RESEARCH ARTICLE

Profiling of Exome Mutations Associated with Progression of HBV-Related Hepatocellular Carcinoma

Hyun Goo Woo[1,3], Soon Sun Kim[2], Hyunwoo Cho[1,3], So Mee Kwon[1,3], Hyo Jung Cho[3], Seun Joo Ahn[2], Eun Sung Park[4], Ju-Seog Lee[5], Sung Won Cho[2], Jae Youn Cheong[2]*

1. Department of Physiology, Ajou University School of Medicine, Suwon, Republic of Korea, **2.** Department of Gastroenterology, Ajou University School of Medicine, Suwon, Republic of Korea, **3.** Graduate School of Biomedical Science, Ajou University School of Medicine, Suwon, Republic of Korea, **4.** Institute for Medical Convergence, Yonsei University College of Medicine, Seoul, Korea, **5.** Department of Systems Biology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas, United States of America

*jaeyoun620@gmail.com

OPEN ACCESS

Abstract

Recent advances in sequencing technology have allowed us to profile genome-wide mutations of various cancer types, revealing huge heterogeneity of cancer genome variations. However, its heterogeneous landscape of somatic mutations according to liver cancer progression is not fully understood. Here, we profiled the mutations and gene expressions of early and advanced hepatocellular carcinoma (HCC) related with Hepatitis B-viral infection. Integrative analysis was performed with whole-exome sequencing and gene expression profiles of the 12 cases of early and advanced HCCs and paired non-tumoral adjacent liver tissues. A total of 293 tumor-specific somatic variants and 202 non-tumoral variants were identified. The tumor-specific variants were found to be enriched at chromosome 1q particularly in the advanced HCC, compared to the non-tumoral variants. Functional enrichment analysis revealed frequent mutations at the genes encoding cytoskeleton organization, cell adhesion, and cell cycle-related genes. In addition, to elucidate actionable somatic mutations, we performed an integrative analysis of gene mutations and gene expression profiles together. This revealed the 48 mutated genes which were differentially mutated with concomitant gene expression enrichment. Of these, *CTNNB1* was found to have a pivotal role in the differential progression of the HCC subgroup. In conclusion, our integrative analysis of whole-exome sequencing and transcriptome profiles could provide actionable mutations which might play pivotal roles in the heterogeneous progression of HCC.

Introduction

Genomic heterogeneity of hepatocellular carcinoma (HCC) makes it difficult to characterize molecular pathogenesis and to develop efficient treatment modalities. During HCC development, aberrant genetic and epigenetic events occurred and accumulated, which provoked heterogeneous pattern of HCC progression. Previously, a number of sequence variations of HCC have been reported. For example, *TP53*, *CTNNB1*, and *AXIN1* have been reported to associate with the diverse progression pattern of HCC [1, 2]. Recently, the advance of high-throughput sequencing technology so called next generation sequencing (NGS) allowed us to profile mutations in a whole-genome scale. By applying NGS technology, genome-wide mutational spectra of HCC have been reported [3–7]. Numerous novel mutations such as *ARID1A* [8], *IRF2* [9], and *JAK1* [10] have been identified. In addition, the differential mutation spectrum of hepatitis B and C-related HCC has been studied [11]. By applying evolution models, stage-specific driver mutations (e.g., *CCNG1* and *P62*) have been noticed [12]. *LEPR* has been reported to be frequently mutated in the HCC-surrounded cirrhotic liver affecting tumor growth [13]. However, previous studies are still limited in the sample numbers, and the effect of mutations on the heterogeneous progression of HCC was not fully considered. Indeed, it is well known that the tumor grade is associated with the heterogeneous gene expressions and clinical outcomes of HCC [14]. Considering such studies, we sought to evaluate the role of mutation profiles with the tumor grader of HCC. In the present study, we performed whole–exome sequencing and compared the mutation profiles of HCC with different tumor grade of the early and the advanced cases. In addition, by performing integrative analysis of mutational profiles with their corresponding gene expression profiles, we sought the potential key regulators which are responsible for the heterogeneous progression of HCC. Our analysis could reveal prioritized candidates of functional and actionable mutations involved in the HCC progression, providing novel insight into the regulatory roles of mutation profile in the heterogeneous progression of HCC.

Methods

1. Sample Preparation

A total of 12 cases of HBV-related HCC samples and paired adjacent non-tumoral tissues were obtained from the Ajou Human Bio-Resource Bank (AHBB), a member of the National Biobank of Korea, which is supported by the Ministry of Health and Welfare. The Institutional Review Board of Ajou University Hospital at Korea has approved this study, and waived the need for informed consent from donors.

2. Exome Capture and High-Throughput Sequencing.

Whole exome sequencing was performed using the 12 cases of frozen HCC tissues. For targeted exome capture, Illumina TruSeq exome enrichment kit was used with the given protocol. The captured samples were sequenced as 110 bp paired-end reads using Illumina GAIIx with the average coverage of $30 \times$. The 75 raw read bases from 5′ end start position were used trimming out the remaining 3′ sequences because the 10th percentile of quality scores in each sample is less than 20. The trimmed reads were mapped to hg19 human reference genome using Burrows-Wheeler Alignment tool (BWA) [15] with default parameters. The mapping quality of the resulting.sam files was inspected, and those with zero quality were filtered out to reduce the false positive mapped reads. The PCR duplicates were identified and removed by using the Genome Analysis Toolkit (GATK) [16]. Then, local realignments of indels were performed using GATK local realignment walker [17], and the read quality was normalized using the GATK recalibration walker. Quality filter was performed using GATK unified genotyper with filter options of Hard to validate (MQ0$>$=4 and MQ0/DP$>$0.1), Low coverage (DP $<$5), Low quality (QUAL $<$50.0), and Low quality-by-depth (QD $<$1.5). Each variant was annotated by using ANNOVAR [18]. Validation of the identified mutations was performed by Sanger (capillary) sequencing method.

3. Gene Expression Profiling.

Total RNA was extracted by using the mirVana total RNA extraction kit (Ambion, Austin, TX, U.S.A) according to manufacturer's instruction and amplified by using Illumina TotalPrep 96 RNA Amplification Kit (IIlumina). Gene expression profiling was conducted with Illumina HumanHT-12 v4 Expression BeadChip kit. Microarray hybridization, image acquisition and processing were performed according to the manufacturer's guidelines. The raw data were log2 transformed, quantile normalized, and centered to each array mean for further analyses.

4. Functional Enrichment Analysis of Gene Sets.

The enriched functions in the mutated gene sets were analyzed using Gene Ontology (GO) categories and KEGG database implemented in DAVID software [19]. Conservative statistical significance of the enrichments was estimated by EASE scores from modified Fisher's exact T-test P-values. For the gene expression profile analysis, the functional enrichment of gene sets in individual patient was determined by applying Kolmogorov-Smirnov (KS) test. For each individual gene expression profile, the directional *P*-values for the estimates D+ and D- were calculated by KS-test, and the enrichment score for a given signature was calculated as -$\log_{10}$ (*P*-value) as described previously [20]. The GO gene sets with more than ten genes were considered for analysis. All the statistical computation was performed using R software (http://www.r-project.org).

Results

1. Identification of tumor-specific and non-tumoral mutations

Whole-exome captured sequencing was performed on the 6 cases of early HCC with tumor grade I or II, and the 6 cases of advanced HCC with grade III or IV, and the paired non-tumoral tissues. Clinical and pathological features of the specimens were described in S1 Table.

Previously, it has been noticed that the current high-throughput NGS technologies still have limits because of substantial error rates of false base calls even with deep coverage profiling of sequence leads [21, 22]. Moreover, intra-tumoral heterogeneity of the mutations may lead to false negative calls especially in case of the use of small amount of tissues for sequencing experiment [23]. Thus, a stringent and careful variant calling pipeline might be required to reduce the false positivity of sequencing errors even for deep coverage data. With respect to this, we built an optimized pipeline for tumor-specific variant calls from our exome-seq data. First, following the classical sequence alignment pipeline using bwa and Genome Analysis Toolkit (for details see Methods), a total of 131,947 variants were mapped to human genome (hg19). Of these, we found a total of 46,510 exonic variants which were mapped with at least ten reads. Next, we tried to identify the most probable tumor-specific and non-tumoral tissue-specific variants, respectively. To estimate the tumor-specificity and/or non-tumor-specificity of the variants, we first calculated the odds between the number of variant counts of the paired tumor and non-tumor tissues (Fisher exact t-test, $P<0.01$), which yielded 1,705 tumor-specific and 1,360 non-tumor-specific variants, respectively. The non-tumoral variants were regarded as the background mutations for each patient. The variants called both in tumor and non-tumor tissues were considered as germline variants or non-specific calls, thus, we filtered them by applying more stringent filtering criteria of read counts $\geq$5 in the paired tissues. Then, to avoid false variant calls generated by possible local mis-alignment errors, the hyper-mutable regions with more than ten variants within a 1,000 base pair-sized window were also removed. Previously, kataegic foci with local enrichment of C>T mutations have been noticed in various cancers [24, 25]. However, the filtered hyper-mutable regions were not likely to be kataegis because the C>T enrichment was not observed. After removing the known SNPs (dbSNP135 and 1000 genome), we finally obtained 293 tumor-specific and 202 non-tumor-specific variants, which we used for further analyses. The details of the filtering pipeline were shown in Fig. 1A.

2. Genomic landscape of the tumor-specific variants

The 293 tumor-specific mutations include 202 nonsynonymous, 78 synonymous variants, and 13 codon stop-signaling variants (Fig. 1B). Previously, the synonymous variants so called silent mutations have been noticed to play driver roles in cancer development and progression [26, 27], therefore we did not exclude the synonymous variants in the subsequent analysis. Of the tumor-specific
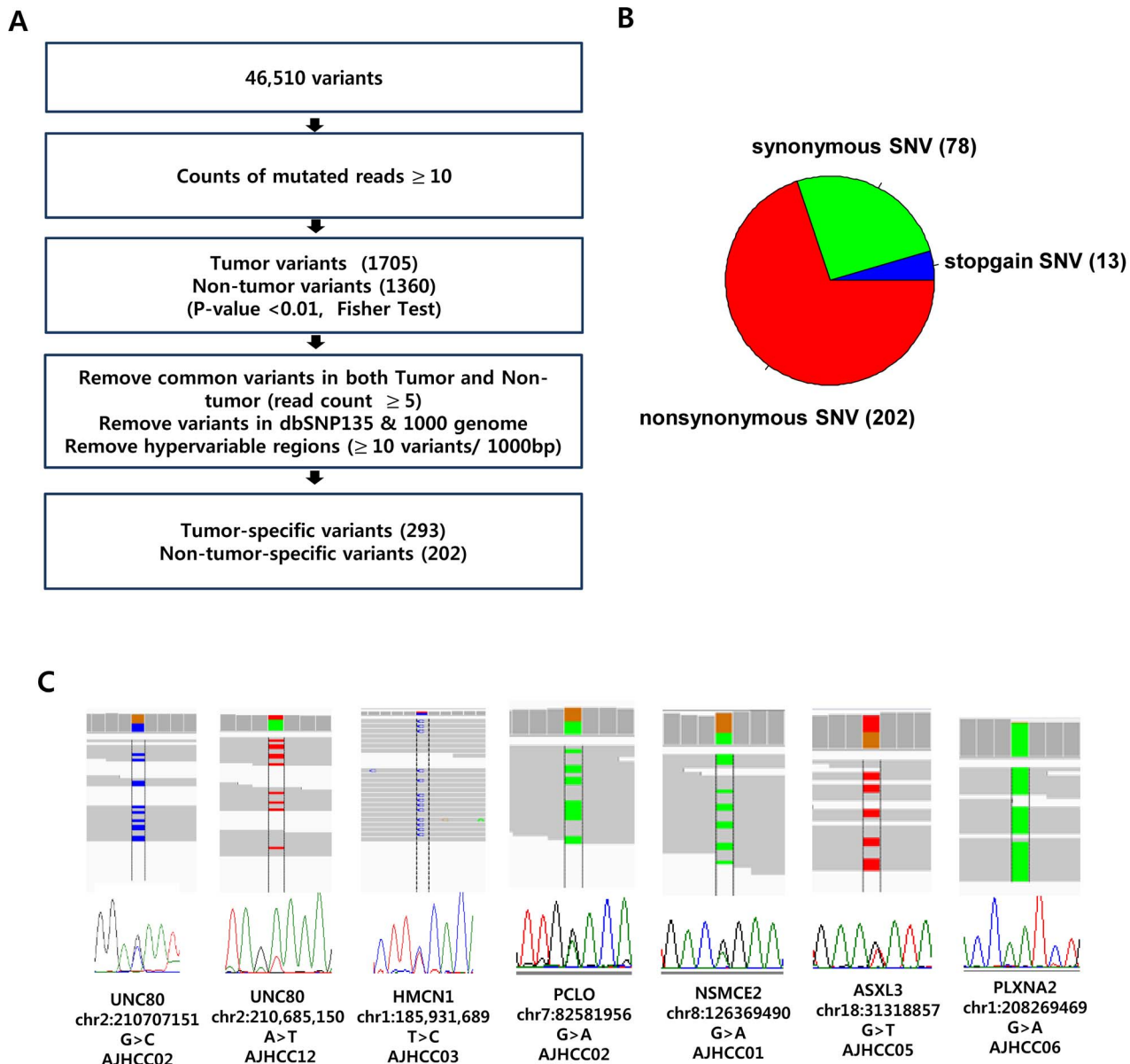
A

B

C

Fig. 1. Identification of tumor-specific and non-tumoral variants. A. The flowchart for algorithms identifying tumor-specific and non-tumor-specific variants are shown. **B.** Distribution of somatic tumor-specific variants in HCC is shown. **C.** The tumor-specific variants are validated by Sanger sequencing method, and their read alignments are evaluated by Integrated Genome Browser (IGV) software.

doi:10.1371/journal.pone.0115152.g001

mutations, 148 genes including *CTNNB1, TTN, SETD2, ALK* have been previously observed in HCC from Catalogue Of Somatic Mutations In Cancer database (COSMIC, http://www.sanger.ac.uk/genetics/CGP/cosmic, v68) [28]. However, some of the well-known recurrent mutations such as *TP53, ARID1,* and *AXIN1* genes [11, 29] were not observed in our data, which might be due to small sample size and the application of high stringent criteria for variant calling. Of the tumor-specific mutations, the recurrent mutations with more than two

observations were found in the 9 genes encoding *TCHHL1*, *PLXNA2*, *UNC80*, *CEP85L*, *PCLO*, *NSMCE2*, *ASXL3*, *SULF1*, and *MARS*. The frequencies of the tumoral and non-tumoral variants in each patient are highly variable ranging from 8 to 99, and 1 to 147, respectively. There was no significant difference of variant frequency between the groups of early and advanced HCCs (S1 Figure). We validated some of these variants by performing Sanger sequencing analysis and manual visualization of the sequence reads using Integrated Genome Viewer (IGV) [30] (Fig. 1C). This may support that our stringent variant calling can identify valid mutations despite of the relative low coverage of the sequencing reads. The list of tumor-specific and non-tumor-specific variants was summarized in S2 Table.

Next, we examined the genomic patterns of the tumor-specific variants. Previously, the C:G>T:A transition is thought as a characteristic mutational signature of HCV-associated HCC [4]. However, another report has shown that the C:G>T:A transition was commonly found in HBV-associated HCC [5]. Consistently, we observed that the frequent C:G>T:A transition in both tumor-specific and non-tumoral variants. The prevalence of C:G>T:A transition was thought to associate with the aflatoxin B1 exposure[5]. However, the Korean patients in our study are not likely to relate with aflatoxin B1 exposure. Therefore, further elaboration might be required to validate and to explain the cause of the prevalence of the C:G>T:A transition. All the T:A>A:T transversions were found as tumor-specific variants (Wilcox test, p<0.001, Fig. 2A), which were consistent with the previous study regardless of different use of variant calling algorithms and patient cohorts[5].

We next examined the existence of any preferential site of the tumor-specific variants. Remarkably, we found that the tumor-specific somatic variants were significantly enriched at chromosome 1q compared to those in the non-tumoral tissues (Fisher's exact test, $P=0.009$, Fig. 2B). The enriched mutations at 1q are not related with the size of the chromosome arms, because the enrichment scores of mutations were calculated by comparing to the number of mutations at the same chromosome arms of the non-tumoral tissues and the entire number of mutations in the whole genome of each patient. To confirm this finding in individual samples, we evaluated the mutation enrichment of chromosome arms for each sample data. Considering the individual variations of the background non-tumoral mutations, the enrichment of the tumor-specific variants was normalized with that of the non-tumor specific variants. In detail, the mutation enrichment score (ES) for each chromosome arm for each patient (j) was estimated by the odds ratios of the numbers of somatic variants (N) in each chromosome arm (i) against those of outside the chromosome arm (-i).

$$ES = N_{Tumor}(i,j) \bullet \pm N_{non-tumor}(-i,j) / N_{Tumor}(-i,j) \bullet N_{non-tumor}(i,j)$$

By calculating the enrichment scores of each chromosome arm, we found that the tumor-specific variants at 1q was significantly enriched in the advanced HCC group compared to early HCC group (Fig. 2C). This result suggests that the frequent mutations at 1q might be actionable with the association of the aggressive HCC progression. The genes with the enriched mutations at 1q
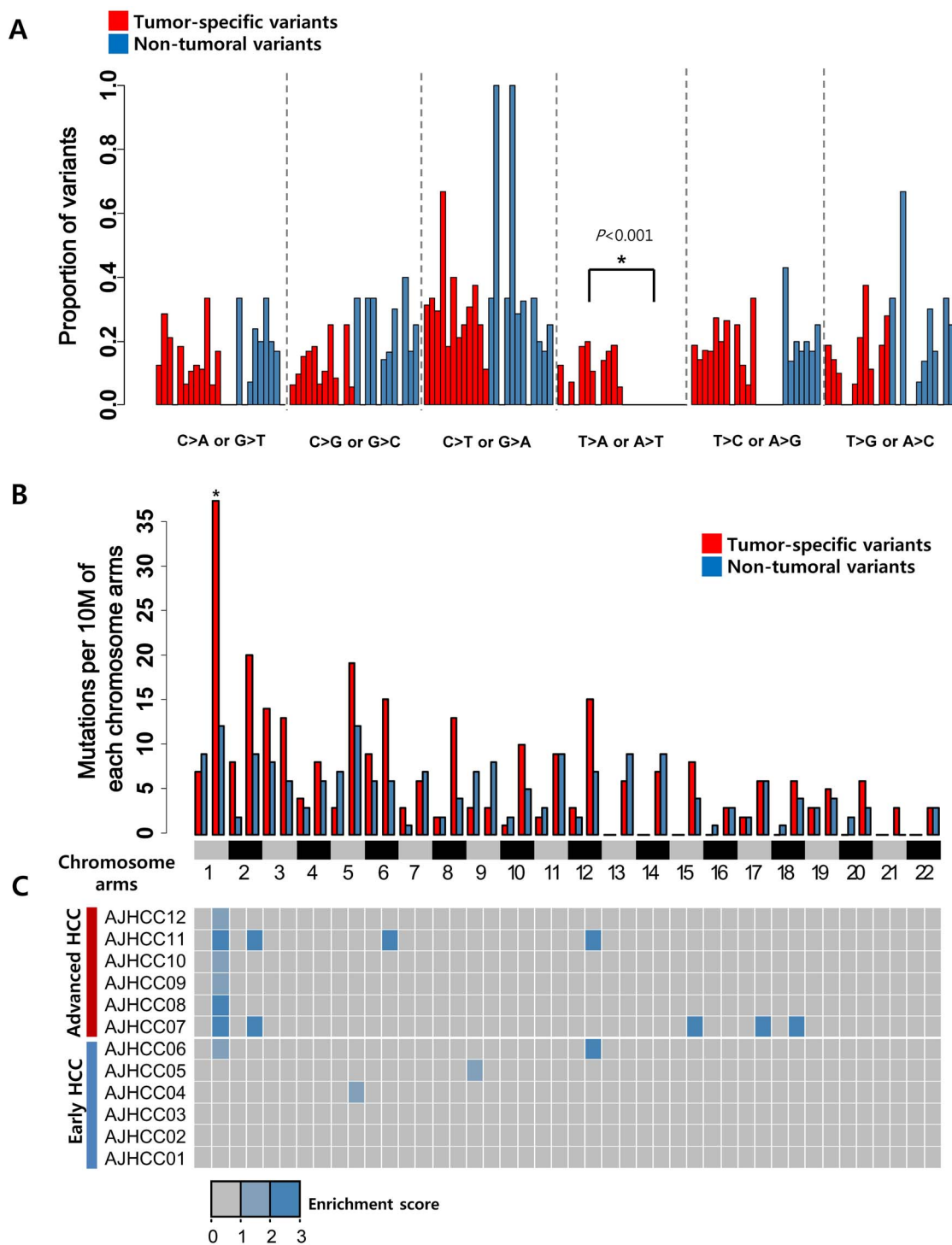
Fig. 2. The mutation spectra of tumor-specific variants. A. Mutation spectrum of tumor-specific and non-tumor-specific variants are shown. **B.** The observed numbers of mutations per 10 million bases in each chromosome arm are plotted for tumor-specific and non-tumoral variants, respectively. Statistical analysis compared the occurrence of tumor-specific variants with that of non-tumoral variants. (*$P<0.01$, **$P<0.001$). **C.** Heatmap indicates the enrichment scores of tumor-specific variants which calculated as the odds ratios of the numbers of variants in each chromosome arm against those of outside the chromosome arm in each patient. The enrichment score less than 1 was truncated to zero.

doi:10.1371/journal.pone.0115152.g002

encoded prevalently the protein and ion transport-related functions (e.g. *KIFAP3*, *MIA3*, *STX6*, *TOMM40L*, *CACNA1E*, and *RYR2*). We also evaluated the spectrum of the non-tumoral variants. The non-tumor-specific variants included 147 nonsynonymous variants, 49 synonymous ones, and 6 codon stop-signaling ones appearing similar distribution of the tumor-specific variants (S2A Figure). However, the chromosomal enrichment at 1q was not observed in non-tumoral mutations, suggesting that the chromosome 1q enrichment is tumor-specific observation (S2B Figure).

Previously, the somatic variants have been reported to occur in association with specific neighbor sequences. For example, A>C transversions at AA dinucleotide were reported to be commonly found in esophageal cancer [31]. With respect to this, we examined the sequence patterns of the flanking regions of the mutated positions by applying sequence logo algorithm which visualize the position-specific probability of the sequence variations [32]. However, we could not observe any base prevalence of the flanking sequences (S3 Figure).

3. Tumor-specific mutations reveal functional enrichment

Next, we evaluated the functional enrichment of the tumor-specific mutations because the enriched mutations are likely to be functional. The tumor-specific variants were found in the 280 genes, while the non-tumoral variants were found in the 198 genes, respectively (S4A Figure). Overall, the genes with tumor-specific mutations showed functional enrichment of cytoskeleton organization, cell adhesion, ion and protein transport, and transcription-related genes (Table 1). Noticeably, the genes coding cytoskeleton organizations showed the most prevalent enrichment ($P=0.0059$), suggesting the enriched mutations of cytoskeleton genes play key roles in cancer development or progression. In addition, pathway analysis using KEGG database demonstrated a significant enrichment of cell cycle pathway ($P=0.005$) (S4B Figure). Such associations of the enriched mutations with cancer aggressiveness implies a pivotal role of tumor-specific mutations in cancer development and progression. The non-tumoral mutations also showed the enrichment of cell cycle-related genes (S3 Table), although the genes were different from the genes with tumor-specific mutations. Only the 29 genes were overlapped between the genes with tumor-specific variants (10%) and the genes with non-tumor-specific variants (14.5%) (S4C Figure). This may suggest that the cell cycle-related genes might be highly mutable regardless of tumor and non-tumor tissues even though different genes being involved. In addition, the chromatin regulators have been noticed to be frequently mutated in HCC [9, 11]. Congruently, we also found the mutations of chromatin regulation-related genes including *XRCC5*, *RAD51C*, *HIST1H2BD*, *SETD2*, *TTN*, *HIST1H2AL*, *BAZ2A*, *MYSM1*, and *SUPT6H*, but no statistical significance was observed by the functional enrichment analysis.

Table 1. Functional categories of the tumor-specific mutations.

Category	Term	Gene counts	EASE score (P-Value)
cytoskeleton organization	cytoskeleton organization	15	5.9×10^{-03}
	actomyosin structure organization	4	8.2×10^{-03}
cell adhesion	cell adhesion	19	1.7×10^{-02}
	calcium-dependent cell-cell adhesion	3	4.6×10^{-02}
cell cycle	hsa04110: cell cycle (KEGG)	8	5.5×10^{-03}
ion transport	calcium ion transport	8	5.4×10^{-03}
	transmembrane transport	16	2.4×10^{-02}
protein transport	vesicle-mediated transport	18	5.8×10^{-03}
	protein localization	23	1.2×10^{-02}
	intracellular transport	18	1.9×10^{-02}
	protein transport	19	3.6×10^{-02}
transcription initiation	transcription initiation from RNA polymerase II promoter	5	1.9×10^{-02}

doi:10.1371/journal.pone.0115152.t001

4. Differential tumor-specific variations of early and advance HCC are related with the transcriptional deregulation.

To evaluate the relationships between the mutations and transcriptional deregulations in HCC, we performed gene expression profiling with the same patients' samples. To address the functional association, we assigned the mutated genes and the gene expressions to the gene ontology (GO) terms. To predict the functional alteration by gene mutations, a total of 682 GO terms were selected which had at least one mutations in our dataset. Then, the functional enrichment of gene expressions for each patient was calculated as described in Methods. Strikingly, we observed that the expression levels of the mutated genes were significantly lower than those of the wild-type genes (Two-tailed Student's T-test, p=0.0006) (Fig. 3A). This may indicate the gene mutations might be the inactivation mutations rather than the activation mutations.

Next, we compared the differential gene expressions between early and advanced HCC subgroups. Of the 682 GO terms, a total of 102 GO terms with the 48 mutated genes were identified which had different enrichment scores ($P<0.05$) with greater than two-fold difference between the subgroups. Interestingly, the advanced HCC were significantly enriched with aggressiveness-related functions including cytoskeleton organization, development, proliferation, and cell migration-related genes, while the early HCCs were enriched with metabolism-related genes (Fig. 3B). Cancer-associated pathways including NGFR, TGFB, Hippo, and Wnt were also deregulated between the HCC subgroups, implying that the differential functions were involved in HCC progression. This result suggests that the different mutation profiles with concomitant transcriptional deregulation may contribute to the heterogeneous progression of the early and the advanced HCCs.

In addition, we sought to identify key regulators among the 48 mutated genes which are likely to associate with the subgroup-dependent mutations and concomitant gene expression. By applying GeneMANIA software implemented in Cytoscape plugin [33], we constructed a network using the genetic and physical
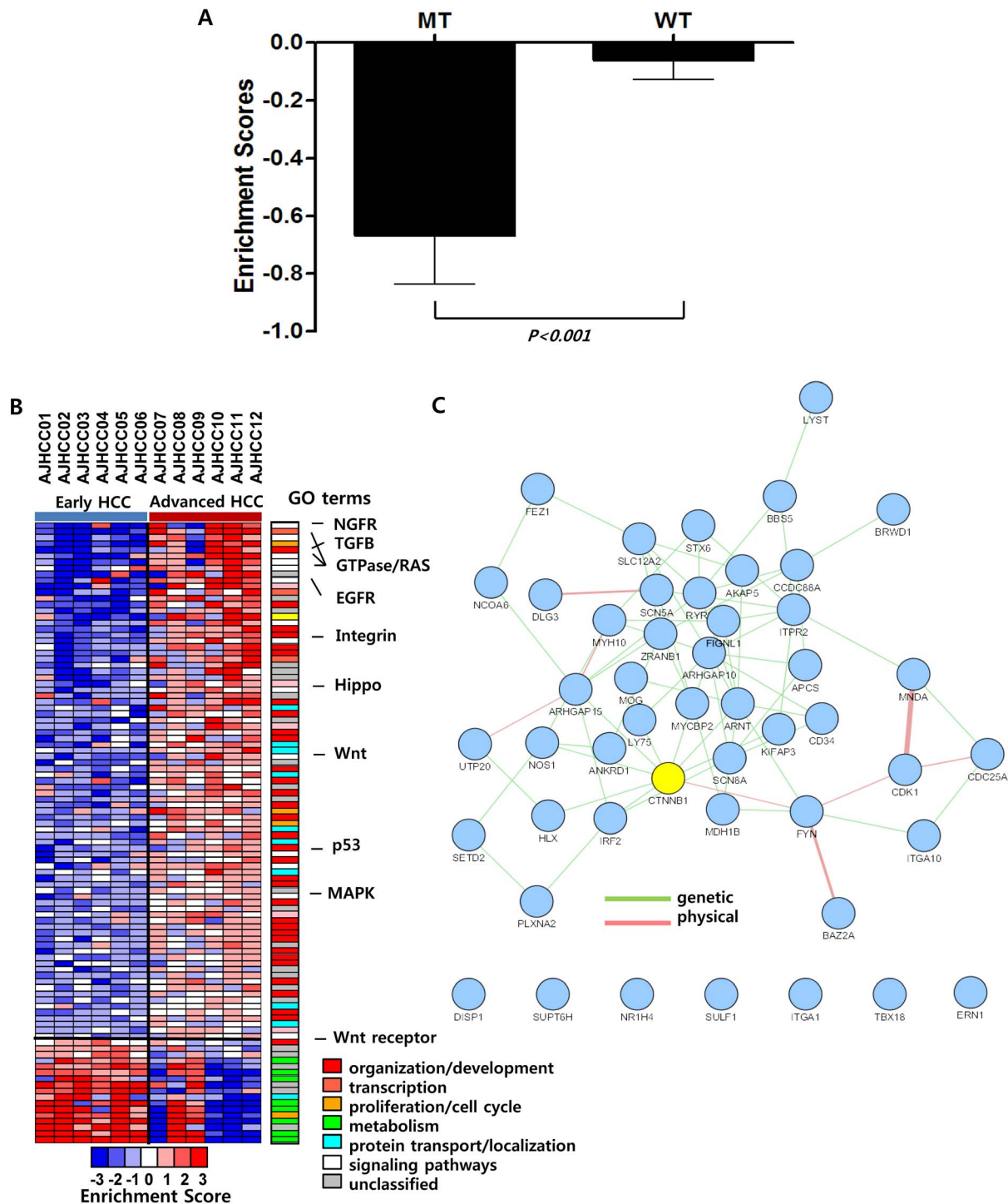
Fig. 3. Differential spectrum of tumor-specific variations are related with the transcriptional deregulation of the early and advanced HCC. A. The barplot shows the average enrichment scores of the gene expressions in the GO terms in each patient. The enrichment scores of the gene sets are calculated as described in Method. The GO terms of the genes with mutations and the ones without mutations are indicated as mutated type (MT) and wild type (WT), respectively. **B.** The heatmap shows the differentially enriched functions of mutated genes with deregulated expression. The categories with similar functions are indicated as a barplot of different colors (*right bar*). **C.** Network view of the 48 mutated genes with differential expression between the early and advanced HCC subgroups. The *CTNNB1* gene harbored the largest interaction partners indicated with yellow color.

doi:10.1371/journal.pone.0115152.g003

interactions among the 48 mutated genes. The mutated genes were found to be closely linked together suggesting that the mutations might be actionable in the differential development of HCCs (Fig. 3C). Particularly, *CTNNB1* gene, which is well known to play critical roles in HCC development and progression, had the largest number of interaction partners in the network. This may indicate that the mutation-derived disruption of *CTNNB1* and its pathways might play important roles in the progression of the HCC subgroups. This observation is concordant with the previous results that the HCCs with *CTNNB1* mutations were early staged and had a differentiated phenotype [34]. It has also been known that the *CTNNB1* mutations are associated with better prognosis of HCC [35], while the activation of Wnt pathways is correlated with poor prognosis [36].

Discussion

Although the NGS technologies could provide large scale profiles of genome-wide mutations in cancers during the past years, there are limitations in the assessment of functional implications of the mutation profile. Somatic mutations of cancers showed marked individual variations, therefore, it might be difficult to find key regulators with frequent mutations by analyzing the small sample sized data. Moreover, the action of the mutants can be diverse according to their base positions. The mutations from different base positions of the same gene can lead to different gene functions such as inactivation and activation mutations. This might be resulted from their effects on the physico-chemical properties and structures of wild-type proteins. Moreover, intra-tumoral heterogeneity may also contribute to false-negative calls in the tumors especially with low copy variations [23]. Besides, low sensitivity and specificity of the current high-throughput sequencing technologies may preclude the acquisition of accurate mutation profiles [4]. In addition, the erroneous mutation calls can be derived from base calling algorithms. For example, the genes encoding extremely large proteins were found to be frequently mutated, demanding further development of more accurate and significant base calling algorithms [37]. Considering these limitations of the mutation profiling studies, large sample and multiple assays might be necessarily required to get accurate information for predicting functional and clinical outcomes. Validation of the individual mutations with different methods might be required to overcome such limitations of NGS technology [38]. Notwithstanding such limitations of the mutation profiling studies, gene set-based analysis of the mutations such as gene set enrichment analysis could have successfully demonstrated their functional significance in cancers. For example, frequent mutations of chromatin regulators [8], Wnt/beta catenin, and Jak/STAT pathways [10] have been noticed in HCC, although the individual variants have low mutation rates. These results suggest that the assessment of the enriched mutations of a certain gene set rather than a single gene mutation is a useful approach to predict the functional role of the mutation

profiles on cancer progression. Of course, these mutations could be prioritized targets for cancer treatment.

In this study, we performed a mutation profiling of HCC by whole-exome sequencing analysis. Comparing the mutations and gene expression profiles between the early and the advanced HCC, we could observe the tumor grade-specific mutation patterns and their associations with the gene expression levels. Considering the high rates of false mutation calls of whole exome-seq data as described above, we applied stringent algorithms for variant calling.

In addition, our mutation set-based analysis could reveal new insights on the somatic mutations of HCC. By applying comparative and integrative approaches, we could reveal regional and functional patterns of the variants. Remarkably, we found that the chromosome 1q had frequent tumor-specific mutations particularly in the advanced HCC. Previously, genomic alterations at chromosome 1q have been noticed to have functional and clinical significance in the HCC progression. For example, recurrent copy number gains at 1q were observed to associate with the aggressive cancer behavior [39, 40]. This may indicate that the frequent genomic aberrations and instability at 1q including genome copy numbers and mutations contribute to the aggressive phenotype of HCC.

Functional enrichment analysis revealed the differential enrichment of mutations between the early and the advanced HCCs. The cytoskeleton and protein transport-related genes were enriched in the early HCC, while the cell migration and ion transport-related genes were enriched in the advanced HCC. These analyses successfully demonstrated that the differential mutations between the early and advanced HCC were well reflected at transcriptional level. Particularly, the mutations of ion-transport genes were largely located at chromosome 1q, which might be associated with the enriched mutations of 1q in the advanced HCC. Furthermore, it is interesting to find that the cancer-associated pathways including NGFR, TGFB, Hippo, and Wnt had the differential mutations and expressions between the HCC subgroups (see Fig. 3B). The integrative network analysis could reveal that the 48 gene mutations including CTNNB1, which might have pivotal roles in the heterogeneous progression of HCC subgroup.

In conclusion, our integrative analysis using exome-seq and gene expression profiles could provide genome-wide landscape of mutations in the early and the advanced HCC. By demonstrating the effects of the mutation sets on the transcriptional level, we could identify the potentially actionable tumor-specific mutations, which could be personalized and therapeutic targets for HCC diagnostics and treatment.

Supporting Information

S1 Figure. The number of mutations in each patient. The number of tumor-specific and non-tumor-specific mutations are plotted.
doi:10.1371/journal.pone.0115152.s001 (PDF)

S2 Figure. Mutation spectrum of non-tumor-specific mutations. Distribution of somatic non-tumor-specific variants in HCC. B. The heatmap indicate the enrichment scores of non-tumor-specific variants which calculated as the odds ratios of the numbers of variants in each chromosome arm against those of outside the chromosome arm in each patient. The enrichment score less than 1 was truncated to zero.
doi:10.1371/journal.pone.0115152.s002 (PDF)

S3 Figure. Probability of the flanking sequences of the tumor-specific mutation sites. Positional probability of the flanking sequences of the each of the reference and the mutated nucleotide base of the tumor-specific mutations are plotted by using Weblogo software.
doi:10.1371/journal.pone.0115152.s003 (PDF)

S4 Figure. Tumor-specific mutations at gene level were enriched with cell cycle-related genes. A. Ben-diagram show the number the tumor-specific and the non-tumor-specific mutations at gene level. **B.** The tumor specific mutated genes in cell-cycle pathway from KEGG are indicated by red color. C. Of the genes harboring the tumor-specific and the non-tumor-specific mutations, cell cycle-related genes are shown.
doi:10.1371/journal.pone.0115152.s004 (PDF)

S1 Table. Clinical and pathological features of HCC samples.
doi:10.1371/journal.pone.0115152.s005 (DOCX)

S2 Table. List of tumor-specific and non-tumor-specific variations in the early and the advanced HCC.
doi:10.1371/journal.pone.0115152.s006 (DOCX)

S3 Table. Functional enrichment of non-tumoral specific variants.
doi:10.1371/journal.pone.0115152.s007 (DOCX)

Author Contributions

Conceived and designed the experiments: JYC. Performed the experiments: JYC HGW SMK. Analyzed the data: HGW JYC ESP JSL HC. Contributed reagents/materials/analysis tools: JYC SWC SSK HJC SJA. Wrote the paper: HGW JYC.

References

1. **Woo HG, Wang XW, Budhu A, Kim YH, Kwon SM, et al.** (2011) Association of TP53 mutations with stem cell-like gene expression and survival of patients with hepatocellular carcinoma. Gastroenterology 140: 1063–1070.

2. **Laurent-Puig P, Legoix P, Bluteau O, Belghiti J, Franco D, et al.** (2001) Genetic alterations associated with hepatocellular carcinomas define distinct pathways of hepatocarcinogenesis. Gastroenterology 120: 1763–1773.

3. **Nakagawa H, Shibata T** (2013) Comprehensive genome sequencing of the liver cancer genome. Cancer Lett 340: 234–240.

4. Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, et al. (2011) High-resolution characterization of a hepatocellular carcinoma genome. Nat Genet 43: 464–469.

5. Huang J, Deng Q, Wang Q, Li KY, Dai JH, et al. (2012) Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. Nat Genet 44: 1117–1121.

6. Sung WK, Zheng H, Li S, Chen R, Liu X, et al. (2012) Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet.

7. Cleary SP, Jeck WR, Zhao X, Chen K, Selitsky SR, et al. (2013) Identification of driver genes in hepatocellular carcinoma by exome sequencing. Hepatology 58: 1693–1702.

8. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, et al. (2012) Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat Genet 44: 760–764.

9. Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, et al. (2012) Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet 44: 694–698.

10. Kan Z, Zheng H, Liu X, Li S, Barber TD, et al. (2013) Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. Genome Res 23: 1422–1433.

11. Li M, Zhao H, Zhang X, Wood LD, Anders RA, et al. (2011) Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. Nat Genet 43: 828–829.

12. Tao Y, Ruan J, Yeh SH, Lu X, Wang Y, et al. (2011) Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data. Proc Natl Acad Sci U S A 108: 12042–12047.

13. Ikeda A, Shimizu T, Matsumoto Y, Fujii Y, Eso Y, et al. (2014) Leptin receptor somatic mutations are frequent in HCV-infected cirrhotic liver and associated with hepatocellular carcinoma. Gastroenterology 146: 222–232 e235.

14. Wurmbach E, Chen YB, Khitrov G, Zhang W, Roayaie S, et al. (2007) Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. Hepatology 45: 938–947.

15. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

16. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.

17. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498.

18. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164.

19. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57.

20. Woo HG, Lee JH, Yoon JH, Kim CY, Lee HS, et al. (2010) Identification of a cholangiocarcinoma-like gene expression trait in hepatocellular carcinoma. Cancer Res 70: 3034–3041.

21. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, et al. (2011) Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 29: 908–914.

22. Wang Q, Jia P, Li F, Chen H, Ji H, et al. (2013) Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med 5: 91.

23. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, et al. (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 366: 883–892.

24. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, et al. (2012) Mutational processes molding the genomes of 21 breast cancers. Cell 149: 979–993.

25. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, et al. (2013) Signatures of mutational processes in human cancer. Nature 500: 415–421.

26. Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B (2014) Synonymous mutations frequently act as driver mutations in human cancers. Cell 156: 1324–1335.

27. **Gartner JJ, Parker SC, Prickett TD, Dutton-Regester K, Stitzel ML, et al.** (2013) Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. Proc Natl Acad Sci U S A 110: 13481–13486.

28. **Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, et al.** (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet Chapter 10: Unit 10 11.

29. **Huang H** (1999) Beta-catenin mutations are frequent in human hepatocellular carcinomas associated with hepatitis C virus infection. Am J Pathol 155: 1795–1801.

30. **Thorvaldsdottir H, Robinson JT, Mesirov JP** (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14: 178–192.

31. **Dulak AM, Stojanov P, Peng S, Lawrence MS, Fox C, et al.** (2013) Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nat Genet 45: 478–486.

32. **Crooks GE, Hon G, Chandonia JM, Brenner SE** (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190.

33. **Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, et al.** (2010) GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics 26: 2927–2928.

34. **Calvisi DF, Factor VM, Loi R, Thorgeirsson SS** (2001) Activation of beta-catenin during hepatocarcinogenesis in transgenic mouse models: relationship to phenotype and tumor grade. Cancer Res 61: 2085–2091.

35. **Hsu HC, Jeng YM, Mao TL, Chu JS, Lai PL, et al.** (2000) Beta-catenin mutations are associated with a subset of low-stage hepatocellular carcinoma negative for hepatitis B virus and with favorable prognosis. Am J Pathol 157: 763–770.

36. **Inagawa S, Itabashi M, Adachi S, Kawamoto T, Hori M, et al.** (2002) Expression and prognostic roles of beta-catenin in hepatocellular carcinoma: correlation with tumor progression and postoperative survival. Clin Cancer Res 8: 450–456.

37. **Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, et al.** (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499: 214–218.

38. **Kwon SM, Cho H, Choi JH, Jee BA, Jo Y, et al.** (2012) Perspectives of integrative cancer genomics in next generation sequencing era. Genomics Inform 10: 69–73.

39. **Woo HG, Park ES, Lee JS, Lee YH, Ishikawa T, et al.** (2009) Identification of potential driver genes in human liver carcinoma by genomewide screening. Cancer Res 69: 4059–4066.

40. **Poon TC, Wong N, Lai PB, Rattray M, Johnson PJ, et al.** (2006) A tumor progression model for hepatocellular carcinoma: bioinformatic analysis of genomic data. Gastroenterology 131: 1262–1270.