Genomics 105 (2015) 76–82

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Genome-wide identification of RNA editing in hepatocellular carcinoma

Lin Kang [a], Xiaoqiao Liu [b], Zhoulin Gong [a,c], Hancheng Zheng [a,c], Jun Wang [a,d,e], Yingrui Li [a,c], Huanming Yang [a,d,f], James Hardwick [g], Hongyue Dai [g], Ronnie T.P. Poon [h], Nikki P. Lee [h], Mao Mao [i,j,*], Zhiyu Peng [a,k,*], Ronghua Chen [b,l,*]

[a] BGI-Shenzhen, Shenzhen, China
[b] Scientific Informatics, MSD R&D (China), Beijing, China
[c] BGI-Tech, BGI-Shenzhen, Shenzhen, China
[d] King Abdulaziz University Jeddah, Saudi Arabia
[e] Department of Biology, University of Copenhagen, Copenhagen, Denmark
[f] James D. Watson Institute of Genome Science, Hangzhou, China
[g] Merck Research Laboratories, Merck & Co., Inc., Boston, MA, USA
[h] Department of Surgery, The University of Hong Kong, Hong Kong
[i] Pfizer Oncology, San Diego, CA, USA
[j] Asian Cancer Research Group, Inc., Wilmington, DE, USA
[k] Guangzhou Key Laboratory of Cancer Trans-Omics Research, BGI-Guangzhou, Guangzhou, China
[l] Scientific Informatics, Merck & Co., Inc., Boston, MA, USA

ARTICLE INFO

Article history:
Received 5 May 2014
Accepted 14 November 2014
Available online 25 November 2014

Keywords:
Hepatocellular carcinoma
RNA-editing
RNA-Seq

ABSTRACT

We did whole-transcriptome sequencing and whole-genome sequencing on nine pairs of Hepatocellular carcinoma (HCC) tumors and matched adjacent tissues to identify RNA editing events. We identified mean 26,982 editing sites with mean 89.5% canonical A → G edits in each sample using an improved bioinformatics pipeline. The editing rate was significantly higher in tumors than adjacent normal tissues. Comparing the difference between tumor and normal tissues of each patient, we found 7 non-synonymous tissue specific editing events including 4 tumor-specific edits and 3 normal-specific edits in the coding region, as well as 292 edits varying in editing degree. The significant expression changes of 150 genes associated with RNA editing were found in tumors, with 3 of the 4 most significant genes being cancer related. Our results show that editing might be related to higher gene expression. These findings indicate that RNA editing modification may play an important role in the development of HCC.

1. Background

RNA editing is a post-transcriptional event that modifies the genetic information from the genome and is most catalyzed by adenosine deaminase acting on RNA (ADAR) family enzymes. By modifying coding (CDS) regions or splice sites, RNA editing can induce amino acids changes, while affecting translation via regulation of mRNA retention in the nucleus [1]. The variations in RNA editing increase genetic diversity in individuals and tissues [2,3], and its potential role in cell physiological processes might affect disease development [4–6].

The potential links between RNA editing and cancer also intrigue many researchers. Global hypoediting of Alu elements was found in several tumors, and reduced editing in *MED13* was discovered in brain tissue [7]. Shah et al. found two new RNA editing events altering amino acid sequences of *COG3* and *SRP9*, which showed a possible important role of RNA editing in breast cancer [8]. More RNA editing events in the CDS region have been observed in seven different tissues [9]. Other studies highlighted the correlations between ADARs and cancer stages, and identified dysregulation of RNA editing in cancer progression [10,11]. Possible contribution of A-to-I RNA editing to carcinogenesis was also reported [4].

Hepatocellular carcinoma (HCC) is the third leading cause of cancer deaths worldwide. An increase of a recoding RNA editing in *AZIN1* was found in some HCC cases, which might influence cell proliferation in tumorigenesis; however, only a single editing site was analyzed in this study [12]. Recently new high-throughput sequencing technologies allow researchers to more thoroughly investigate RNA editing events

Abbreviations: HCC, (hepatocellular carcinoma); ADAR, (adenosine deaminase acting on RNA); NGS, (next-generation sequencing); CDS, (coding DNA sequence); UTR, (untranslated region); RPKM, (reads per kilobase per million mapped reads); RNA-Seq, (next generation sequencing of RNA).

* Corresponding authors.
E-mail addresses: kanglinxm@gmail.com (L. Kang), xiao.qiao.liu@merck.com (X. Liu), gongzhuolin@bgitechsolutions.com (Z. Gong), zhenghch@genomics.cn (H. Zheng), wangj@genomics.org.cn (J. Wang), liyr@genomics.org.cn (Y. Li), yanghm@genomics.cn (H. Yang), james_hardwick@merck.com (J. Hardwick), hongyue_dai@merck.com (H. Dai), poontp@hkucc.hku.hk (R.T.P. Poon), nikkilee@hku.hk (N.P. Lee), mao_m@yahoo.com (M. Mao), pengzhiyu@genomics.cn (Z. Peng), ronghua_chen@merck.com (R. Chen).

http://dx.doi.org/10.1016/j.ygeno.2014.11.005
0888-7543/

at a whole-transcriptome level and a significant number of new RNA edits was identified [9,13–15]. These studies revealed that the prevalence of RNA editing in human transcriptome might be greatly underestimated. While these studies improve our understanding of the scale of RNA editing events, few studies have investigated the functional RNA edits in human diseases at a genome-wide scale. Thus, deeply sequencing whole genomic DNA and transcriptome in disease tissues with advanced bioinformatics analysis might provide biological insight into the relationship between potentially functional RNA edits and human disease.

Here we sequenced the transcriptome and whole genome of nine pairs of HCC tumors and matched liver tissues for RNA editing analysis to fully investigate the global RNA editing events in HCC. We identified a large number of RNA editing events ranging from 10,262 to 47,753 using our improved RNA editing identification pipeline. We then focused on the difference of RNA editing between normal and tumor samples to uncover the possible ways that RNA editing events shape biological processes. Our results showed 292 edits with significant differences in editing degree between tumor and adjacent normal tissues, 11 specific edits located in the CDS region, plus significant expression changes of 114 genes associated with RNA editing.

2. Results

2.1. RNA sequencing for 9 HCC patients

We used samples from a previous HCC study [16], which sequenced whole tumor and adjacent non-tumor liver genomes at least at 30× coverage from 88 HCC patients who underwent curative hepatectomy. We randomly selected 9 HCC patients from 88 cases and 18 high-quality RNA samples (paired tumor tissue and adjacent normal tissue, respectively) were carried on to transcriptome sequencing. The clinicopathological data for those patients are listed in Supplementary Table S1. For the 18 RNA samples, an average of 146.6 million reads or 13.19G RNA-Seq sequencing bases were generated with 90 bp read length (paired-ends) and 200 bp expectation insertion size (Supplementary Table S2). The alignment rate (against genome reference and junction database, see Materials and methods) for those data ranged from 78.73% (sample 11T) to 85.44% (sample 65N), with an average mapping rate of 81.92% and 81.13% for normal and tumor samples, respectively. Supplementary Table S3 shows the digital gene expression for each sample based on RNA-Seq data.

2.2. RNA editing callings for 9 HCC patients

The RNA editing sites calling pipeline was first developed by Peng et al. [17] and further refined by adding bioinformatics filters for this study (see Materials and methods). This refined pipeline had been applied to YH poly(A) + RNA sequencing analysis. The sensitivity was improved by identifying 88,059 RNA editing sites, compared with 11,467 sites by the previous version, with similar proportion of A → G editing type (91.5% vs. 90.2%).

In total, we identified 485,684 editing sites from 18 samples, ranging from 10,262 to 47,753 in each sample, with an average of 26,982 sites per sample. The percentages of A → G editing ranged from 85.02% to 94.38% (average 89.51%). The T → C editing sites identified by a non-strand specific RNA-Seq protocol are possibly wrong strand-defined A → G editing sites. We combined A → G and T → C editing sites and the combined percentages were increased to an average 96.73% (Supplementary Table S4), indicating high performance of our RNA editing identification pipeline. Detailed attributes of editing sites by annotation for each sample are shown in Supplementary Table 5 (Supplementary Table S5). Due to low validation rate and lack of strong evidence on the existence of non-A → G types of editing reported in previous studies [17,18], we excluded non-A → G RNA editing sites from further analyses. We noticed that higher numbers of editing sites were identified in tumor samples than in normal samples (Supplementary Table S4). To exclude possible sequencing and/or expression difference among samples, we calculated editing rate for each sample (Fig. 1). Consistently, there was a higher editing rate in tumor samples in addition to more editing sites ($P = 0.048$, paired t-test; $P = 0.039$, Wilcoxon signed-rank test). Moreover, we compared expression of the ADAR enzyme family between tumor and normal samples and found that ADAR1 (or *ADAR*) was expressed significantly higher in tumors ($P = 1.2 \times 10^{-3}$, Fig. 2), which is in line with previous observation [19]. Finally, overexpression of ADAR2 (or *ADARB1*) and ADAR3 (or *ADARB2*) was found in tumor samples, but without significant difference from normal samples ($P = 0.34$ and 0.28, respectively). This increased expression of ADAR family proteins may explain why more RNA editing events occur in tumor samples than in corresponding normal samples. Previous cancer studies have also reported higher levels of ADAR enzymes within cancer samples [20,21] and reduced expression of ADAR3 in brain tumor samples [7].

2.3. Normal- and tumor-specific editing

Editing in coding regions may change amino acids later in the translation process, and thus may influence biological processes in a direct way. We first focused on the tissue specific editing sites, which were only edited in one tissue but not edited in the adjacent tissue (see Materials and methods) in the CDS region (annotation retrieved from RefSeq). Altogether, we found 101 edits in the CDS region, 68 of which were non-synonymous changes, and two of which may induce stop codon disruption (Supplementary Table S6). Among the edits in the CDS region, only 11 showed tissue specific editing. More precisely, 5 were tumor-specific edits and 6 were normal-specific edits. Seven of them were non-synonymous changes, but no recurrent specific edits were found in our samples (Table 1). The low number of specific edits and recurrent edits found in the CDS region might be due to the sample size limitation and is consistent with previous studies [17,18,22–24]. For the nucleotide substitution at the genomic DNA level, we also hypothesized that higher somatic mutation at the genome level may cause/suffer more RNA editing events, but the low correlation (0.1477, Pearson's r) found between the number of RNA edits and the number of somatic mutations or SNVs ruled out this hypothesis.

Further, a huge proportion of RNA editing sites is identified in non-coding regions, UTR and introns. In addition to the CDS specific editing, a total of 2582 tumor-specific and 2947 normal-specific editing sites were found in the specific editing sites in UTR and introns (Supplementary Table S7). Of these sites, 218 tumor-specific and 245 normal-specific editing sites were specifically found in more than one sample pair (Supplementary Table S8) and dispersed among 233 genes. According to the annotation of these genes, we found an enrichment in catabolic processes such as protein maturation, blood coagulation and immune or stimulus responses (Supplementary Table S9), as well as in the apoptosis signaling pathway ($P = 2.84 \times 10^{-2}$). These enrichments indicate that RNA editing could play a role in RNA stability or other modifications in cancer progression.

In light of elevated editing events found in tumor samples, we expected to uncover more tumor-specific edits. There were 166,386 sites in total after sample merging and discarding sites with heterozygous DNA in any sample (Supplementary Table S7). Among them, we found only 5529 (3.33%) specific editing sites in either tumor or normal samples, and surprisingly there were more normal-specific sites (2947) than tumor-specific sites (2582). Upon checking overlaps of editing sites between normal and tumor sets (after merging editing sites into normal sets and tumor sets, respectively), we found that 112,727 sites covered at least 10 reads in either the tumor or normal set, 7006 edits (6.22%) were at positions with total number of reads mapped in all normal samples less than two (the minimal depth requirement for edits detection in our method), and only 176 (0.16%) edits were at positions in which tumor samples showed less than two covered reads

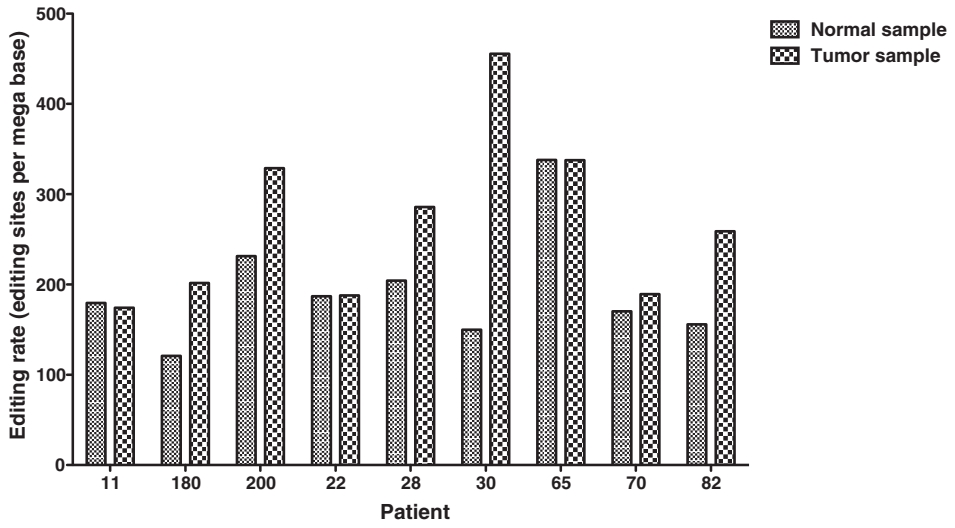

Fig. 1. Overall editing rate in each sample. Editing rate was calculated by number of RNA editing sites divided by total length of sequencing (minimal covered depth ≥ 2) in each sample.

(Supplementary Fig. S1). One possible explanation is that the "over-expressed" regions allow us to detect more RNA editing sites, or vice versa, that RNA editing causes the "over-expression" in tumors. To include those positions "unexpressed" in normal samples but with editing in tumor samples, we then modified the definition of a specific editing site to allow specific edits to occur in one sample with minimal editing degree of 10% and minimal depth of 10, regardless of the coverage of its corresponding pair. And under this definition, 6764 sites were classified to tumor-specific and 3,643 sites were normal-specific editing.

2.4. Comparison of editing degree between tumor and normal tissues

Common edits shared in all normal and tumor samples with high degree (>20%), such as those in *AZIN1*, *ZNF669* and *COPA*, may form a connection between editing and functional progress in the liver. For instance, *AZIN1* was reported to have increased editing in tumor samples and might result in tumor-initiating potential and aggressive behavior [12]. We further surveyed the difference of editing degree between normal and tumor samples from the same patient.

There were 292 sites that showed significant differences between tumor and normal groups ($P < 0.01$, paired *t*-test, see Materials and methods, Supplementary Table S10). Among the 292 edits, 163 showed "tumor preferred editing" (editing degree in tumor is greater than in normal), and 129 showed "normal preferred editing" (editing degree in normal is greater than in tumor). 13 cancer related genes (oncogene or tumor suppressor) with 18 editing sites and one gene with editing in the CDS region showed significant differences in editing degree ($P < 0.01$, Supplementary Table S11). Two oncogenes, *EIF2AK2* (eukaryotic translation initiation factor 2-α kinase 2) and *HNF4A* (hepatocyte nuclear factor 4, α) were presented in more than one site with different editing degrees (3 and 4 respectively, Supplementary Table S11). Finally, *CTSB*, which has been previously reported to be associated with esophageal adenocarcinoma and other tumors [25–29], showed a significant editing degree difference between tumor and normal tissues ($P = 4.17 \times 10^{-3}$, chr8:11701458, in 3-UTR region). Annotation of these genes with different editing degrees uncovers the enriched pathways in immune or virus response, activation of MAPK activity, membrane organization, etc. (Supplementary Table S12).

We also noticed that there was a slight higher editing degree in tumors than in normal samples in our data. More precisely, among the 42,191 sites covered at least by 4 sample-pairs, 22,078 editing sites show "up-regulation" (average of editing degrees in tumor sample higher than normal sample) while 19,907 show "down-regulation" (higher degree showed in normal sample than tumor sample). This is consistent with the finding of over-expression of ADARs enzymes in tumor samples, but how the enzymes function and how they change the cancer progress remain unknown.

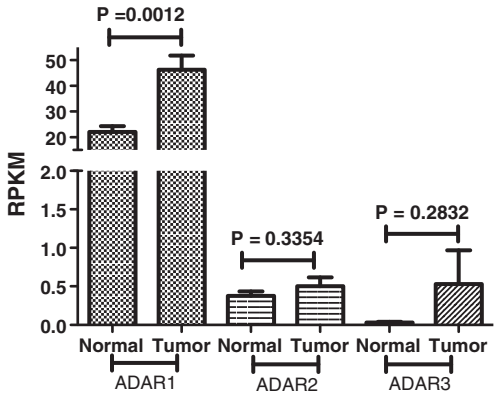

Fig. 2. Comparison of gene expression between normal and tumor samples in ADAR family. The most important and expressed RNA editing associated enzyme in humans, ADAR1 showed significant expression difference between normal and tumor groups ($P = 0.0012$, paired *t*-test). RPKM mean/SEM bar is shown in figure.

2.5. Relation between editing and gene expression

One possible role of editing on RNA is to affect or serve as a target for regulator factors such as miRNA, and thus may cause differential gene expression. Therefore, we checked the expression in tumor samples between two groups of genes carrying any editing site (edit$^+$, with minimal number of read shows edit bases: 2 and minimal editing degree: 10%) or not (edit$^-$) to investigate the potential correlation between RNA editing and expression in HCC. We found 150 genes with significant differences in expression (*t*-test, $P < 0.01$, Supplementary Table S13). The top 4 genes with significant alternation in expression are shown in Fig. 3. Among the 4 genes, 3 of them are linked with cancer (*CASP2*, *MAFK* and *ULK2*). *CASP2* may function in stress-induced cell death pathways, cell cycle maintenance, and the suppression of tumorigenesis; *MAFK* is v-maf musculoaponeurotic fibrosarcoma oncogene homolog K (avian); *ULK2* is a protein kinase involved in autophagy in response to starvation. We also found enrichments of these genes in the insulin signaling pathway and the B cell receptor signaling pathway (Supplementary Table S14). Interestingly, all top 6 alterations in

Table 1
Normal- or tumor-specific RNA editing sites in CDS.

Type	GeneSymbol:Chr:Pos	Amino acid change	No.	Patient ID	Gene name
NS	COPA:chr1:160302244[a]	ATT → GTT; Ile → Val	1	180	Coatomer protein complex, subunit α
NS	ZNF587B:chr19:58355651	AAG → GAG; Lys → Glu	1	70	Zinc finger protein 587B
NS	LYRM4:chr6:5187135	AAG → AGG; Lys → Arg	1	200	LYR motif containing 4
TS	NCSTN:chr1:160319987	AGT → GGT; Ser → Gly	1	200	Nicastrin
TS	IL12RB1:chr19:18182135	ATC → GTC; Ile → Val	1	22	Interleukin 12 receptor, β 1
TS	TMEM230:chr20:5092186[a]	CAC → CGC; His → Arg	1	28	Transmembrane protein 230
TS	FLNB:chr3:58141791[a]	ATG → GTG; Met → Val	1	82	Filamin B, β (actin binding protein 278)

[a] Found in previous studies.

expression show higher expression in the edit$^+$ group, When we then checked all 150 genes, 145 showed higher expression in the edit$^+$ group while the rest 5 showed higher expression in the edit$^-$ group. To eliminate the possible bias of less RNA edits due to low expression genes, we excluded the low expression genes with RPKM (reads per kilobase per million mapped reads, see Materials and methods) less than 10 in both groups and 29 genes were remaining. Among the 29 genes, 27, or 93.10%, showed higher expression while only 2, or 6.90%, showed lower expression in the edit$^+$ group. Since the RNA editing may affect the stability of RNA by influencing the binding effectiveness of miRNA or other regulators, this finding indicates that RNA editing may enhance the stability of RNA in most cases.

Same analyses were also carried out in normal samples, and we found 75 genes with significant differences in expression (Supplementary Table S15). Annotation of those genes is shown in Supplementary Table S16. Interestingly, only one gene, *THOC5*, was shared between significant gene lists in tumor and normal samples, and it might indicate that RNA editing changes gene expression on different genes in normal and tumor tissues. The same tendency for higher expression (62 out of 75 genes) in the edit$^+$ group compared to the edit$^-$ group was also found in normal samples.

We also calculated the correlation between the ratio of editing degrees and the ratio of expression levels in normal/tumor samples (see Materials and methods). A total of 361 edits showed high correlation (|Pearson's r| >0.9) in 278 genes, with 240 out of 361 having positive correlation and 121 having negative correlation (Supplementary Table S17). Annotations of these 278 genes are shown in Supplementary Table S18. Mainly positive correlation between editing degree ratio and gene expression ratio may suggest a trend of increasing gene expression through RNA editing, but we should note the possible bias of edit detection in low expression genes or regions.

3. Discussions

Recently, a large amount of RNA editing events has been reported in human transcriptome. This, in addition to its potential contribution to genetic diversity, emphasizes the importance of RNA editing characterization. Here, we surveyed RNA editing events across 9 HCC patients using a revised RNA editing identification method to compare the difference between normal and tumor samples, and presented a comprehensive RNA editing landscape in relation to HCC. The large quantity and high percentage of A → G RNA editing acknowledged the well-

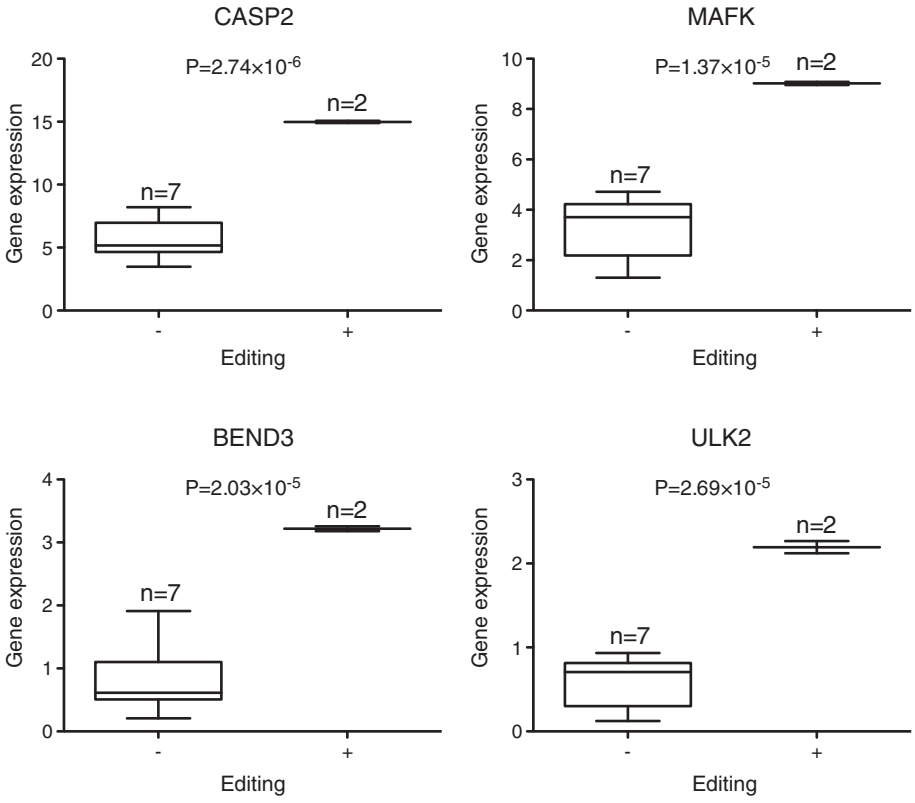

Fig. 3. RNA editing and gene expression. Gene expression is presented as RPKM. "+"/"−" indicates harboring editing site or not in a given sample, and samples were grouped to either edits$^+$ or edits$^-$; *P*-values of *t*-test are shown.

designed bioinformatics approach, which has been tailored to cancer research.

A robust method of RNA editing event identification is basic and important for further RNA editing analysis. The number of edits detected in a single sample varies from 5695 to 395,943 in recent studies using NGS technology [17,18,22–24], but from our revised RNA editing identification pipeline we found an average of 26,982 edits, 89.51% in A → G percentage, in 9 HCC patients. Comparing edits found in this study with the RNA editing database DARNED [30], only 38,939 or 23.40% was included in DARNED. This demonstrates the individual diversity in RNA editing events and also indicates that there is still large proportion of RNA undiscovered in humans.

The finding that RNA editing in *AZIN1* plays an important role in HCC tumorigenesis [12] enhances our view of RNA editing as a potential driver in HCC. Three patients (11T, 200T and 82T) showed higher editing degree in tumors in our data (see Supplementary Table S7), while the other 6 patients showed higher editing degree in normal samples. Two editing sites inducing alternation of amino acids in *COG3* and *SRP9* were reported in a previous breast cancer study [8]. We also identified an editing site in COG3 in our HCC study (chr13:46090371, Ile → Val). However, no significant difference was found between normal and tumor samples ($P = 0.95$), and no editing site was found in *SRP9* in our data. Some previously reported genes related to RNA editing and brain or neuron disorders (e.g.: *GRIA2*, *GRIK2*, *FLNA*, *GABRA3*, *CYFIP2*, *HTR2C*, *KCNA1* etc.) were rarely found in our data (only two sites in intron of GRIK2 showed editing among aforementioned genes). This might indicate diversity of editing events among tissues or individuals, and may also point to important RNA editing roles in development that we have yet to discover. However, another gene *FLNB*, filamin B β, which encodes a member of the filamin family of proteins and is associated with many diseases, such as Larsen's syndrome and pancreatitis [31,32], was found with an RNA editing event (in codon #2293, Met2293Val) in all 9 tumor samples and 7 normal samples (except 22 and 82) and showed tumor-specific editing in patient 82. The average editing degree among 9 tumor samples of 28.88% is also slightly higher than the average degree of 24.66% in 7 normal samples, although no significant *P*-value was found in our current data ($P = 0.59$). Patient 22 did not count for tumor-specific editing due to the low depth (9 reads) in its normal sample, which did not meet our criteria of specific edits (minimal depth requirement 10, see Materials and methods). Interestingly, this codon #2293 editing site was also reported in a breast cancer sample [22] and may indicate its prevalence in cancer development. Another tumor-specific editing gene found in patient 22, *IL12RB1*, was reported to be related to virulent mycobacterium tuberculosis [33], but it was predicted to be a tolerated mutation in SIFT analysis [34].

Consistent with previous studies, our results did not show many editing events in the CDS region (101 or 0.06%), which may directly change amino acids and be considered a more important function. Moreover, even less specific editing and significantly different editing degree between normal and tumor samples ($P < 0.01$) was noted in the CDS region (42 and 2, respectively). This may indicate that RNA editing events act more often in an indirect way. The common editing events found in all samples, even without significant differences, may also play a potential functional role in the liver, as well as other tissues, or in a temporary way. We compared our data of RNA editing to the COSMIC point mutation database (v62, http://cancer.sanger.ac.uk), and only one site was shared among the two sources (chr1:145100740, intron of *SEC22B*). This site showed editing in 3 samples (200T, 22T and 70T) exclusively in tumor stages, but they are not categorized as tumor-specific edits due to the low coverage in sequencing depth in both tumor and normal samples. More intensive validation could be done on such a set of potential candidates. In regards to RNA editing in non-coding regions, our data also showed the significant RNA editing discrepancies in non-coding regions between normal and tumor samples in HCC (especially in those cancer correlated genes). The disease associated SNPs are enriched in non-coding functional elements, according to the newest ENCODE (Encyclopaedia of DNA Elements) group's research [35], and thus our findings could be used to better and more fully investigate RNA editing function.

4. Conclusions

In this study, we surveyed RNA editing events across 9 HCC patients and compared the difference between normal and tumor samples using a revised RNA editing identification method. We were able to report a comprehensive RNA editing landscape in relation to HCC. Massive differentials in RNA editing events found in normal and tumor and the correlation between editing and expression may indicate an important role of RNA editing in HCC. This study provides a way to quickly scan the RNA editing in tumor samples at a whole-genome scale. More studies on experimental validation and biological assays, as well as application of other technologies like ultra-high throughput sequencing should also be carried out while we focus on particular and smaller regions. Based on our knowledge, this is the first study to survey RNA editing events at a whole genome level in cancer, and thus provides a comprehensive view of the role of RNA editing on HCC.

5. Materials and methods

5.1. Samples

We sequenced 18 RNA samples (tumor tissue and adjacent non-tumor tissue) from 9 Chinese individuals diagnosed with HCC and who underwent curative hepatectomy at Queen Mary Hospital as previously described [16]. Approval for the use of clinical specimens for research was obtained from Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB). All subjects gave written informed consents to use liver tissues for research as previously described [16].

5.2. RNA preparation, library construction and RNA-Seq

Total RNA isolated with TRIzol reagent was treated with RNase-free DNaseI (New England BioLabs) at 37 °C for 10 min. The Dynabeads mRNA Purification Kit (Life Technologies) was used to isolate mRNA from the total RNA samples.

For library construction and RNA-Seq, both types of experiments were performed based on the manufacturer's instructions (Illumina Inc., USA). Libraries were prepared according to the Illumina's protocol (Preparing Samples for Sequencing of mRNA, Part #1004898, Rev. A). Poly(A)$^+$ RNA was isolated using the oligo(dT) beads (Dynabeads mRNA Purification Kit; Invitrogen, Cat. #610-06). Libraries were sequenced as paired-end 90-bp sequence tags using the standard Illumina pipeline and sequenced by Illumina HiSeq™ 2000.

5.3. RNA editing calling pipeline

We updated our method previously described in Ref. [17], combined with the pipeline mentioned in Ref. [18]. We modified the previous mapping procedure by including sequences that surround all currently known splicing junctions from Gencode (http://www.gencodegenes.org/, release 16), RefSeq (http://www.ncbi.nlm.nih.gov/refseq/, release 55), UCSC (http://genome.ucsc.edu/) and Ensembl (http://www.ensembl.org/, release 71) as targets in the mapping procedure. The paired-end reads obtained from HiSeq™ 2000 were aligned to the reference genome and junction database using the SOAP2 [36] program (with parameters: *-m* 0 *-x* 10,000 *-s* 40 *-l* 35 *-v* 5 *-r* 0). PCR duplicates were removed using Samtools [37] with default parameters, and RNA-centric SNVs were identified from pileup result by keeping all sites with discrepancies between cDNA and reference. The new workflow is described in Supplementary Fig. S2. Improved accuracy and sensitivity

in YH sample, which is indicated by higher percentage of A → G edits, were shown in Supplementary Table S19.

The initially identified SNVs were then filtered by the following quality-aware steps: 1) RNA–DNA variants filter, 2) genome variants filter: we required corresponding genome position to be strict homozygous in genotype, 3) read parameter filter, 4) MES filter; it is to remove misaligned reads that arise from mapping error inherent to the mapping algorithm (MES). As described in Ref. [17], the MES set was generated as follows: read sequences were simulated based on all human genes using MAQ without mutation (−*r* parameter). The simulation reads were input to the RNA editing pipeline, and the resultant collection of RNA-centric SNVs is termed MES and represents an inherently error-prone set of sites that are incorrectly called owing to the nature of mapping and/or calling algorithms. Any SNVs that matched the MES were then removed. The last step of the initial filtering is 5) known SNPs filter. All filters except the read parameter filter were kept the same as we described in our previous method, and the parameters used in the read parameter filter were set as follows: distance cutoff = 5 (*m*, the minimal distance of an SNV site to its supporting reads' ends); quality score cutoff = 20 (*q*, minimal sequencing quality score of SNV-corresponding nucleotide); and supporting reads number cutoff = 2 (*n*, minimal number of reads that support the variant and that meet the above two cutoff parameters).

All SNVs that passed through the above filters were chosen as candidate edits. Subsequently, they were divided into two sets — *Alu* candidates and non-*Alu* candidates. We did not apply any further filtering steps for *Alu* candidates and considered them as true editing sites, while non-*Alu* candidates were subject to the following strict filters: 1) strict read parameter filter, 2) simple repeat filter, 3) BLAT filter, and 4) editing degree filter. We set $m = 8$, $q = 20$ and $n = 3$ in the strict read parameter filter, then removed sites if they matched the simple repeat annotation in the simple repeat filter. BLAT (v. 35) was used to search for SNVs whose supporting reads are multiply aligned to the reference genome with the same mismatch tolerance used in the SOAP2 alignment. We discarded all supporting reads with more than one hit within the mismatch tolerance. Subsequently, we filtered SNVs that were supported by less than two reads or if less than half of the reads remained after the BLAT procedure. Finally, sites with extreme degree of variation were excluded in the editing degree filter as previously described. Non-*Alu* candidates that passed through all above filters were combined with *Alu* edits for further analysis.

The reference genome version used in this study is hg19. Gene information from RefSeq and Ensembl was used for strand annotation of RNA editing events, and thus the editing types were assigned.

5.4. Specific edits

In a given sample-pair (paired tumor sample with adjacent normal control from same patient), a tumor-specific editing site was only found in tumor samples with a minimal editing degree of 10% and minimal edit bases (carried edited nucleotide) of 2, while paired normal samples were covered at a minimal depth of 10 without carrying any edit base. A normal-specific editing site was only found in normal samples with a minimal editing degree of 10% and minimal edit base (carried edited nucleotide) of 2, while paired tumor samples were covered at a minimal depth of 10 without carrying any edit base. An editing site ambivalently marked as both tumor-specific and normal-specific in 9 patients was deleted from specific edits list and thus excluded from further count and analyses.

5.5. Editing rate

Editing rate was calculated as the total number of edits divided by the coverage length (in Mega base) for a given sample.

5.6. Editing degree

Degree of editing for a given site was calculated as the ratio of reads supporting the edit base to the total number of reads covering the site.

5.7. Editing degree comparison

Editing sites with a minimal coverage of 2 in both tumor and normal samples in sample-pairs were taken into degree calculation. At least 4 sample-pairs conducting degree calculations were carried on to degree comparison, and a paired *t*-test was used to estimate *P*-value.

5.8. Gene expression

Gene expression was calculated using RPKM (reads per kilobase per million mapped reads) from RNA-Seq data. The formula to calculate RPKM is, $\text{RPKM} = (\text{number of mapping reads}) \times 10^3 \times 10^6 / [(\text{length of transcript}) \times (\text{number of total reads})]$ [38].

5.9. Correlation between ratio of editing degrees and ratio of expression levels

For a given editing site, we calculated the ratio of RNA editing degree of normal/tumor for each sample-pair, and the expression ratio of normal to tumor gene harboring, such as editing site. Sample-pairs with covered reads lower than 2 in either normal or tumor sample were excluded, and edits with more than 3 excluded sample-pairs were eliminated from this analysis. Pearson's *r* was used for correlation calculation.

5.10. GO/KEGG pathway annotation

All the GO and pathway analyses were carried on DAVID site (http://david.abcc.ncifcrf.gov/). For functional Annotation Tool of DAVID, GO biological process categories and KEGG pathways were chosen for enrichment analysis, and human gene sets was chosen as the background.

5.11. Accession code

RNA-Seq data have been uploaded in the NCBI Sequence Read Achieve (SRA) under the accession SRA074279. Their corresponding DNA sequencing data can be accessed in the European Genome-phenome Archive (EGA) under accession ERP001196.

Competing interests

The authors declare to no competing interests.

Authors' contributions

N.P.L. and R.T.P. coordinated the collection of specimens and clinical data. L.K. and Z.P. revised the RNA editing identification pipeline and carried out the data analysis. X.L., H.Z. and Y.L. contributed genome sequencing data analysis. The article was written by L.K. and revised by R.C., Z.P. and M.M. The study was initiated and designed by R.C., M.M., H.D., J.H. and J.W., and coordinated by Z.G. All authors read and approved the final manuscript.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2014.11.005.

Acknowledgments

This work is supported by Asian Cancer Research Group, Inc. (ACRG), a not-for-profit organization formed by Eli Lilly, Merck and Pfizer, the National High Technology Research and Development Program of China – 863 Program (no. 2012AA02A201), the Guangdong Enterprise

Key Laboratory of Human Disease Genomics and Guangdong Innovative Research Team Program (no. 2009010016).

References

[1] K.V. Prasanth, S.G. Prasanth, Z. Xuan, S. Hearn, S.M. Freier, C.F. Bennett, M.Q. Zhang, D.L. Spector, Regulating gene expression through RNA nuclear retention, Cell 123 (2005) 249–263.

[2] K. Nishikura, Functions and regulation of RNA editing by ADAR deaminases, Annu. Rev. Biochem. 79 (2010) 321–349.

[3] S. Maas, Gene regulation through RNA editing, Discov. Med. 10 (2010) 379–386.

[4] D. Dominissini, S. Moshitch-Moshkovitz, N. Amariglio, G. Rechavi, Adenosine-to-inosine RNA editing meets cancer, Carcinogenesis 32 (2011) 1569–1577.

[5] Y. Kawahara, K. Ito, H. Sun, H. Aizawa, I. Kanazawa, S. Kwak, Glutamate receptors: RNA editing and death of motor neurons, Nature 427 (2004) 801.

[6] P.H. Seeburg, A-to-I editing: new and old sites, functions and speculations, Neuron 35 (2002) 17–20.

[7] N. Paz, E.Y. Levanon, N. Amariglio, A.B. Heimberger, Z. Ram, S. Constantini, Z.S. Barbash, K. Adamsky, M. Safran, A. Hirschberg, et al., Altered adenosine-to-inosine RNA editing in human cancer, Genome Res. 17 (2007) 1586–1595.

[8] S.P. Shah, R.D. Morin, J. Khattra, L. Prentice, T. Pugh, A. Burleigh, A. Delaney, K. Gelmon, R. Guliany, J. Senz, et al., Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution, Nature 461 (2009) 809–813.

[9] J.B. Li, E.Y. Levanon, J.K. Yoon, J. Aach, B. Xie, E. LeProust, K. Zhang, Y. Gao, G.M. Church, Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing, Science 324 (2009) 1210–1213.

[10] C. Cenci, R. Barzotti, F. Galeano, S. Corbelli, R. Rota, L. Massimi, C. Di Rocco, M.A. O'Connell, A. Gallo, Down-regulation of RNA editing in pediatric astrocytomas: ADAR2 editing activity inhibits cell migration and proliferation, J. Biol. Chem. 283 (2008) 7251–7260.

[11] J.C. Hartner, C. Schmittwolf, A. Kispert, A.M. Muller, M. Higuchi, P.H. Seeburg, Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1, J. Biol. Chem. 279 (2004) 4894–4902.

[12] L. Chen, Y. Li, C.H. Lin, T.H. Chan, R.K. Chow, Y. Song, M. Liu, Y.F. Yuan, L. Fu, K.L. Kong, et al., Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma, Nat. Med. 19 (2013) 209–216.

[13] T. He, Q. Wang, G. Feng, Y. Hu, L. Wang, Y. Wang, Computational detection and functional analysis of human tissue-specific A-to-I RNA editing, PLoS One 6 (2011) 18129.

[14] E.Y. Levanon, E. Eisenberg, R. Yelin, S. Nemzer, M. Hallegger, R. Shemesh, Z.Y. Fligelman, A. Shoshan, S.R. Pollock, D. Sztybel, et al., Systematic identification of abundant A-to-I editing sites in the human transcriptome, Nat. Biotechnol. 22 (2004) 1001–1005.

[15] M. Blow, A survey of RNA editing in human brain, Genome Res. 14 (2004) 2379–2387.

[16] W.K. Sung, H. Zheng, S. Li, R. Chen, X. Liu, Y. Li, N.P. Lee, W.H. Lee, P.N. Ariyaratne, C. Tennakoon, et al., Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma, Nat. Genet. 44 (2012) 765–769.

[17] Z. Peng, Y. Cheng, B.C. Tan, L. Kang, Z. Tian, Y. Zhu, W. Zhang, Y. Liang, X. Hu, X. Tan, et al., Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome, Nat. Biotechnol. 30 (2012) 253–260.

[18] G. Ramaswami, W. Lin, R. Piskol, M.H. Tan, C. Davis, Li JB: accurate identification of human Alu and non-Alu RNA editing sites, Nat. Methods 9 (2012) 579–581.

[19] Y. Midorikawa, S. Tsutsumi, H. Taniguchi, M. Ishii, Y. Kobune, T. Kodama, M. Makuuchi, H. Aburatani, Identification of genes associated with dedifferentiation of hepatocellular carcinoma with expression profiling analysis, Jpn. J. Cancer Res. 93 (2002) 636–643.

[20] C. Pilarsky, M. Wenzig, T. Specht, H.D. Saeger, R. Grutzmann, Identification and validation of commonly overexpressed genes in solid tumors by comparison of microarray data, Neoplasia 6 (2004) 744–750.

[21] G.C. Jayan, J.L. Casey, Increased RNA editing and inhibition of hepatitis delta virus replication by high-level expression of ADAR1 and ADAR2, J. Virol. 76 (2002) 3819–3827.

[22] J.H. Bahn, J.H. Lee, G. Li, C. Greer, G. Peng, X. Xiao, Accurate identification of A-to-I RNA editing in human by transcriptome sequencing, Genome Res. 22 (2012) 142–150.

[23] E. Park, B. Williams, B.J. Wold, A. Mortazavi, RNA editing in the human ENCODE RNA-seq data, Genome Res. 22 (2012) 1626–1633.

[24] R. Piskol, Z. Peng, J. Wang, J.B. Li, Lack of evidence for existence of noncanonical RNA editing, Nat. Biotechnol. 31 (2013) 19–20.

[25] O. Vasiljeva, A. Papazoglou, A. Kruger, H. Brodoefel, M. Korovin, J. Deussing, N. Augustin, B.S. Nielsen, K. Almholt, M. Bogyo, et al., Tumor cell-derived and macrophage-derived cathepsin B promotes progression and lung metastasis of mammary cancer, Cancer Res. 66 (2006) 5242–5250.

[26] S. Yan, M. Sameni, B.F. Sloane, Cathepsin B and human tumor progression, Biol. Chem. 379 (1998) 113–123.

[27] A. Khan, M. Krishna, S.P. Baker, R. Malhothra, B.F. Banner, Cathepsin B expression and its correlation with tumor-associated laminin and tumor progression in gastric cancer, Arch. Pathol. Lab. Med. 122 (1998) 172–177.

[28] Y. Nakamura, M. Takeda, H. Suzuki, H. Hattori, K. Tada, S. Hariguchi, S. Hashimoto, T. Nishimura, Abnormal distribution of cathepsins in the brain of patients with Alzheimer's disease, Neurosci. Lett. 130 (1991) 195–198.

[29] I. Podgorski, B.F. Sloane, Cathepsin B and its role(s) in cancer progression, Biochem. Soc. Symp. 263–276 (2003).

[30] A. Kiran, P.V. Baranov, DARNED: a DAtabase of RNa EDiting in humans, Bioinformatics 26 (2010) 1772–1776.

[31] Y. Wang, S. Naruse, M. Kitagawa, H. Ishiguro, Y. Nakae, T. Hayakawa, Urinary excretion of trypsinogen activation peptide (TAP) in taurocholate-induced pancreatitis in rats, Pancreas 22 (2001) 24–27.

[32] L.S. Bicknell, C. Farrington-Rock, Y. Shafeghati, P. Rump, Y. Alanay, Y. Alembik, N. Al-Madani, H. Firth, M.H. Karimi-Nejad, C.A. Kim, et al., A molecular and clinical study of Larsen syndrome caused by mutations in FLNB, J. Med. Genet. 44 (2007) 89–98.

[33] M. Zhang, J. Gong, D.H. Presky, W. Xue, P.F. Barnes, Expression of the IL-12 receptor beta 1 and beta 2 subunits in human tuberculosis, J. Immunol. 162 (1999) 2441–2447.

[34] P. Kumar, S. Henikoff, P.C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nat. Protoc. 4 (2009) 1073–1081.

[35] I. Dunham, A. Kundaje, S.F. Aldred, P.J. Collins, C.A. Davis, F. Doyle, C.B. Epstein, S. Frietze, J. Harrow, R. Kaul, et al., An integrated encyclopedia of DNA elements in the human genome, Nature 489 (2012) 57–74.

[36] R. Li, C. Yu, Y. Li, T.W. Lam, S.M. Yiu, K. Kristiansen, J. Wang, SOAP2: an improved ultrafast tool for short read alignment, Bioinformatics 25 (2009) 1966–1967.

[37] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.

[38] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nat. Methods 5 (2008) 621–628.