JA, Morgan RO, Shaib Y, McGlynn KA and El-Serag HB: Hepatitis C infection and the increasing incidence of hepatocellular carcinoma: a population-based study. Gastroenterology *127*: 1372-1380, 2004.

4 El-Serag HB, Davila JA, Petersen NJ and McGlynn KA: The continuing increase in the incidence of hepatocellular carcinoma in the United States: an update. Ann Intern Med *139*: 817-823, 2003.

5 Starley BQ, Calcagno CJ and Harrison SA: Nonalcoholic fatty liver disease and hepatocellular carcinoma: a weighty connection. Hepatology *51*: 1820-1832, 2010.

6 Bugianesi E, Vanni E andMarchesini G: NASH and the risk of cirrhosis and hepatocellular carcinoma in type 2 diabetes. Curr Diab Rep *7*: 175-180, 2007.

7 Qureshi K and Abrams GA: Metabolic liver disease of obesity and role of adipose tissue in the pathogenesis of nonalcoholic fatty liver disease. World J Gastroenterol *13*: 3540-3553, 2007.

8 Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*: 621-628, 2008.

9 Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF *et al*: Stem cell transcriptome profiling *via* massive-scale mRNA sequencing. Nat Methods *5*: 613-619, 2008.

10 Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M *et al*: A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science *321*: 956-960, 2008.

11 Roessler S, Jia HL, Budhu A, Forgues M, Ye QH, Lee JS, Thorgeirsson SS *et al*: A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer Res *70*: 10202-10212, 2010.

12 Edmonson MN, Zhang J, Yan C, Finney RP, Meerzaman DM and Buetow KH: Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. Bioinformatics *27*: 865-866, 2011.

13 Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM *et al*: VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics *25*: 2283-2285, 2009.

14 Rozen S and Skaletsky H: Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol *132*: 365-386, 2000.

15 Ewing B, Hillier L, Wendl MC and Green P: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res *8*: 175-185, 1998.

16 Clifford RJ, Edmonson MN, Nguyen C and Buetow KH: Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics *20*: 1006-1014, 2004.

17 Ng PC and Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res *31*: 3812-3814, 2003.

18 Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T and Buetow KH: PID: the Pathway Interaction Database. Nucleic Acids Res *37*: D674-679, 2009.

19 Benjamini Y, Drai D, Elmer G, Kafkafi N and Golani I: Controlling the false discovery rate in behavior genetics research. Behav Brain Res *125*: 279-284, 2001.

20 Olshen AB, Venkatraman ES, Lucito R and Wigler M: Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics *5*: 557-572, 2004.

21 Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ *et al*: An anatomy of normal and malignant gene expression. Proc Natl Acad Sci USA *99*: 11287-11292, 2002.

22 Davila JA, Morgan RO, Shaib Y, McGlynn KA and El-Serag HB: Diabetes increases the risk of hepatocellular carcinoma in the United States: a population based case control study. Gut *54*: 533-539, 2005.

23 Sanyal A, Poklepovic A, Moyneur E and Barghout V: Population-based risk factors and resource utilization for HCC: US perspective. Curr Med Res Opin *26*: 2183-2191, 2010.

24 Hayama S, Daigo Y, Yamabuki T, Hirata D, Kato T, Miyamoto M, Ito T *et al*: Phosphorylation and activation of cell division cycle associated 8 by aurora kinase B plays a significant role in human lung carcinogenesis. Cancer Res *67*: 4113-4122, 2007.

25 Wierstra I and Alves J: FOXM1, a typical proliferation-associated transcription factor. Biol Chem *388*: 1257-1274, 2007.

26 Shimo A, Tanikawa C, Nishidate T, Lin ML, Matsuda K, Park JH, Ueki T *et al*: Involvement of kinesin family member 2C/mitotic centromere-associated kinesin overexpression in mammary carcinogenesis. Cancer Sci *99*: 62-70, 2008.

27 Guo J, Cagatay T, Zhou G, Chan CC, Blythe S, Suyama K, Zheng L *et al*: Mutations in the human naked cuticle homolog NKD1 found in colorectal cancer alter Wnt/Dvl/beta-catenin signaling. PLoS One *4*: e7982, 2009.

28 Wang X, Quail E, Hung NJ, Tan Y, Ye H and Costa RH: Increased levels of forkhead box M1B transcription factor in transgenic mouse hepatocytes prevent age-related proliferation defects in regenerating liver. Proc Natl Acad Sci USA *98*: 11468-11473, 2001.

29 Zhao WM and Fang G: MgcRacGAP controls the assembly of the contractile ring and the initiation of cytokinesis. Proc Natl Acad Sci USA *102*: 13158-13163, 2005.

30 Gammie AE, Kurihara LJ, Vallee RB and Rose MD: DNM1, a dynamin-related gene, participates in endosomal trafficking in yeast. J Cell Biol *130*: 553-566, 1995.

31 Camarero N, Mascaro C, Mayordomo C, Vilardell F, Haro D and Marrero PF: Ketogenic HMGCS2 Is a c-Myc target gene expressed in differentiated cells of human colonic epithelium and down-regulated in colon cancer. Mol Cancer Res *4*: 645-653, 2006.

32 Musi N and Goodyear LJ: Insulin resistance and improvements in signal transduction. Endocrine *29*: 73-80, 2006.

33 Boden G: Fatty acid-induced inflammation and insulin resistance in skeletal muscle and liver. Curr Diab Rep *6*: 177-181, 2006.

34 Longato L, de la Monte S, Califano S and Wands JR: Synergistic premalignant effects of chronic ethanol exposure and insulin receptor substrate-1 overexpression in liver. Hepatol Res *38*: 940-953, 2008.

35 Ning J and Clemmons DR. AMP-activated protein kinase inhibits IGF-I signaling and protein synthesis in vascular smooth muscle cells *via* stimulation of insulin receptor substrate 1 S794 and tuberous sclerosis 2 S1345 phosphorylation. Mol Endocrinol *24*: 1218-1229, 2010.

36 Leclerc GM, Leclerc GJ, Fu G and Barredo JC: AMPK-induced activation of Akt by AICAR is mediated by IGF-1R dependent and independent mechanisms in acute lymphoblastic leukemia. J Mol Signal *5*: 15, 2010.

37 Horike N, Takemori H, Katoh Y, Doi J, Min L, Asano T, Sun XJ *et al*: Adipose-specific expression, phosphorylation of Ser794 in insulin receptor substrate-1, and activation in diabetic animals of salt-inducible kinase-2. J Biol Chem *278*: 18440-18447, 2003.

38 Kim YD, Park CH, Kim HS, Choi SK, Rew JS, Kim DY, Koh YS *et al*: Genetic alterations of Wnt signaling pathway-associated genes in hepatocellular carcinoma. J Gastroenterol Hepatol *23*: 110-118, 2008.

39 Machida K, Cheng KT, Sung VM, Shimodaira S, Lindsay KL, Levine AM, Lai MY *et al*: Hepatitis C virus induces a mutator phenotype: enhanced mutations of immunoglobulin and protooncogenes. Proc Natl Acad Sci USA *101*: 4262-4267, 2004.

40 Azzalin CM and Lingner J: The human RNA surveillance factor UPF1 is required for S phase progression and genome stability. Curr Biol *16*: 433-439, 2006.

41 Lee H, Sengupta N, Villagra A, Rezai-Zadeh N and Seto E: Histone deacetylase 8 safeguards the human ever-shorter telomeres 1B (hEST1B) protein from ubiquitin-mediated degradation. Mol Cell Biol *26*: 5259-5269, 2006.

42 Hussain SP, Schwank J, Staib F, Wang XW and Harris CC: TP53 mutations and hepatocellular carcinoma: insights into the etiology and pathogenesis of liver cancer. Oncogene *26*: 2166-2176, 2007.

43 Imbeaud S, Ladeiro Y and Zucman-Rossi J: Identification of novel oncogenes and tumor suppressors in hepatocellular carcinoma. Semin Liver Dis *30*: 75-86, 2010.

Received February 6, 2014
Revised February 21, 2014
Accepted February 24, 2014