Recurrently deregulated IncRNAs in hepatocellular carcinoma

Yang Yang1,*, Lei Chen2,3,*, Jin Gu4,*, Hanshuo Zhang5,*, Jiapei Yuan1, Qiuyu Lian4, Guishuai Lv2, Siqi Wang1, Yang Wu1, Yu-Cheng T. Yang1, Dongfang Wang4, Yang Liu1, Jing Tang6, Guijuan Luo2, Yang Li1, Long Hu1, Xingao Sun1, Dong Wang1, Mingzhou Guo7, Qiaoran Xi1, Jianzhong Xi5, Hongyang Wang2,3,**, Michael Q. Zhang1,4,8,** & Zhi John Lu1,**

Hepatocellular carcinoma (HCC) cells often invade the portal venous system and subsequently develop into portal vein tumour thrombosis (PVTT). Long noncoding RNAs (lncRNAs) have been associated with HCC, but a comprehensive analysis of their specific association with HCC metastasis has not been conducted. Here, by analysing 60 clinical samples’ RNA-seq data from 20 HCC patients, we have identified and characterized 8,603 candidate lncRNAs. The expression patterns of 917 recurrently deregulated lncRNAs are correlated with clinical data in a TCGA cohort and published liver cancer data. Matched array data from the 60 samples show that copy number variations (CNVs) and alterations in DNA methylation contribute to the observed recurrent deregulation of 235 lncRNAs. Many recurrently deregulated lncRNAs are enriched in co-expressed clusters of genes related to cell adhesion, immune response and metabolic processes. Candidate lncRNAs related to metastasis, such as HAND2-AS1, were further validated using RNAi-based loss-of-function assays. Thus, we provide a valuable resource of functional lncRNAs and biomarkers associated with HCC tumorigenesis and metastasis.
Hepatocellular carcinoma (HCC) is one of the most common and aggressive human malignancies. The dismal clinical outcome of HCC is largely due to the high incidence of intrahepatic and extrahepatic metastasis in HCC patients. HCC cells are highly likely to develop into portal vein tumour thrombosis (PVTT), which is the main route for intra-hepatic metastasis of HCC (ref. 3). Therefore, PVTT is closely associated with poor prognosis for HCC patients.

Several long noncoding RNAs (lncRNAs), including H19 (ref. 5), HOTAIR (ref. 6) and HULC (ref. 7), are directly involved in tumorigenesis and metastasis of various types of cancer. Recent studies have also revealed the pre-metastasis mechanisms through which some lncRNAs contribute to the activation of epithelial-to-mesenchymal transition networks, including activation of the WNT (ref. 8) and TGF-β signalling pathways.

Moreover, efforts at systematic identification and characterization of candidate lncRNAs involved in HCC, especially those involved in HCC metastasis, remain at an early stage. A recent study based on The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) revealed the existence of more than 50,000 lncRNAs (designated MiTranscriptome lncRNAs) in the human transcriptome (generated from various tumours, normal tissues and cell lines), of which more than 80% were not reported in previous studies or found in databases (that is, GENCODE (ref. 13) and RefSeq (ref. 14)). This study demonstrates the genomic diversity and expression specificity of lncRNAs, while suggesting that more lncRNAs will be discovered as additional tumour and cell types (for example, metastatic samples) are sequenced.

Remarkably, studies suggest that approximately 88% of single-nucleotide polymorphisms (SNPs) in the human genome are within noncoding regions, suggesting that many noncoding nucleotide polymorphisms (SNPs) in the human genome are relatively stable local secondary structures, facilitating their detection in body fluids. For instance, lncRNA PVT1 has been used as a diagnostic and prognostic biomarker for HCC (ref. 11).

Here, 60 matched samples (primary tumour, PVTT and adjacent normal tissue) from 20 Chinese HCC patients were subjected to total RNA-seq (rRNA depleted), followed by integrative analysis at the genomic, transcriptomic and epigenomic levels, with the goal of identifying and characterizing deregulated lncRNAs in HCC patients. Approximately 76% of the lncRNAs identified in the samples were not annotated by the MiTranscriptome or GENCODE transcriptome databases. Next, approximately 1,000 lncRNAs that were recurrently deregulated in primary tumours and/or PVTTs were identified. Their expression levels were correlated with TCGA clinical data and additional published liver cancer data. We also showed that one of the recurrently deregulated lncRNAs was suitable as a prognosis and metastasis biomarker in HCC patients. Furthermore, copy number variations (CNVs) and DNA methylation alterations were shown to be responsible for the aberrant expression patterns of 147 and 93 recurrently deregulated lncRNAs, respectively. Finally, a coding-noncoding co-expression network was used to predict candidate lncRNAs related to metastasis, after which the predictions were validated experimentally.

Results
Identification of candidate lncRNAs in HCC clinical samples.

To systematically identify lncRNAs related to HCC tumorigenesis and metastasis, approximately 9.6 billion reads for 60 samples from 20 HCC patients were sequenced using total RNA-seq (rRNA depleted) (Supplementary Data 1). Three matched samples were collected from each patient: primary tumour, adjacent normal tissue and PVTT.

We first found that 95.1% of 13,870 lncRNAs (including 23,898 transcripts) annotated by GENCODE (V19) (ref. 13) were detected (FPKM > 0.5 for single-exon lncRNAs and FPKM > 0 for multi-exon lncRNAs) in our samples, indicating that our sequencing depth was good. Next, using these GENCODE lncRNAs as a reference annotation, we assembled 8,603 candidate lncRNAs (including 10,196 transcripts) (Supplementary Data 2) with a bioinformatics pipeline (Fig. 1a and Supplementary Table 1): (1) assembling RNA transcripts from RNA-seq reads; (2) filtering potential noise based on genomic location, length and expression level; (3) removing transcripts with coding potential, which was calculated by two computational tools, CPC (ref. 22) and COME (ref. 23) (see details in Methods). These candidate lncRNAs were designated as newly assembled lncRNAs. Only a small number of the newly assembled lncRNAs were reported in other studies. For example, 76% of them were not found in the MiTranscriptome database, which was mainly derived from TCGA data (Fig. 1b). The newly assembled lncRNAs were also compared with two other representative annotation databases: a high quality set, RefSeq (ref. 14), and a comprehensive set, NONCODE (over 50,000 lncRNA transcripts included) (ref. 24). Only 2% and 10% of the newly assembled lncRNAs were found in the RefSeq (Release 72) and NONCODE (V4) databases, respectively (Supplementary Fig. 1). These results indicate the depth of our sequencing and expression specificity of the lncRNAs identified in our samples. We showed that the number of lncRNAs increased when the number of sequenced samples was increased (Supplementary Fig. 2), while the detection ability for protein-coding genes and GENCODE lncRNAs was saturated at approximately 10 and 20 out of 60 samples, respectively.

Characterization of the candidate lncRNAs.

We characterized genomic location, expression abundance, transcript length, conservation and SNP enrichment for the newly assembled lncRNAs (Fig. 1c–f, Supplementary Data 3 and 4). We first focused on the genomic locations of newly assembled lncRNAs (Supplementary Fig. 3A). The majority (74%) of the lncRNAs were located in intergenic regions; 16% were antisense to protein-coding genes, whereas 3% were located in the introns of protein-coding genes. Moreover, 1.39% and 24.94% of the lncRNAs overlapped with pseudogenes and transposable elements, respectively, whereas 2.24% and 5.06% of the lncRNAs contained local domains conserved with canonical ncRNAs (for example, rRNA, tRNA and snRNA) at the sequence and structure levels, respectively. Similarly, the GENCODE lncRNAs also overlapped with or included these elements (Supplementary Fig. 3b,c).

Next, we characterized the basic features of the newly assembled lncRNAs and compared them with protein-coding genes and GENCODE lncRNAs. Because they had fewer putative exons, we found that the newly assembled lncRNA transcripts were shorter than protein-coding genes, but longer than GENCODE lncRNAs (Fig. 1c, Supplementary Fig. 4). This result
indicates that the high sequencing depth of our analysis (Supplementary Data 1) enabled us to assemble long transcripts close to their full length, even though they were expressed at low levels (Fig. 1d, Supplementary Data 4 and 5). Notably, the newly assembled lncRNAs were less evolutionarily conserved in comparison with protein-coding genes and GENCODE lncRNAs, while exonic regions were more conserved in comparison with intronic regions (Fig. 1e).

A previous study reported that almost 90% of SNPs are located in non-coding regions15. To investigate the relationship between lncRNAs and diseases, we capitalized on the GWAS SNP Catalog from NHGRI (ref. 25). We intersected the lncRNAs with GWAS SNPs from the NHGRI catalogue and randomly selected SNPs from the dbSNP (ref. 26). Interestingly, we found that GWAS SNPs were significantly enriched in the newly assembled lncRNAs in comparison with a random set (Fig. 1f). These data suggested that our newly assembled lncRNAs were likely to be functionally associated with human diseases.

To further validate the activity of the newly assembled lncRNAs, we used published ChIP-seq data for the HepG2 cell line27 to depict activity markers around transcription start sites (TSSs). Different epigenetic signatures (H3K4me3, H3K27ac, Pol II binding, DNase I hypersensitivity) indicated active transcription of the newly assembled lncRNAs in liver cancer cell lines (Supplementary Fig. 5). Peaks of these markers were found at the TSSs of lncRNAs, suggesting that the promoters of these transcripts are actively regulated in HepG2 cells.

Recurrently deregulated lncRNAs in tumours and PVTTs. We used three statistical methods, GFOLD (ref. 28) followed by counting recurrences in multiple patients (Supplementary Figs 6–7), DESeq2 (ref. 29) and Wilcoxon signed-rank test to define lncRNAs (including both GENCODE lncRNAs and newly assembled lncRNAs) that were differentially expressed recurrently (Fig. 2a,b, Supplementary Data 6) (see details in Methods).

For the comparison of primary tumours and adjacent normal tissues, we found that the results of DESeq2 (ref. 29) and Wilcoxon signed-rank test were more similar to each other than to the results of GFOLD (Fig. 2a), because the former two methods both treated the patients as replicates, whereas GFOLD

Figure 1 | Identification of candidate lncRNAs in 60 HCC samples. (a) Overview of the comprehensive experimental and computational scheme for the systematic identification of lncRNAs in samples from HCC patients. (b) Venn diagram showing the overlap between newly assembled lncRNAs and MiTranscriptome lncRNAs. Characterization of GENCODE lncRNAs, newly assembled lncRNAs and protein-coding genes: (c) transcript length distribution; (d) cumulative distribution curve of maximum gene expression level (RPKM); (e) conservation of exons and introns; (f) enrichment of GWAS SNPs (circle) over randomly selected SNPs (triangle).
assessed differential expression based on individual patients. Finally, we identified 525 and 323 lncRNAs as recurrently downregulated or upregulated, respectively, in primary tumours by overlapping the predictions of all three methods (Fig. 2a,c).

The identified lncRNAs were considered to represent a group of recurrently deregulated lncRNAs potentially associated with tumorigenesis. These lncRNAs and their P-values, q-values and expression fold-changes are listed in Supplementary Data 6.

For the comparison of PVTTs and primary tumours, we found that DESeq2 and Wilcoxon signed-rank test identified very few differentially expressed lncRNAs (Fig. 2b), because of the relatively high heterogeneity of PVTTs (Fig. 2d). Only one lncRNA (HAND2-AS1) was identified as a downregulated candidate in PVTTs. Paired primary tumours and PVTTs from individual patients (average correlation coefficient: 0.99) were more similar than PVTTs from different patients (average correlation coefficient: 0.76). Therefore, we used GFOLD to evaluate differential expression by treating patients individually, followed by counting recurrences in multiple patients (see Methods), revealing 107 lncRNAs that were defined as recurrently deregulated lncRNAs potentially associated with metastasis (Fig. 2d, Supplementary Data 6). Notably, of the 107 metastasis-associated candidates, 38 were also recurrently deregulated in primary tumours in comparison with adjacent normal tissues (see details in Discussion).

Examples of tumorigenesis-associated lncRNAs are shown in Fig. 2e. Some of the tumorigenesis-associated lncRNAs identified in this study have been reported by previous studies. For instance,
PVT1 promotes cell proliferation, cell cycle progression and development of stem cell-like properties in HCC (ref. 11). In addition, several lncRNAs (for example, TEX41, XLOC_014515 and XLOC_030220) were recurrently upregulated in PVTT samples, whereas others (for example, HAND2-AS1, RP11-731F5.2 and XLOC_055355) were recurrently downregulated in PVTTs (Fig. 2f). For example, HIF1A-AS2, a lncRNA antisense to hypoxia-inducible factor 1-alpha, is overexpressed in gastric cancer cells and involved in gastric cancer development20. Notably, we identified some newly assembled lncRNA candidates potentially related to metastasis, including XLOC_014515 and XLOC_030220 (Supplementary Data 6).

Association of deregulated lncRNAs with public clinical data.

Based on Gene Set Enrichment Analysis (GSEA) (see Methods) of recurrently deregulated lncRNAs, we found that the expression levels of the tumorigenesis- and metastasis-associated lncRNA sets were consistent with their expression levels in another published data set from 11 matched normal tissue samples, primary tumours and PVTTs (ref. 21) (Fig. 3a,b). In addition, we explored the expression levels of recurrently deregulated lncRNAs in a TCGA liver hepatocellular carcinoma (LIHC) cohort (see Methods). Tumorigenesis-associated lncRNAs were also aberrantly expressed between normal tissues and primary tumours. Because the TCGA LIHC cohort had no PVTT or metastatic tumour samples, we compared expression levels of the tumorigenesis- and metastasis-associated lncRNA candidates potentially related to metastasis, using XLOC_014515 and XLOC_030220 (Supplementary Data 6).

Figure 3 | Association of recurrently deregulated lncRNAs with TCGA clinical data and other published data. (a) Gene set enrichment analysis (GSEA) of recurrently deregulated tumorigenesis-associated lncRNAs based on TCGA LIHC data and published liver cancer data. lncRNAs were rank-ordered by differential expression between adjacent normal tissue and primary tumour samples. (b) GSEA of recurrently deregulated metastasis-associated lncRNAs. lncRNAs were rank-ordered by differential expression between primary tumours with and without vascular invasion in the TCGA LIHC data, as well as by differential expression between primary tumours and PVTTs in published liver cancer data. (c) Kaplan-Meier analysis of overall survival in the TCGA LIHC cohort. Subjects were stratified according to the expression of lncRNA RP11-166D19.1. The P value for Kaplan-Meier analysis was determined using log-rank test. (d) Multivariate analysis using additional clinical information. Forest plot depicting correlations between the indicated clinical criteria and the expression level of RP11-166D19.1. (e) Expression levels (TCGA data) of RP11-166D19.1 in three HCC subclasses (S1, S2 and S3). ***P value < 0.001, Wilcoxon rank-sum test.
DNA methylation alterations and CNVs of lncRNAs. We categorized the recurrently deregulated lncRNAs based on their correlations with CNVs and/or DNA methylation alterations assayed in our matched samples (Fig. 4a). We listed all recurrently deregulated lncRNAs (Supplementary Data 8) with expression patterns correlated with CNV data (that is, upregulated lncRNAs were found to be located in a DNA amplification region) and/or DNA methylation data (that is, downregulated lncRNAs had strong DNA methylation signals at their promoter regions). Several lncRNAs were recurrently upregulated in some patients and recurrently downregulated in other patients; such lncRNAs were designated as bimorphic lncRNAs (Fig. 4a).

The CNVs of recurrently deregulated lncRNAs in HCC cells were analysed using Cytoscan HD arrays. Based on a GISTIC analysis\(^\text{33}\), 4 significantly amplified genome regions and 70 significantly deleted genome regions were revealed in our samples (Fig. 4b). To characterize candidate CNV-driven lncRNAs, we mapped recurrently deregulated lncRNAs to amplified and deleted genome regions. In total, 147 recurrently deregulated lncRNAs were identified in deleted regions, whereas none were found in amplified regions (Fig. 4a, Supplementary Data 8). For example, FENDRR, which was reported to be a prognostic biomarker in gastric cancer\(^\text{34}\), had a pattern of decreased expression driven by copy number deletion (Fig. 4b).

Based on DNA methylation microarrays of 60 matched samples, we analysed DNA methylation patterns to identify recurrently deregulated lncRNAs that were affected by alterations in DNA methylation. We applied several separate filtering criteria (see Methods) to define recurrently deregulated lncRNAs driven by alterations in DNA methylation. In total, 93 (10.1%) recurrently deregulated lncRNAs had significant correlations between DNA methylation and expression levels (Fig. 4a,c, Supplementary Data 8), suggesting that their expression levels in tumour and/or PVTT samples were probably regulated by DNA methylation. As an example, we showed that the expression level of a recurrently deregulated lncRNA, HAND2-ASI, was inversely correlated with the DNA methylation level at its promoter region (\(\text{R}^2 = 0.16694\)) (Fig. 4d). The promoter region of HAND2-ASI was hypermethylated in primary tumours and PVTT samples.

Inference of lncRNA function using a co-expression network. To predict the potential functional and regulatory mechanisms of lncRNAs with respect to the molecular aetiology of HCC, we constructed a co-expression network\(^\text{35}\) of protein-coding genes and lncRNAs (see Methods). The resulting co-expression network consisted of 7,367 protein-coding genes, 6,377 GENCODE lncRNAs and 5,612 newly assembled lncRNAs. There were 1,286 recurrently deregulated lncRNAs in the network. All protein-coding genes and lncRNAs were grouped into 43 clusters, each of which had at least 100 highly inter-connected genes (Supplementary Data 9). In addition to interaction edges within a cluster, there were also 337,609 edges between nodes in different clusters, which could indicate their functional relatedness or regulatory relationships (Fig. 5a, Supplementary Fig. 11).

Among the 43 clusters, we found four clusters containing protein-coding genes with interesting functions: clusters 4, 9, 18 and 25 (Fig. 5b). The recurrently deregulated lncRNAs are highly enriched in these gene clusters (Fisher’s exact test, all \(P < 0.01\)). For example, Gene Ontology and KEGG pathway enrichment analyses suggest that the protein-coding genes in cluster 4 are mostly associated with metabolic processes in the liver, such as organic acid metabolism and degradation of fatty acids. The protein-coding genes in cluster 9 mainly function in cell cycle processes such as DNA repair, DNA replication and DNA metabolism, which influence cell migration\(^\text{36}\). Cluster 18 is enriched with immune response genes involved in the T-cell and B-cell receptor signalling pathways, consistent with reports that immune and inflammatory responses play decisive roles in tumour development by influencing the processes of invasion and migration\(^\text{37}\).

Another intriguing example is cluster 25 (Fig. 5c), which includes protein-coding genes enriched in functional terms related to metastasis, such as cell adhesion and the TGF-\(\beta\) signalling pathway, which have been shown to play essential roles in diverse processes, including cell proliferation, differentiation, motility and adhesion\(^\text{38}\). Furthermore, many HCC-related driver genes\(^\text{39}\) are also found in cluster 25. For example, the FLT3 receptor plays a role during the late stages of liver regeneration and is involved in GPCR signalling pathways\(^\text{40}\). FLT3 was co-expressed with other driver genes in the sub-network, including WDFY4 and FAT4. Moreover, three recurrently deregulated lncRNAs (HAND2-ASI, ACO96579.7 and FENDRR) were strongly correlated with FLT3 at the expression level, suggesting that these lncRNAs have functions related to that of FLT3. Another interesting gene in cluster 25, FAT4, encodes a cadherin (a calcium-dependent cell adhesion protein) that serves as a tumour-suppressor gene\(^\text{41}\). FAT4 was closely associated with some recurrently deregulated lncRNAs, including HAND2-ASI, FENDRR and WDFY3-AS2, all of which were differentially expressed during cell migration. Cell adhesion was one of the most significantly enriched processes during tumour metastasis. These co-expression relationships provide functional evidence demonstrating that adhesion-related lncRNAs likely have roles in tumour metastasis. Furthermore, driver gene ZFPM2 was also highly involved in the sub-network; it was significantly correlated with seven driver genes and five recurrently deregulated lncRNAs. Some migration-related
recurrently deregulated lncRNAs (HAND2-AS1, AC096579.7 and FENDRR) had expression patterns similar to that of PTPRB, which plays an important role in blood vessel remodelling and angiogenesis, indicating that these lncRNAs could have related functions.

**Functional assay for metastasis-related lncRNA candidates.** Transwell migration assays were used to assess whether putative candidate lncRNAs might function in the progression of HCC. We first selected ten lncRNA candidates that were associated with cell adhesion based on the co-expression network described above (Supplementary Table 2), in which four lncRNAs (WDFYS3-AS2, HAND2-AS1, RP11-166D19.1 and XLOC_055355) were also significantly co-expressed with genes in the TGF-β signalling pathway. Note that seven candidate lncRNAs (shown in Fig. 6a) were recurrently deregulated lncRNAs in PVTT samples and designated as metastasis-associated lncRNAs in the sections above. Three siRNAs were synthesized for each candidate lncRNA and mixed as a pool (Supplementary Table 3). Three liver cancer cell lines, HepG2, SMMC-7721 and HCCLM9, were used to conduct loss-of-function RNAi assays (Fig. 6a).

Remarkably, knockdown of seven of the ten candidate lncRNAs significantly affected cell migration in at least one cell line; knockdown of three lncRNAs produced accordant alterations in cell migration in at least two cell lines by suppressing or promoting cell migration (Fig. 6a, Supplementary Table 3). RNAi knockdown efficiency was confirmed using
**Figure 5 | Inference of potential functions for recurrently deregulated lncRNAs using a co-expression network.** (a) Network representation of 18 selected inter-connected clusters in the coding-non-coding co-expression network. (b) GO and KEGG pathway enrichment for four selected clusters (4, 9, 18 and 25). Heatmap showing enrichment scores (−log10(P value)) for GO terms and KEGG pathways in four selected clusters. The most significantly enriched GO terms and KEGG pathways are displayed. (c) Sub-network showing important genes/lncRNAs in cluster 25. The subnetwork depicts the relationships among four lncRNAs and liver cancer-related driver genes.

qRT-PCR (Fig. 6b, Supplementary Fig. 12, Supplementary Table 4). The results from three representative transwell assays are shown on the right side of the figure (Fig. 6c,d). In order to document the reproducibility of the results, we repeated the RNAi knockdown and transwell migration assays using three siRNAs separately in all three cell lines (Supplementary Fig. 13A–B and Supplementary Table 5); the results of these experiments were consistent with those of the mixed siRNA experiments. Moreover, to avoid mistaking differences in cell proliferation for differences in cell migration, we performed CCK8 cell proliferation assays, revealing that knockdown of these lncRNAs hardly affected cell proliferation (Supplementary Fig. 13A–B and Supplementary Table 5); the results of repeated the RNAi knockdown and transwell migration assays using three siRNAs separately in all three cell lines (Supplementary Fig. 13A–B and Supplementary Table 5).

For some lncRNA knockdown experiments, changes in migration ability were consistent with the deregulation patterns of HCC patients. For example, *RP11-166D19.1* was recurrently downregulated in PVTT samples from four patients. The loss-of-function assay showed that knockdown of *RP11-166D19.1* enhanced the migration ability of HCC cells. However, some other lncRNAs, such as *HAND2-ASI*, demonstrated an inconsistent trend between deregulation patterns in HCC patients and experimentally validated functions in cancer cell lines; silencing of *HAND2-ASI* suppressed cell migration, although it was downregulated in 8 of 20 patients’ PVTT samples. It has been reported that nearly half of HCC cell lines do not resemble primary tumours, so the intrinsic differences between cancer cells lines and clinical samples might explain the discrepancies between the samples’ gene expression patterns and experimentally validated functions in cell lines. Overall, the high validation rate of the candidate lncRNAs showed that the co-expression network, based on previous knowledge of signalling pathways and supplemented by recurrent aberrant expression patterns in matched clinical samples, identified candidate lncRNAs that potentially play functional roles in the sophisticated regulation of cancer development and progression.

**Discussion**

Based on an analysis of genomic, epigenomic and transcriptomic data of HCC primary tumours and PVTTs, this study reports several findings. First, based on high-throughput sequencing technology and bioinformatics analysis of 60 matched samples, including primary tumours, PVTTs and adjacent normal tissue, we discovered and characterized an expanded landscape of lncRNAs. Using PVTTs from Chinese HCC patients and deep sequencing data enabled us to detect many candidate lncRNA
transcripts. Moreover, we identified lncRNAs that were recurrently deregulated during HCC tumorigenesis and metastasis. Second, integrative multi-omics analysis revealed that recurrent deregulation of lncRNA expression was often associated with alterations in DNA methylation and CNV. In addition, lncRNA expression levels were correlated with clinical data from the TCGA and other published liver cancer data sets. Lastly, using network analysis and loss-of-function assays, we identified functional lncRNAs potentially related to cell adhesion, immune responses and metabolic processes. For example, our paired RNA-seq data showed that lncRNA HAND2-AS1 was recurrently deregulated; its expression levels among 60 samples were inversely correlated with matching DNA methylation data. Based on our co-expression network, we inferred that HAND2-AS1 might be related to HCC metastasis. Finally, using an RNAi approach on our co-expression network, we inferred that HAND2-AS1 could potentially serve as a promising single-gene HCC biomarker.

In addition, we have shown that RP11-166D19.1 could potentially serve as a promising single-gene HCC biomarker. We also demonstrated that knockdown of RP11-166D19.1 promoted cell migration. RP11-166D19.1 is an isoform of lncRNA MIR100HG, a leukemia-related oncogene hosting three miRNAs (let-7a, miR-100 and mir-125b) in its introns. As reported previously, lncRNAs are more tissue- and cell-type-specific in comparison with protein-coding genes. Moreover, in comparison with protein-coding genes, the local secondary structures of lncRNAs confer greater stability and provide a greater likelihood of detection. Therefore, translation of these results into candidate lncRNA biomarkers might impact clinical decision-making and ultimately improve clinical outcomes for patients with HCC.

By exploring lncRNA transcriptome alteration, we found that the lncRNA landscapes of PVTTs were indistinguishable from those of matched primary tumours, consistent with previous studies. We employed principal component analysis to assess the expression profiles of recurrently deregulated lncRNAs in different samples. Principal component analysis showed that the recurrently deregulated lncRNAs could be used to clearly distinguish primary tumours from adjacent normal tissues, while PVTTs were more similar to primary tumours (Supplementary Fig. 8). This observation showed that the lncRNA expression profile of the PVTTs was very similar to that of their matched primary tumours, consistent with studies on protein-coding genes, CNV and DNA methylation. These findings suggest that (1) primary tumours of HCC patients with PVTT may contain sub-clones with the potential to invade the portal vein and develop into PVTTs; (2) many metastasis-associated lncRNAs were deregulated in these sub-clones. These findings are consistent with clinical observations, because all of our sequenced patients had stage IV HCC with serious PVTT. Although the overall lncRNA expression patterns of PVTTs were similar to those of their matched primary tumours, approximately 100 lncRNAs were significantly and recurrently deregulated in PVTTs in comparison with their expression levels in paired

Figure 6 | Loss-of-function assay of candidate lncRNAs regulating cell migration. Transwell migration assays were conducted to test the effects of siRNA-mediated RNAi of candidate lncRNAs in three liver cancer cell lines: HepG2, SMMC-7721 and HCCLM9. (a) The value in the heatmap is the fold-change (P value<0.05) of the transwell cell numbers for knockdown cells over those of control cells. All results are expressed as the mean derived from three independent experiments. The unpaired Student’s t-test (two-tailed) was used for comparisons of two groups. Seven of the ten candidate lncRNAs were metastasis-associated lncRNAs that were recurrently deregulated in PVTTs (in bold font). *P value<0.05, **P value<0.01, ***P value<0.001, t-test, n = 3. (b) RNAi was validated by qRT-PCR. Examples of migration phenotype: transwell cells (DAPI staining) (c) (Scale bar, 500 µm) and their counts (d). Error bars represent the s.d. of three experiment replicates.
primary tumours. These lncRNAs could play essential roles in metastasis, because they were deregulated further as primary HCC cells invaded the portal vein.

PVTT has been considered as a type of intrahepatic HCC metastasis by several previous genomic studies. Moreover, PVTTs have also been used to study HCC metastasis in functional and mechanistic studies, although some researchers have not regarded PVTT as a solid metastatic model. In this genomic study, we provide putative associations and predicted candidates at the transcriptome level, but their functions and the mechanisms in which they play a role must be confirmed by experimental validation.

In this study, we identified recurrently deregulated tumorigenesis- and metastasis-associated lncRNAs, many of which were experimentally validated and mechanistically linked to cancer development and progression. We anticipate that the recurrently deregulated lncRNAs identified in this report could provide a valuable resource for studies aimed at delineating the relationship between functional lncRNAs and HCC tumorigenesis/metastasis.

In addition, recent studies suggest that lncRNAs can code for small peptides. Although COME can detect metastasis, in functional and mechanistic studies many of which were experimentally validated and mechanistically linked to cancer-

Annotation of candidate lncRNAs. We annotated the genomic locations of the identified candidate lncRNAs. GENCODE lncRNAs and MiTranscriptome lncRNAs (TCGA) by overlapping them with annotated coding genes. Intronic lncRNAs were defined as those located in the intronic regions of coding genes on the sense strand. Antisense lncRNAs were those that overlapped at least 1 nt with any exon (including both coding genes and ncRNAs) on the antisense strand. Cis-lncRNAs (also called sense lncRNAs) were those that were close to (within 2,000 nt of the 5′ or 3′-ends) a protein-coding gene. The remaining lncRNAs that did not overlap with any coding genes or annotated ncRNAs were designated as intergenic lncRNAs.

We also assessed whether any of the candidate lncRNAs overlapped with pseudogenes or transposable elements, because previous studies suggested that some lncRNAs could be derived from such sequences. Annotations of pseudogenes and transposable elements were derived from GENCODE and the UCSC Genome Browser, respectively.

Furthermore, we also annotated lncRNAs with domains/motifs conserved with annotated gRNA (gRNA) coding regions and with differentially expressed lncRNAs. These lncRNAs were defined as those located in the intronic regions of coding genes on the sense strand.

Methods Transcriptome assembly for 60 samples from HCC patients. Total RNA from 60 samples from 20 Chinese HCC patients was sequenced (GSE77509). Each patient had three matched samples: primary HCC tumour, adjacent normal liver tissue and PVTT. The patients were ordered using alphabetic labels (A to T) in this paper, but the patients were originally numbered as 3, 6, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 25 and 26. The PVTT sample of one patient (14) was not distinguishable from normal tissue, so we did not use it in our migration and metastasis analyses. The ethical committee of EHBH hospital approved this study. Informed consent was obtained from each patient.

We first evaluated RNA-seq quality using FastQC (version 0.10.1) and found that all raw reads qualified for the analysis. We aligned the RNA-seq reads to the reference genome using Bowtie with one mismatch in order to estimate the rRNA ratio. Most of the rRNAs were removed by our experiments; only a few remained and generated rRNA reads. Next, the RNA-seq reads were aligned to the human reference genome (hg19) using TopHat (version 2.0.10) (ref. 53) with default parameters. The human genome sequence was downloaded from Ensembl (Homo sapiens GRCH37/hg19). After mapping, we further removed PCR duplicates using sambamba (ref. 24).

Further details of the preprocessing results are described in Supplementary Data 1. Subsequently, based on the mapped reads, we re-assembled a transcriptome using Cufflinks (version 2.2.1) (ref. 55) by providing reference annotations (option ‘--g’) from GENCODE (v19) for each data set of 60 samples. Next, we used Cuffmerge to merge all meta-assemblies to generate a final transcriptome (Supplementary Data 2).

Identification of candidate lncRNAs. After the transcriptome was assembled, we used several stringent filters to identify a set of candidate lncRNAs, newly assembled lncRNAs, in addition to GENCODE lncRNAs:

[1] Transcripts that overlapped (> = 1 nt) on the same strand with the exons of protein-coding genes or noncoding RNAs (both canonical ncRNAs and long lncRNAs (lncRNAs)) annotated by GENCODE (V19) were removed. Canonical ncRNAs include rRNA, tRNA, miRNA, snoRNA, snRNA, misc_RNA, mitochondria tRNA and rRNA. Six biotypes were defined as ‘long non-coding RNAs’ by GENCODE: ‘lincRNA’, ‘processed_transcript’, ‘sense_intronic’, ‘sense_overlapping’, ‘antisense’, and ‘preambly_overlapping_ncrna’. Note that ‘processed transcript’ means that a transcript does not contain an open reading frame, although it could have a historical protein-coding-style name. Note that some lncRNAs were updated as protein-coding genes in recently released GENCODE annotation versions.

[2] Transcripts that overlapped with long noncoding RNAs (lncRNAs) on the same strand were filtered.

[3] To remove fragments of annotated RNA, single-exon transcripts proximal (within 2000 bp) to protein-coding genes or other noncoding RNAs on the same strand were filtered.

We calculated transcript expression levels using Cufflinks with rRNA reads masked. Single-exon transcripts with low expression levels (FPKM < 0.5) in all samples were removed.

To ensure stringent evaluation of coding potential, we calculated the coding potential of each transcript using two computational tools, CPC (ref. 22) and COME (ref. 23). CPC calculates coding potential based on sequence features, whereas COME integrates expression, RNA secondary structure, conservation and epigenetic signals (Supplementary Table 1). COME has been successfully applied to noncoding RNA prediction in worm, fly, human, mouse and Arabidopsis. Transcripts with CPC score > 0 or COME score > 0.5 were removed.

Annotation of candidate lncRNAs. We annotated the genomic locations of the identified candidate lncRNAs. GENCODE lncRNAs and MiTranscriptome lncRNAs (TCGA) by overlapping them with annotated coding genes. Intronic lncRNAs were defined as those located in the intronic regions of coding genes on the sense strand. Antisense lncRNAs were those that overlapped at least 1 nt with any exon (including both coding genes and ncRNAs) on the antisense strand. Cis-lncRNAs (also called sense lncRNAs) were those that were close to (within 2,000 nt of the 5′ or 3′-ends) a protein-coding gene. The remaining lncRNAs that did not overlap with any coding genes or annotated ncRNAs were designated as intergenic lncRNAs.

We also assessed whether any of the candidate lncRNAs overlapped with pseudogenes or transposable elements, because previous studies suggested that some lncRNAs could be derived from such sequences. Annotations of pseudogenes and transposable elements were derived from GENCODE and the UCSC Genome Browser, respectively.

Furthermore, we also annotated lncRNAs with domains/motifs conserved with annotated gRNA (gRNA) coding regions and with differentially expressed lncRNAs. These lncRNAs were defined as those located in the intronic regions of coding genes on the sense strand.

To assess SNPs in different genomic elements, we downloaded two SNP databases:

(1) 12,891 SNPs from the National Human Genome Research Institute’s GWAS catalogue (https://www.genome.gov/26525384);

(2) 14,416,369 common SNPs from dbSNP Build 142 (downloaded from the UCSC Genome Browser) (treated as background variation). We calculated the number of SNPs that overlapped with the transcripts using the BEDTools intersect function. We first calculated the fraction, frac(transcripts), of the amount of overlapped SNPs from the GWAS catalogue to the number of overlapped background SNPs for different categories of genomic elements (for example, lncRNAs and protein-coding genes). Next, we shuffled the transcripts’ positions on the whole genome 100 times and re-calculated frac(shuffled transcripts). Subsequently, we calculated the odds ratio (OR) as

OR = frac(transcripts) / frac(shuffled transcripts)

An OR (control) was calculated by replacing SNPs from the GWAS catalogue with control SNPs:

OR(control) = frac(transcripts) / frac(shuffledtranscripts)

where the control SNPs were randomly selected from the background SNPs shuffled over the whole genome. The significance of comparison for the OR over OR (control) was tested via paired Student’s t-test.

Differential expression analyses. From 8,603 newly assembled lncRNAs and 13,870 known lncRNAs annotated in GENCODE, we identified lncRNAs that were differentially expressed between primary tumours and adjacent normal tissues, as well as between PVTTs and matched primary tumours. We used two different strategies: treating all patients as biological replicates (DESeq2 (ref. 29) and Wilcoxon signed-rank test) and treating each patient individually (GFOID (ref. 28)) followed by recurrence count.

In DESeq2 and Wilcoxon signed-rank test, significantly differentially expressed lncRNAs were defined as satisfying two criteria: log2 (fold-change) > 1 and q-value (Benjamini-Hochberg adjusted P value) < 0.05. In GFOID (V1.1.3), which was especially useful for assessing samples without replicates, differentially expressed lncRNAs were first identified for each individual patient. GFOID calculated its own statistics (that is, significance cutoff and
GFOLD value) for expression level changes based on the posterior distribution of the log fold-changes in expression. The significance cutoff was set at 0.01 (< q-value > 0.01), also requiring [GFOLD value]>1. After differentially expressed lncRNAs were identified in individual patients using GFOLD (Supplementary Fig. 6), lncRNAs that were not recurrently deregulated in multiple patients were filtered out from the results, yielding 1,112 lncRNAs that were recurrently differentially expressed in at least 8 of 20 patients (permutation test, FDR < 0.05) when primary tumours were compared with adjacent normal tissues, as well as 107 lncRNAs that were recurrently differentially expressed in at least 4 patients (permutation test, FDR < 0.05) when PVTVs were compared with matched primary tumours (Supplementary Fig. 7). We used a permutation test to estimate the distribution of the number of recurrences among multiple patients. First, we randomly sampled size-matched lncRNAs for each patient according to the number of differentially expressed lncRNAs identified by GFOLD and calculated the number of recurrences. Next, we repeated the sampling and counting 10,000 times to estimate a null distribution for the number of recurrences. Finally, we calculated the FDRs for the recurrences based on the null distribution. We set the minimum number of recurrences to 8 and 4 for primary tumours versus adjacent normal tissues and PVTVs versus primary tumours, respectively, to ensure that the FDR of each recurrence was smaller than 0.05.

**Integrative analyses of TCGA LIHC data.** We downloaded RNA-seq data for 157 LIHC patients with clinical data in the TCGA from the NCBI Cancer Genomics Hub (CGHub). We calculated the expression level of each gene/lncRNA (FPKM) for each TCGA sample using Cufflinks. In the Kaplan-Meier survival analysis, the survival data included vital status, days to death, and other variables, which were available for 151 of 157 patients. We first divided the samples into two groups (51 low-expression and 100 high-expression) according to the expression level of a marker gene/lncRNA (for example, RP11-166D19.1). Next, we used Kaplan-Meier survival analysis to perform a 5-year survival analysis via the survival package (https://cran.r-project.org/web/packages/survival) in the R environment for statistical computing. We computed significance using the log-rank test. Additional clinical information (age, gender, AFP, serum albumin, prothrombin time, cirrhosis, vascular invasion, and so on) for the LIHC patients in the TCGA was downloaded for the multivariate analysis. Based on two groups defined by the expression level of a particular lncRNA (for example, RP11-166D19.1), the odds ratio of each clinical criterion was calculated for each class of patients (low expression and high expression). A forest plot was drawn with odds ratios and 95% confidence intervals for each clinical criterion.

We used GSEA (v2.0.13) (ref. 63) to assess enrichment of sets of recurrently deregulated lncRNAs in other data sets. GSEA requires three input files: a gene set, expression data, and a phenotype label. Because the TCGA LIHC cohort had no PVTT or metastasis data, we used sample information (adjacent normal tissue, primary tumour and metastasis) as the gene set. Expression data were provided by Professor Jianzhong Xi from Peking University. All cell lines were derived from the total RNA-seq data for the 60 samples were used. Genes with a maximum expression level among all data sets that ranked in the bottom 20% were used as the gene set. We used GSEA to find the gene clusters associated with the expression of the genes of interest. We also annotated 271 driver genes of liver cancer, which were derived from a recent study.29

**Co-expression network construction.** We adapted a published method35 to construct a co-expression network of lncRNAs (including both GENCODE lncRNAs and lncRNAs identified in the HCC samples). The expression levels derived from the total RNA-seq data for the 60 samples were used. Genes with a maximum expression level among all data sets that ranked in the bottom 20% were used as the gene set. We used GSEA to find the gene clusters associated with the expression of the genes of interest. We also annotated 271 driver genes of liver cancer, which were derived from a recent study.39

**Knockdown of candidate lncRNAs and transwell migration assay.** Three human liver cancer cell lines, HepG2, SMCC-7721 and HCCLM9, were used to conduct functional assays (HepG2 was purchased from the American Type Culture Collection (ATCC, Manassas, VA, USA), SMCC-7721 and HCCLM9 were provided by Professor Jianzhong Xi from Peking University). All cell lines were tested for mycoplasma contamination and no contamination was found. Three siRNAs were designed for each candidate lncRNA (Supplementary Table 3) and obtained from Genepharma. qRT-PCR was used to monitor siRNA knockdown efficiency. Primer sequences used for qRT-PCR are listed in Supplementary Tables 3 and 4. Transwell migration assays were used to test the effects of candidate lncRNAs on cell migration. CCK8 cell proliferation assays were used to assess the effects of candidate lncRNAs on cell proliferation.

For the cell migration assays, cells were first transfected with 30 nM siRNA mixtures of three designed siRNAs in 24-well plates and incubated for 48 h, followed by resuspension and washing with phosphate-buffered saline buffer. Next, for each experiment, approximately 40,000 cells were seeded into the upper chamber of a transwell insert (pore size, 8 µm, Costar) in 100 µl of serum-free medium per well. Medium (600 µl) containing 10% serum was placed in the lower chamber to act as a chemottractant. The seeded cells were incubated for 24 h to allow them to migrate. Subsequently, non-migratory cells were removed from the upper chamber by scraping it with cotton. The cells remaining on the lower surface of the insert were fixed with 4% formaldehyde (Sigma) and stained with DAPI for counting. Each type of cell was assessed in triplicate. Moreover, we repeated the RNAi knockdown and migration assays for all candidate lncRNAs using each siRNA at the same concentration.

To assess whether cell proliferation affected cell migration activity in the transwell assays, we performed CCK8 cell proliferation assays after knockdown of candidate lncRNAs using each individual siRNA. For the cell proliferation assays, cells were transfected with 30 nM of each siRNA and allowed to grow for 48 h. Next, cells were incubated with 10% CCK8 reagent (DoInDi Laboratories, Japan) for 1 h at 37°C. The absorbance of the solution in each well at 450 nm was detected using an automatic spectrometer (Multimode Reader; Enspire). Each experiment was performed in triplicate.

**Data availability.** The high-throughput sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE77275 (RNA-seq, GSE77276; RNA-seq, GSE77599; small RNA-seq, GSE76905; 450k array, GSE77269; Cytoscan HD, GSE77275). The authors declare that all other data are available in the article and its Supplementary Information Files or from the corresponding author on reasonable request.
References

1. Jemal, A. et al. Global cancer statistics. CA Cancer J. Clin. 61, 69–90 (2011).
2. Aldrichetti, L. et al. Liver resection with portal vein thrombectomy for hepatocellular carcinoma with vascular invasion. Ann. Surg. Oncol. 16, 1254–1254 (2009).
3. Mitsuobu, M., Toyosaka, A., Oryiama, T., Okamoto, E. & Nakao, N. Intrahepatic metastases in hepatocellular carcinoma: the role of the portal vein as an efferent vessel. Clin. exp. metastasis 14, 520–529 (1996).
4. Uka, K. et al. Clinical features and prognosis of patients with extrahepatic metastases from hepatocellular carcinoma. World J. Gastroenterol. 13, 414–420 (2007).
5. Li, H. et al. Overexpression of IncRNA H19 enhances carcinogenesis and metastasis of gastric cancer. Oncotarget 5, 2318–2329 (2014).
6. Gupta, R. A. et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464, 1071–1076 (2010).
7. Panzitt, K. et al. Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. Gastroenterology 132, 330–342 (2007).
8. Lai, C. C. et al. Viral-human chimeric transcript predisposes risk to liver cancer development and progression. Cancer Cell 25, 335–349 (2014).
9. Yuan, J. H. et al. A long non-coding RNA activated by TGF-beta promotes the invasion-metastasis cascade in hepatocellular carcinoma. Cancer Cell 25, 666–681 (2014).
10. Huang, J. et al. Up-regulation of DLK1 as an imprinted gene could contribute to human hepatocellular carcinoma. Carcinogenesis 28, 1094–1103 (2007).
11. Wang, F. et al. Oncostal long noncoding RNA PVT1 promotes proliferation and stem cell-like property of hepatocellular carcinoma cells by stabilizing NOP2. Hepatology 60, 1278–1290 (2014).
12. Iyer, M. K. et al. The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. 47, 199–208 (2015).
13. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774 (2012).
14. Pruitt, K. D. et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 42, D756–D763 (2014).
15. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–1195 (2012).
16. Fatica, A. & Bozzoni, I. Long non-coding RNAs in human cancer. Nat. Rev. Genet. 14, 570–574 (2013).
17. Rehaji, S. et al. Recurrent PTPRB and PLCG1 mutations in angiosarcoma. Nat. Genet. 46, 376–379 (2014).
18. Shen, B., Sirotta, M., Fan-Minogue, H., Hadley, D. & Butte, A. J. Relating hepatocellular carcinoma tumor samples and cell lines using gene expression data in translational research. BMC Med. Genomics 8, 5 (2015).
19. Emmrich, S. et al. LincRNAs NONC and MIR100HG act as oncogenes in acute megakaryoblastic leukemia. Mol. cancer 13, 1–12 (2014).
20. Zhang, Z. T. et al. Predicting hepatitis B virus-positive metastatic hepatocellular carcinoma using gene expression profiling and supervised machine learning. Nat. Med. 9, 416–423 (2003).
21. Zhang, Z. Genomic landscape of liver cancer. Nat. Genet. 44, 1075–1077 (2012).
22. Liu, S. et al. MicroRNA-135a contributes to the development of portal vein tumor thrombus by promoting metastasis in hepatocellular carcinoma. J. Hepatol. 56, 389–396 (2012).
23. Luo, Z.-L. et al. A splicing variant of Merlin promotes metastasis in hepatocellular carcinoma. Nat. Commun. 6, 8457 (2015).
24. Andrews, S. R. & Rothnagel, J. A. Emerging evidence for functional peptides encoded by short open reading frames. Nat. Rev. Genet. 15, 193–204 (2014).
25. Jeng, D. M. et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell 160, 595–606 (2015).
26. Hsu, P. Y. et al. Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. Proc. Natl Acad. Sci. USA 113, E7126–E7135 (2016).
27. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111 (2009).
28. Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
29. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515 (2010).
30. Liu, L. et al. A common set of distinct features that characterize noncoding RNAs across multiple species. Nucleic Acids Res. 43, 104–114 (2015).
31. Gerstein, M. B. et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE Project. Science 330, 1775–1787 (2010).
32. Trapnell, C. et al. Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome Res. 21, 276–285 (2011).
33. Gerstein, M. B. et al. Comparative analysis of the transcriptome across distant species. Nature 512, 445–448 (2014).
34. Di, C. et al. Characterization of stress-responsive IncRNAs in Arabidopsis thaliana by integrating expression, epigenetic and structural features. Plant J. 80, 848–861 (2014).
35. Wilks, C. et al. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. Database 2014, baou93 (2014).
Acknowledgements
This work was supported by the National High-Tech Research and Development Program of China (2014AA021103), National Key Basic Research and Development Plan of China (2014YFA0500803), National Key Basic Research Program (2012CB316503), National Natural Science Foundation of China (31522030, 91329000, 91129000, 31671384, 91519326, 31361163004, 81422032, 81627280, 61370335, 31271402) and the Tsinghua University Initiative Scientific Research Program (2014Z21045). This work was also supported by the Computing Platform of National Protein Facilities (Tsinghua University). The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. We thank GenoArray (Suzhou 215123, China) for the RNAi experiment support.

Author contributions
Z.J.L., M.Q.Z., H.W. and L.C. conceived the project and established and coordinated the studies. Y.Y., J.G., J.Y., Q.L., Y.-C.T.Y., D.W. and L.H. analysed the data. L.C., G.L. and J.T. collected the clinical samples. H.Z. and J.X. performed the functional assay and experimental validation. All authors wrote the paper.

Additional information
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications
Competing financial interests: The authors declare no competing financial interests.
Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/
How to cite this article: Yang, Y. et al. Recurrently deregulated IncRNAs in hepatocellular carcinoma. Nat. Commun. 8, 14421 doi: 10.1038/ncomms14421 (2017).
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/
© The Author(s) 2017