Published in final edited form as:
Nat Methods. 2013 February ; 10(2): 155–161. doi:10.1038/nmeth.2331.

Lentiviral vector-based insertional mutagenesis identifies genes associated with liver cancer

Marco Ranzani[1], **Daniela Cesana**[1,*], **Cynthia C. Bartholomae**[2,*], **Francesca Sanvito**[3,4], **Mauro Pala**[5], **Fabrizio Benedicenti**[1], **Pierangela Gallina**[1], **Lucia Sergi Sergi**[1], **Stefania Merella**[1], **Alessandro Bulfone**[5], **Claudio Doglioni**[3,6], **Christof von Kalle**[2], **Yoon Jun Kim**[7], **Manfred Schmidt**[2], **Giovanni Tonon**[8], **Luigi Naldini**[1,6], and **Eugenio Montini**[1,†]

[1]San Raffaele-Telethon Institute for Gene Therapy, San Raffaele Scientific Institute, Milan, Italy.

[2]National Center for Tumor Diseases (NCT), im Neuenheimer Feld 350, 69120 Heidelberg, Germany

[3]Department of Pathology, San Raffaele Scientific Institute, via Olgettina 60, 20132, Milan, Italy.

[4]Mouse Histopathology, San Raffaele Scientific Institute, via Olgettina 60, 20132, Milan, Italy.

[5]Bio))flag Ltd., Piscinamanna 09010 Pula, Cagliari, Italy

[6]Vita Salute San Raffaele University, via Olgettina 58, 20132, Milan, Italy.

[7]Department of Internal Medicine and Liver Research Institute, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul 110-799, Korea

[8]Functional Genomic of Cancer Unit, San Raffaele Scientific Institute, via Olgettina 58, 20132, Milan, Milan, Italy

Abstract

Transposons and γ-retroviruses have been efficiently used as insertional mutagens in different tissues to identify molecular culprits of cancer. However, these systems are characterized by recurring integrations that accumulate in tumor cells, hampering the identification of early cancer-driving events amongst bystander and progression-related events. We developed an insertional mutagenesis platform based on lentiviral vectors (LVV) by which we could efficiently induce hepatocellular carcinoma (HCC) in 3 different mouse models. By virtue of LVV's replication-deficient nature and broad genome-wide integration pattern, LVV-based insertional mutagenesis allowed identification of 4 new liver cancer genes from a limited number of integrations. We validated the oncogenic potential of all the identified genes in vivo, with different levels of penetrance. Our newly identified cancer genes are likely to play a role in human disease, since they are upregulated and/or amplified/deleted in human HCCs and can predict clinical outcome of patients.

†To whom correspondence should be addressed: San Raffaele-Telethon Institute for Gene Therapy, via Olgettina 58, 20132, Milan, Italy. Phone: +390226433869; fax: +390226434621, montini.eugenio@hsr.it.
*These authors contributed equally to the work.

AUTHOR CONTRIBUTIONS M.R. designed and performed experiments and wrote the manuscript. D.C. performed all gene expression analyses and revised the manuscript. C.B. performed 454 pyrosequencing. F.S. and C.D. performed histopathological analyses. M.P., A.B. and G.T. analysed microarray data. F.B., P.G., L.S. and S.M. performed experiments. C.vK. and M.S. supervised 454 pyrosequencing and mapping of vector integrations. Y.J.K. provided clinical data of HCC patients. L.N. supervised the project and revised the manuscript. E.M. supervised the project and wrote the manuscript.

E.M. and L.N. share senior authorship.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts

INTRODUCTION

The approaches most frequently used to discover genes that are altered in cancer are high-throughput 'omics' technologies. However, since bystander lesions are also frequent, the cause-effect relationships of cancer-associated alterations, especially in late-stage tumors, are not always obvious[1]. Insertional mutagenesis approaches use oncoretroviruses or transposons to trigger cancer in mice by widespread integration into the cellular genome and activation of oncogenes near the integration site. Mapping the genomic integration sites in tumors allows the identification of genomic regions that are recurrently hit in independent tumors (defined as Common Insertion Sites, CIS), which host genes likely involved in cancer development[2].

We have shown that HIV-derived Lentiviral Vectors (LVVs) with Long Terminal Repeats (LTR) containing strong enhancer-promoter sequences are prone to induce insertional mutagenesis[3]. Since LVVs are able to efficiently transduce quiescent cells and a variety of tissues and organs in vivo[4-11], including liver[12], here we developed a LVV specifically tailored to induce HCC in mice (LV.ET.LTR) by activating and tagging cancer genes in hepatocytes. We used LV.ET.LTR to screen for liver cancer genes in three mouse models. First, we screened in *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice, which combine the high sensitivity to genotoxic mutations conferred by the *Cdkn2a* deficiency[13] with the high permissiveness to hepatocyte gene transfer by LVV conferred by the *Ifnar1* deficiency[14]. *CDKN2A* and its targets – pRB and p53 – are frequently inactivated or silenced in human HCCs[15]. Second, as the inflammatory microenvironment plays a fundamental role in the pathogenesis of human HCC[16], we used a mouse model of liver-specific *Pten* deficiency (*Pten* liver-null) that recapitulates several aspects of human non-alcoholic steatohepatitis and is characterized by chronic liver oxidative damage which, after a long latency period, results in the development of hepatic adenomas and HCCs[17]. *PTEN* expression is reduced or absent in almost 50% of human HCCs and it is associated to a poor prognosis[18]. Third, we set up an experimental model of chronic liver injury in wild type (WT) mice by carbon tetrachloride (CCl_4) administration, which results in waves of hepatocyte necrosis and regeneration that cause liver damage without progression to cancer. By LVV-based insertional mutagenesis we induced HCC in these three mouse models and identified four HCC genes that figure prominently in human hepatocarcinogenesis.

RESULTS

LVV-based insertional mutagenesis

We constructed a transgeneless LVV with highly-active hepatospecific enhancer-promoter sequences (Enhanced Transthyretin, ETr[19]) in the LTR (LV.ET.LTR, Fig. 1a) in order to activate genes upon integration in hepatocytes and avoid unwanted effects in non-parenchymal cells. LV.ET.LTR was administered to newborn mice by temporal vein injection, a protocol chosen because substantial levels of hepatocytes transduction can be achieved by a single injection (up to 60% of hepatocytes, Supplementary Fig. 1a). We tested LV.ET.LTR in three different mouse models of hepatocarcinogenesis: *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice, *Pten* liver-null mice and WT mice with or without CCl_4 treatment (Fig. 1b, Supplementary Fig. 1b-f, see also Online Methods).

Upon LVV-administration, mice of all three models developed HCCs at a frequency significantly higher than genotype-matched control mice (Fig. 1c-f and Supplementary Fig. 1g-n). All the HCCs that arose in LVV-treated mice were vector-marked (Supplementary Table 1). LVV integrations were retrieved from 30 LVV-induced liver tumors by Linear Amplification Mediated (LAM)-PCR (Supplementary Fig. 2a), resulting in a total of 172 unique integration sites (Supplementary Table 2a).

We considered as putative HCC causal genes those recurrently targeted by LVV integrations in independent tumors at a frequency significantly higher than expected for a random distribution (defined as CIS). Based on previous statistical definitions[20, 21], four CIS were identified and targeted *Fign* (targeted by 9 LVV integrations), *Braf* (4 integrations), *Sos1* (4 integrations) and the *Dlk1-Dio3* region (9 integrations) (Fig. 2a-d and Supplementary Table 2a). None of the CIS found in tumors was targeted by LVV integrations retrieved from tumor-free livers ($n = 162$) and no CIS were identified from these control dataset of insertions (Supplementary Table 2b), indicating that the CIS identified in HCCs are not determined by an intrinsic genomic integration bias of LVV in hepatocytes.

All the integrations within CISs were in the same transcriptional orientation as the targeted gene. By RT-PCR we detected chimeric LVV-CIS gene fusion transcripts which contained LVV sequence from the transcription start site in the 5′LTR to the major HIV splice donor site fused to the splice acceptor site of an exon of the targeted gene and its remaining coding sequence (Fig. 2a-d and Supplementary Fig. 2b). In the case of *Braf* and *Sos1*, the products encoded by these fusion transcripts are truncated proteins with increased activity due to the lack of the N-terminal regulatory domains[22, 23] (Fig. 2e and Supplementary Fig. 2c). LVV-*Fign* fusion transcripts encode for a putative FIGN protein lacking 11 amino acids from the N-terminus. LVV integrations within the *Dlk1-Dio3* region generated fusion transcripts with the full-length *Rtl1* transcript.

The newly identified CIS genes can cause HCC

We tested the oncogenic potential of the newly identified putative cancer genes upon forced expression in mouse hepatocytes *in vivo*. We generated LVVs with self-inactivating (SIN) LTRs in which the expression of the putative oncogene is regulated by the ETr promoter in internal position and by a 3′UTR bearing target sequences for microRNA-142. These SINLV constructs allow high levels of transgene expression restricted to hepatocytes[24], while the SIN LTR design prevents insertional mutagenesis[3]. Newborn *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ ($n = 25$) or WT mice ($n = 6$) were systemically injected with SINLV preparations (2×10^7-4×10^8 transducing units (TU) per mouse; see Online Methods) (Fig. 2f and Supplementary Table 3). One of five *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice treated with the vector encoding for truncated SOS1 and two of seven *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice treated with vector encoding for truncated BRAF developed multifocal HCCs. All *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice (nine out of nine) transduced with the vectors coding for full length or truncated FIGN developed multiple early-onset HCCs that were lethal by 9 weeks of age. While unable to induce HCCs in *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice, *Rtl1* overexpression induced HCC in two out of four WT mice treated with CCl_4. SINLVs expressing full length SOS1 (n = 4 *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice) or the neutral EGFP gene (n = 8 *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice and n = 2 WT mice treated with CCl_4) did not detectably induce tumors (Fig. 2g-l, Supplementary Fig. 2d and Supplementary Table 3)

The newly identified cancer genes dictate the HCC phenotype

Integrations targeting *Braf*, *Rtl1* and *Fign* were found mainly in independent tumors and significantly associated to grade 3, 2 and 1 HCCs, respectively (*Braf* P = 0.0026, *Rtl1* $P < 0.0001$ and *Fign* P = 0.0016, two tailed Fisher's exact test). HCCs with integrations targeting *Sos1* were mainly grade 1. *Braf* integrations were found only in HCCs from *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice (P = 0.0004) (Fig. 3a).By microarray we interrogated the whole transcriptome of 21 HCCs and 8 non-tumor livers (GEO: GSE31409). CIS genes targeted by integration always showed a significant upregulation compared to other HCCs or normal livers (Fig. 3b and Supplementary Fig. 3a-b). Analyzing the signal intensity of the single probes spanning the mRNAs of the different CIS genes, we verified that vector integrations

within *Braf*, *Fign* and *Sos1* induced the significant overexpression of transcripts containing only exons downstream the integration (Fig. 3c, P-values by unpaired t-test).

Hierarchical unsupervised clustering identified four main clusters, each composed of HCCs with the same CIS, such as *Rtl1* (cluster of 8 HCCs out of 8 HCCs harboring integrations targeting *Rtl1*), *Braf* (3 out of 3), *Sos1* (2 out of 3) and *Fign* (5 out of 7) (Fig. 4a). By Gene Set Enrichment Analysis (GSEA) (see Online Methods) we compared the expression profiles of *Braf, Fign* and *Rtl1* HCCs to the profiles of non-tumor tissues and observed the common downregulation of genes involved in hepatic metabolism and the upregulation of genes involved in cancer, cell cycle, growth and proliferation (Supplementary Fig. 3c-d and Supplementary Table 4a-b). We also found relevant differences among the different groups (Fig. 4b-c, Supplementary Fig. 3e-g). *Fign* and *Braf* HCCs shared the upregulation of *E2f* and *Yy1* transcriptional targets (Supplementary Fig. 3g and Supplementary Table 4a-b) and the downregulation of oxidative phosphorylation genes (Supplementary Fig. 3e). Several oxidative phosphorylation genes and *Sf1* transcription factor target genes were upregulated in *Rtl1* HCCs with respect to other groups and normal liver (Fig. 4b, Supplementary Fig. 3e and Supplementary Table 4a-b). WNT pathway was exclusively upregulated in *Fign* HCCs (Fig.4c).

The novel cancer genes are relevant in human HCCs

We analyzed a microarray dataset (MSSM collection[25]) constituted by 75 expression profiles from normal and diseased livers and HCV-induced HCCs. We also performed RT-Q-PCR on HCCs and normal livers from a tissue collection of our institution (HSR collection) (Fig. 5a-c). *SOS1* was upregulated in 60-70% of all HCCs of both collections, a frequency significantly higher than non-tumor samples ($P < 0.01$ and $P = 0.02$ in MSSM[25] and HSR collections, respectively; two tailed Fisher's exact test). In both collections *SOS1* was significantly upregulated in HCCs with respect to non-tumor liver ($P < 0.001$ in MSSM[25] and HSR collections, unpaired t-test) (Fig. 5a). Furthermore, we found a significant *SOS1* upregulation by analyzing an independent HCC microarray dataset[26] ($P < 0.0001$, Supplementary Fig. 4a).

FIGN in MSSM HCCs[25] was overexpressed at a frequency significantly higher than non-tumor samples ($P = 0.008$; Fig. 5b). *FIGN* is embedded in a chromosomal region that is amplified in human HCCs (Fig. 5d) and it is also significantly upregulated in human glioblastoma, astrocytoma, oligodendroglioma, melanoma, testicular teratoma, and in ovarian endometriosis (Supplementary Fig. 4b-c). In the HSR collection, *BRAF* expression is significantly higher in HCCs compared to normal liver ($P < 0.05$) (Fig. 5c). Moreover, the *BRAF* locus mapped within a chromosomal region that is amplified in 5 independent studies on human HCCs (Fig. 5e). We did not determine the expression level of *RTL1* because of the absence of probes in human microarrays and the lack of a RT-Q-PCR assay that reliably detects the transcript. We found that the chromosomal region containing *RTL1* was lost in 4 independent HCC studies (Fig. 5f).

We tested if the gene expression signatures found in our mouse tumor cohorts by Significance Analysis of Microarray (FDR < 0.01, see Online Methods) also had relevance in human hepatocarcinogenesis. GSEA analysis showed that these signatures were significantly enriched in HCCs compared to healthy tissues or non-tumor diseased tissues in MSSM collection[25]. The genes downregulated in murine HCCs with *Braf* or *Rtl1* integration were also downregulated in human liver diseases (cirrhosis, dysplasia and HCCs) compared to normal liver. The genes upregulated in HCCs with *Fign* integrations were significantly enriched in human disease samples (Supplementary Fig. 5a-b). We then considered the expression levels of the human orthologs of the merged lists of genes from murine CIS-specific gene expression signatures for unsupervised clustering analysis on human liver

samples[25]. By unsupervised clustering (see Online Methods) the human samples were grouped accordingly to the different phenotypes (normal, cirrhotic, dysplastic, HCC) and to the different HCC stages (Supplementary Fig. 5c).

We investigated the clinical outcomes in a dataset of 70 HBV-induced HCCs[27]. Patients with HCC with higher expression of *SOS1* had a significantly increased overall survival compared to patients with HCC with lower *SOS1* levels (Fig. 5g). We observed a similar trend in the disease-free survival (Supplementary Fig. 5d). Finally, we used the gene expression signature associated to each of the novel cancer genes to classify HCC patients[27] by clustering analysis (see Online Methods). *Fign, Rtl1* and *Braf* signatures identified subgroups of HCC patients characterized by a significantly decreased disease-free survival. These HCC subgroups with poor prognosis displayed aberrant gene expression patterns reminiscent of the gene expression alterations found in the murine HCCs (Fig. 5h-i and Supplementary Fig. 5e-j).

DISCUSSION

We used for the first time, to our knowledge, a replication-defective LVV as insertional mutagen to induce cancer in mice and to identify molecular culprits of cellular transformation. We efficiently induced HCCs in three different mouse models, generating a collection of 30 HCCs which covered the different grades of the disease (from G1 to G3) and displayed gene expression signatures reminiscent of those in human HCCs. We identified four CIS (from 172 integrations retrieved from 30 tumors) and observed a strong enrichment for tumors bearing a single CIS integration (83%), when compared to other insertional mutagenesis systems[22, 28-31]. This high efficiency of CIS retrieval is likely to be the combined result of the non-replicative nature of LVV, the high efficiency of LVV transduction in hepatocytes, and the efficient coverage of genes by integrations.

LVV integrations are produced in a short time window after injection and before the *in vivo* selection of transformed clones occurs. This results in a lower total number of integrations than γ-retroviruses and transposons[22, 28-31], which may reduce the incidence of tumor induction and the total yield of identified cancer genes. While this may represent a limitation of LVVs compared to transposon and γ-retroviruses as insertional mutagens, this characteristic of LVVs may also facilitate the identification of early-mutated genes in carcinogenesis since it eliminates bystander and progression-related integrations. Identification of early lesions that initiate and drive cellular transformation may be important to unravel pathways essential to the neoplastic phenotype since they may represent early key steps in transformation. Differently from previous studies[22, 28, 30, 31], three of the cancer genes we identified have not so far been causally implicated in HCC. However, our murine data-driven reanalysis of human HCC data showed that all 4 of the newly identified genes have a clinical relevance in the human disease as well.

Transposable elements are insertional mutagens that, despite overcoming the limited tissue tropism of γ-retroviruses, require the generation of multiple knock-in or transgenic mouse strains. Conversely LVVs have wide tissue tropism[4-11] and can be engineered to be oncogenic in different tissues by adapting the specificity of the enhancer-promoter sequences, as we have shown in liver and hematopoietic tumors[3]. Therefore our system complements and extends the insertional mutagenesis screening performed with retroviruses and transposons.

Although they represent a promising insertional mutagenesis tool, LVVs have some potential limitations. First, with LVVs, which are replication deficient, an extensive transduction of the targeted organ is required to obtain significant levels of mutagenesis and

eventually cell transformation. Therefore, LVV-mediated insertional mutagenesis may be inefficient in organs that are difficult to access. Second, as with other insertional mutagens, LVV display a bias for the detection of gain-of-function mutations, leading mainly to the identification of oncogenes rather than tumor suppressor genes. Alternative oncogenic LVV designs may be required to overcome these limits. Third, LVVs display integration biases toward expressed genes and gene-dense regions, which could skew the repertoire of identified oncogenes.

Although we could qualitatively validate the oncogenic potential of the four cancer genes by in vivo SINLV-mediated forced expression in hepatocytes, the incidence of HCC induction was low in some cases. Several variables could explain this low penetrance: (i) some animals may have been transduced at lower levels; (ii) some oncogenes, to induce cell transformation, may require very high levels of expression that are reached in only a minority of hepatocytes in vivo; (iii) conversely, very high SINLV-mediated overexpression of constitutively active oncogenes could induce apoptosis and/or counterselection even in *Cdkn2a*$^{-/-}$ *Ifnar1*$^{-/-}$ hepatocytes. Additional studies aimed at quantitatively studying the oncogenic potential of these liver cancer genes are warranted.

The possibility to study HCC induced by integration in a single CIS allowed us to explore the molecular mechanisms of deregulation of the targeted gene. LVV integrations targeting *Braf* induced aberrantly spliced mRNAs encoding for a constitutively active protein which was previously reported in insertional mutagenesis studies describing sarcomas in *Arf*$^{-/-}$ mice[22] and a myeloid tumor in *Cdkn2a*$^{-/-}$ mice[3]. These data indicate that BRAF activation cooperates with *Cdkn2a* deficiency to induce cell transformation in different tissues. Furthermore, mutations which constitutively activate BRAF protein were recently found in human HCCs[32]. Finally, the aberrant BRAF form we observed is similar to those found in human thyroid carcinoma, melanoma, prostate and gastric cancer as a result of translocations or microdeletions[33, 34].

We showed that ectopic expression of truncated SOS1 protein was able to induce HCCs, while the wild type form was not. *SOS1* was overexpressed in human HCCs and high levels of *SOS1* expression correlated with significantly increased overall survival, making *SOS1* a candidate prognostic marker for human HCC. Consistent with the human data, murine HCCs induced by LVV-mediated *SOS1* truncation were mainly grade 1, less aggressive tumors. Loss of function mutations in *RPS6KA3* gene[35], which encodes for an inhibitor of SOS1 signaling[36, 37], were recently found in human HCCs, further suggesting that activation of SOS1-BRAF axis has a relevant role in human hepatocarcinogenesis. Moreover, clinical trials have shown that Sorafenib, a multikinase inhibitor acting on BRAF, PDGFR and VEGFR, improves the survival of patients with HCC[38]. This may be at least in part mediated by the inhibition of hyper-activated BRAF or SOS1 signaling.

We identified *Fign* as a target in all mouse models tested, and its overexpression in *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice triggered rapid HCC onset with 100% penetrance. The oncogenic potential of the truncated FIGN protein is indistinguishable from that of the full-length protein upon LVV-mediated liver gene transfer, suggesting that FIGN overexpression itself has a major impact in oncogenesis. We found that the WNT pathway was specifically deregulated in HCCs induced by integration in *Fign*. This may shed some light over the putative effectors of this enigmatic gene that we found to be highly relevant in human tumors.

Differently from previous studies which showed deregulation or targeting of maternally expressed genes within the imprinted *Dlk1-Dio3* region[39, 40], we found that LVV integrations targeting the region induced the overexpression of the paternally expressed gene *Rtl1* in HCCs. Forced expression of *Rtl1* induced HCCs in two out of four wild type mice

treated with CCl_4, validating the role in hepatocarcinogenesis of this gene, whose pathway and functions are still elusive.

In summary, we developed LVVs carrying highly-active enhancer-promoter sequences in the LTRs that are genotoxic in hepatocytes and used these tools as insertional mutagens to identify four genes implicated in HCC. The intrinsic versatility, the wide tissue tropism and the high *in vivo* transduction efficiency of LVVs will permit effective insertional mutagenesis for the screening of early tumorigenic events in different tissues. Our approach should help identify candidate prognostic markers and therapeutic targets for human HCC and other tumors.

ON LINE METHODS

Vector production

We cloned the transfer plasmid for the production of LV.ET.LTR as it follows. We eliminated the expression cassette containing hPGK and EGFP from pCCLSIN.cPPT.hPGK.EGFP.wPRE[41] by removing a 1289 bp XhoI-SalI fragment. We obtained the intermediate plasmid pCCLSIN.cPPT.wPRE after blunting the DNA ends and performing intramolecular re-ligation. To eliminate a residual ORF within the mutated wPRE sequence (mwPRE)[42], we included by PCR an additional stop codon and then we cloned the 599 bp product in SalI-EcoRI of pCCLSIN.cPPT.wPRE, thus generating pCCLSIN.cPPT.mwPRE. We blunt-cloned a 31 bp polylinker containing PstI, BamHI, EcoRV, XbaI, NsiI sites at the BbsI site in the –18 SIN LTR. We amplified by PCR a 632 bp a fragment containing the ETr enhancer-promoter sequence from pCCLsin.cPPT.ET.EGFP.wPRE[24] and carrying NsiI and SpeI at the 5′ and 3′ ends. Then we cloned it at the PstI and XbaI sites of the previously described polylinker. We then used the resulting plasmid pCCL.ET.LTR.cPPT.mwPRE (see also Supplementary Note 1) for the production of LV.ET.LTR.

For liver gene transfer of putative cancer genes, we replaced in pCCLsin.cPPT.ET.hFIX.wPRE.142-3pT[24] the wPRE sequence with the mutated mwPRE with the strategy described above. The full length Open Reading Frame (ORF) of *Sos1*, *Fign*, *Rtl1* and the truncated ORF of *Braf*, *Sos1* and *Fign* starting from the exon downstream vector integrations, were amplified by RT-PCR on RNA from murine liver with primers that added the restriction sites for MluI and SalI at the 5′ and 3′ of the ORFs, respectively. We then cloned the PCR products in MluI-SalI digested pCCLsin.cPPT.ET.hFIX.mwPRE. 142-3pT plasmid, thus replacing hFIX transgene with the ORF of the candidate cancer gene. In these LVVs carrying self-inactivating LTRs, the expression of the transgene is regulated by the hepatospecific ETr enhancer-promoter in internal position and by a 3′UTR bearing target sequences for microRNA 142 (microRNA142-target sequences). These SINLV constructs allow high levels of transgene expression in hepatocytes by the activity of the ETr promoter cloned in an internal position, while the SIN LTR design prevents the occurrence of insertional mutagenesis[3]. The microRNA142-target sequences prevent any leaky expression of the transgene in hematopoietic cells in which microRNA 142 activity is high (i.e. Kupffer cells), thus avoiding a confounding transformation of non-hepatocyte cells and preventing immune response against the transgene which may cause clearance of the transduced cells[24].

In order to generate vectors expressing neutral transgene as negative controls, we used pCCLSIN.cPPT.hPGK.EGFP.wPRE[41] to produce SINLV.PGK.EGFP.

We produced concentrated LVV stocks, pseudotyped with the VSV-G envelope, by transient co-transfection of four plasmids in 293T cells and titered on 293T cells as described[43].

Mouse models generation and characterization

Cdkn2$^{-/-}$ *Ifnar1*$^{-/-}$ mice were generated to couple the high sensitivity to genotoxic mutations conferred by the *Cdkn2a* deficiency[3, 41] to the high permissiveness to liver gene transfer by LVV conferred by the *Ifnar1* deficiency[24]. Additionally, this non inflammatory tumor prone mouse model has a clinical relevance, since *CDKN2A* and its targets – pRB and p53 – are frequently inactivated or silenced in human cancer [44] including HCCs [15]. *Cdkn2*$^{-/-}$ (C57BL6/J) mice were obtained from NCI-Frederick MMHCC Repository, while *Ifnar1*$^{-/-}$ (129SVEV) mice were obtained from B&K Universal Limited. F1 *Cdkn2a*$^{+/-}$*Ifnar1*$^{+/-}$ mice were generated by crossing *Cdkn2*$^{-/-}$ mice with *Ifnar1*$^{-/-}$ mice. By further crossing of F1 *Cdkn2a*$^{+/-}$ *Ifnar1*$^{+/-}$ mice, F2 mice were generated with mendelian ratios for each genotype; by allele-specific PCR screening, F2 *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice were identified (Supplementary Fig. 1b) and further crossed to get F3 mice that were used for the insertional mutagenesis experiments. The phenotype of *Cdkn2*$^{-/-}$ [13] and *Ifnar1*$^{-/-}$ [45] mice was previously described. *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice were grown in the mouse facility and their survival curve was characterized. Survival curves of *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ and *Cdkn2*$^{-/-}$ *Ifnar1*$^{+/-}$ mice were overlapping (median survival 255 and 229 days, respectively, Supplementary Fig. 1c). The survival curve of *Cdkn2a*$^{-/-}$ *Ifnar1*$^{+/+}$ mice was previously described (median survival ≈ 250 days)[8, 41] and it is overlapping the ones of *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ and *Cdkn2*$^{-/-}$ *Ifnar1*$^{+/-}$ mice. Therefore *Ifnar1* deficiency does not influence the development of *Cdkn2a* knockout-driven tumors. Histopathological analyses of untreated *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice showed that they develop the same spectrum of hematopoietic malignancies and sarcomas described for *Cdkn2*$^{-/-}$ mice[13].

Mice with *Pten* tissue specific knockout in hepatocytes were generated since they provide a model of inflammatory carcinogenesis associated to steatosis that mimics the non-alcoholic steatohepatitis disease[17] described in humans. Moreover, the model has also a clinical counterpart, since *PTEN* expression is reduced or absent in several human cancers, including almost 50% of advanced HCCs and it is associated with a poor prognosis[18]. *Pten*$^{flox/flox}$ (129S4) mice that carry exon 5 of *Pten* surrounded by loxP sequences were obtained from Jackson Laboratories Mouse Repository. *AlbCre*$^{+}$ (C57BL6/J) mice that express Cre recombinase under the control of the hepatocyte-specific Albumin promoter were obtained from Weizmann Institute of Science. By crossing *Pten*$^{flox/flox}$ with *AlbCre*$^{+}$ mice, F1 *AlbCre*$^{+}$ *Pten*$^{+/flox}$ mice were generated. By further crossing of *AlbCre*$^{+}$ *Pten*$^{+/flox}$ mice, F2 mice were generated with mendelian ratios for each genotype; by allele-specific PCR screening, F2 *AlbCre*$^{+}$ *Pten*$^{flox/flox}$ mice were generated (Supplementary Fig. 1d, left). By PCR, the liver specific deletion of *Pten* exon 5 was confirmed (Supplementary Fig. 1d, right). By further crossing of *AlbCre*$^{+}$ *Pten*$^{flox/flox}$ mice, experimental *AlbCre*$^{+}$ *Pten*$^{flox/flox}$ mice (*Pten* liver-null throughout the text) were generated and used for the insertional mutagenesis experiments. *AlbCre*$^{+}$ *Pten*$^{flox/flox}$ mice were grown in our mice facility and euthanized at different age to investigate the phenotype. *AlbCre*$^{+}$ *Pten*$^{flox/flox}$ mice generated in our laboratory matched the phenotype of a previous *Pten* liver-null model [17], recapitulating the features of non-alcoholic steatohepatitis (Supplementary Fig. 1e) which, after 40 weeks of age, progresses to hepatocellular adenoma and HCC (Supplementary Fig. 1f). The sequences of the primers used to perform PCR-based mouse genotyping are provided in Supplementary Table 5.

The effects of carbon tetrachloride (CCl_4) administration on the mouse liver are well-known [46, 47]. From the previous breeding, also *AlbCre*$^{-}$ mice were generated and included in the experimental outline as wild type mice. Therefore, we setup an experimental model of chronic liver injury in wild type mice by CCl_4 administration which results in waves of hepatocytes necrosis and regeneration that cause liver damage without progression to cancer. In our rationale, in a chronic inflammatory microenvironment caused by CCl_4 treatment, cell clones harboring genotoxic LVV integrations will acquire additional

synergizing genetic lesions which ultimately will be selected and lead to malignant transformation. Eight weeks old wild type mice transduced or not with LV.ET.LTR at neonatal stage, were administered CCl_4 1mg/kg twice weekly for 6 weeks in a 10% mineral oil solution (Sigma). Histopathological analyses on the livers of mice analyzed at different time points after the end of the treatment showed that CCl_4 treatment induced mild steatosis and chronic inflammation (prototypical example in Fig. 1f). The treatment with CCl_4 did not influence mice survival by 1 year of age.

An additional panel of LVV-induced HCCs in the different mouse models is shown in Supplementary Figure 1g-n.

Mouse treatments

For insertional mutagenesis experiments, newborn (24-48 hours old) mice from the three different genotypes (*Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$; *AlbCre*$^{+}$ *Pten*$^{flox/flox}$ and wild type) were administered 10μl of highly concentrated LV.ET.LTR preparations by temporal vein injection (10^8 TU/mouse). In order to generate a dataset of unselected integrations, newborn mice of the three different genotypes (tot $n = 5$) were administered LV.ET.LTR as described above and then euthanized at 2 weeks of age for liver samples collection.

To validate the oncogenicity of the newly identified cancer genes by liver gene transfer, newborn *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice or wild type mice were administered 20μl of highly concentrated LVVs (2×10^7 - 4×10^8 TU/mouse, according to vector titer, see Supplementary Table 3) that express the different cancer genes specifically in hepatocytes (see above). Wild type mice were then administered the CCl_4 regimen as described above.

In our screening for cancer genes, we decided to transduce newborn mice in order to increase our chances of success in inducing HCC by LVV-based insertional mutagenesis. In a newborn mouse the hepatocytes are highly proliferating, thus promoting the accumulation of additional genetic and epigenetic lesions that may complement with the “time-zero” integrations in leading hepatocyte to transformation. Since carcinogenesis is a multistep process and the LVV integration just provide an early event, we transduced newborn mice with highly proliferating hepatocytes to favor the accumulation of mutation that usually take years in the natural history of human hepatocarcinogenesis. Additionally, it was previously reported that the newborn mice are more sensitive than adult mice to hepatocyte transformation[42].

All mice were bred and kept in a dedicated pathogen-free animal facility, and were euthanized when they showed signs of severe sickness or at the defined time points: 20-30 weeks of age for *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mice, before significant mortality due to spontaneous hematopoietic malignancies is reached; 35 weeks of age for *AlbCre*$^{+}$ *Pten*$^{flox/flox}$ mice, to examine the effect of vector administration before the incidence peak of spontaneous liver tumors resulting from the genetic background; 52 weeks of age for wild type mice, since no spontaneous tumors are expected. All procedures were performed according to protocols approved by the Animal Care and Use Committee of the San Raffaele Institute (IACUC 353 and 463) and communicated to the Ministry of Health and local authorities according to Italian law.

Mouse sample collection and histopathology

By autoptical analysis we could identify grossly appearing masses in the liver parenchyma that were collected independently as well as non tumoral liver for microscopic and molecular analyses. One or two HCCs per liver were detected in *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$mice at euthanasia, while up to five and up to seven liver tumors were collected from LVV-treated *Pten* liver-null and wild type mice, respectively. Samples for DNA and RNA extraction were

also collected from normal liver parenchyma. A sample for DNA extraction was collected from every liver mass that was identified at autopsy; tumoral margins were estimated by gross appearance. A sample for RNA extraction was collected only from the tumoral masses whose size allowed the sampling (> 3mm diameter).

For histopathological analysis, normal liver lobes and all the collected liver masses were fixed in buffered 4% formalin, embedded in paraffin and 3μm sections stained with hematoxylin and eosin. Each specimen was evaluated in blinded fashion and independently by two pathologists (Francesca Sanvito, Claudio Doglioni) with expertise in human and mouse histopathology. The liver tumors were classified according to World Health Organization Classification of Tumors[48] and were graded according to the modified Edmonson-Steiner grading system[49]. Photomicrograghps were taken using the AxioCam HRc (Zeiss) with the AxioVision System 6.4 (Zeiss). Only samples collected from the liver masses that were identified as HCC by histopathology were used for the molecular analyses.

We evaluated the association between the integration within a specific CIS and HCC grade or the genetic background (see Fig. 3a) by two tailed Fisher's exact test. HCCs with integrations targeting *Sos1* were mainly of G1 (3 out of 4 HCCs with *Sos1* integration) but, given the small number of HCCs with *Sos1* integration, the co-occurrence of integration at *Sos1* and *Rtl1* in one HCC (that was grade 2) and the concomitant association between *Fign* integration and G1, no significant association to the tumor grade was found in this HCC collection ($P = 0.2903$, two tailed Fisher's exact test).

AlbCre^{+} *Pten*$^{flox/flox}$ mice developed also adenoma both from hepatocellular (mainly) and cholangiocellular (rare) origin. However, due to the high incidence of spontaneous adenomas in *AlbCre*$^{+}$ *Pten*$^{flox/flox}$ untreated mice (Supplementary Fig. 1f), only the hepatocellular carcinoma found in the LVV-treated *AlbCre*$^{+}$ *Pten*$^{flox/flox}$ mice were used for the further analyses. We did not detect any cholangiocellular carcinomas in *AlbCre*$^{+}$ *Pten*$^{flox/flox}$ at 35 weeks of age.

Evaluation of liver transduction upon systemic LVV administration by immunofluorescence

To test liver transduction by LVV systemic administration, we administered 5×10^7 TU/mouse of LV.PGK.EGFP[41] to 1 day old *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$mice ($n = 5$) that were then euthanized at 2-6 weeks of age. We fixed liver samples were in 4% paraformaldehyde and equilibrated them in sucrose gradients for inclusion in OCT compound and freezing. We blocked 20 μm sections in PBS containing 1% bovine serum albumin and 5% fetal bovine serum. We performed the staining with unconjugated rabbit anti-EGFP primary antibody (A11122, Invitrogen, dilution 1:200) and revealed by AlexaFluor488 donkey anti-rabbit secondary antibody (A21206, Invitrogen, dilution 1:500); cell nuclei were labeled by TO-PRO-3 (Invitrogen, 1:10,000). Confocal microscopy used an Axioskop 2 plus direct microscope (Zeiss) equipped with a Radiance 2100 three-laser confocal device (Bio-Rad). Percentage of transduced hepatocytes was calculated as EGFP positive polygonal cells with round-shaped euchromatic nuclei (i.e. hepatocytes) among total cells with round-shaped euchromatic nuclei. Three independent 20X field were analyzed for each mouse (about 2,000 total hepatocytes *per* mouse). Administration of 5×10^7 TU/mouse of LV.PGK.EGFP resulted in an efficient transduction of the liver parenchyma, up to 60% of total hepatocytes.

Vector copy number analysis

We extracted genomic DNA from normal liver and liver masses using the Qiagen blood and cell culture and tissue DNA Kits (Qiagen). We performed Q-PCR analysis with primers and probes complementary to mouse genomic *β-actin* and a common LVV sequence in the Ψ-

signal region as described[41]. VCN was determined as the ratio between the relative amounts of LVV versus total DNA (number of diploid genome) evaluated by *β-actin*. This calculation is possible since hepatocytes are still diploid in newborn mice at the time of transduction and then the integrated provirus replicates together with the cellular genome. A standard curve was made using dilutions from murine DNA with a known LVV VCN determined by Southern blot[50]. Reactions were carried out according to manufacturer's instructions and analyzed using the ABI Prism 7900 HT Sequence Detection System (Applera – Life Technologies).

LAM-PCR and genomic integration site analysis

LAM-PCR was performed as described[51] on all the histopathologically confirmed HCCs and representative non-tumor control samples. In order to favor the amplification of integrations occurring in the putatively oligoclonal tumor parenchyma versus the ones occurring in the tumoral stroma and contaminating surrounding tissue, we used an *ad hoc* designed LAM-PCR amplification protocol that uses limiting amounts of DNA to favor the amplification and retrieval of dominant insertions. For tumor samples, we used different amounts of DNA as template for LAM-PCR, according to the VCN that was detected in the sample by Q-PCR: 100 ng if VCN < 1; 50 ng if VCN between 1 and 3; 10 ng if VCN > 3. For non-tumor samples, 100 ng of DNA was always used as template for LAM-PCR. LAM-PCR was initiated with a 25-cycle linear PCR and restriction digest using Tsp509I, or HpyCHIV4. LAM-PCR primers for LVV were previously described[52-55]. LAM-PCR amplicons were separated on spreadex gels (Elchrom Scientific) to evaluate PCR efficiency and the bands pattern for each sample. Products of the second exponential amplification were tagged and then high-throughput sequenced with the 454 GS Flx platform (Roche).

Sequences were aligned to the mouse genome (assembly July 2007, mm9) using the NCBI BLAST genome browser (http://blast.ncbi.nlm.nih.gov/Blast.cgi) coupled to bioinformatic analyses. Identification of the nearest gene was performed by bioinformatic analyses. A Common Insertion Site (CIS) is identified if at least 4 different integrations from independent tumors targeted a genomic region < 100Kb, based on the statistical definition of CIS developed in other studies[20, 21]. The closest gene to the CIS genomic region was considered as CIS gene.

Overall, LAM-PCR products from 30 LVV-induced HCCs were subjected to 454-pyrosequencing and generated 18,702 sequencing reads that, upon mapping on the murine genome, identified a total of 172 unique integration sites.

Comparing the data on the VCN from Supplementary Table 1 with the integration data presented in Supplementary Table 2, it appears that there is not always a perfect match between the VCN and the number of integration mapped from each HCC mass. However, since tumoral margins were defined by gross appearance (and surrounding normal liver parenchyma could have been collected as contaminant together with the HCC mass, especially when the HCC diameter measures < 5 mm) and tumoral masses may contain several stromal cells, the VCN that we measured from the tumoral mass may be an inaccurate estimate of the actual VCN of tumoral cells. Additionally, LAM-PCR and mapping have intrinsic limitations that reduce their efficiency and can also identify integrations from transduced non-tumor cells that contaminate the tumoral mass. Therefore, we do not expect to find a perfect correlation between the VCN and the number of univocally mapped integrations for each single tumoral mass. Nonetheless, considering the whole collection of LVV-induced HCCs, we could find a significant correlation ($P < 0.0001$, R squared = 0.5039, by Pearson correlation) between the VCN and the number of univocally mapped integrations retrieved from each HCC mass.

RNA isolation and transcriptome analyses of murine samples

We isolated total RNA from tumor masses and normal livers with the miRNeasy Mini Kit (Qiagen). For the analysis of LV.ET.LTR generated aberrant transcript, we performed RT-PCR reactions. We performed cDNA preparation using Mo-MLV reverse transcriptase and random hexamers primers (Invitrogen, Superscript III First-Strand Synthesis System for RT-PCR). We performed PCR amplification using a sense primer on the vector transcript annealing few nucleotides downstream LTR transcription start site (LV.LTR_S), and antisense primers annealing on the exon downstream vector integrations in the targeted genes (Supplementary Table 5). PCR products were purified (Qiagen), cloned into the TOPO TA vector (Invitrogen) and sequenced (Primm). We detected the generation of LVV-driven chimeric transcript from all the representative HCCs bearing integration at CIS that we analyzed by RT-PCR. We also detected the LVV-*Braf* fusion transcript from an additional HCC, even if integration studies failed to retrieve the LVV integration within *Braf*.

For whole transcriptome studies, we performed microarray analysis (data deposited in Gene Expression Omnibus Repository, GSE31409). We used total RNA (100ng) for GeneChip analysis. We carried out the preparation of terminal-labeled cDNA, hybridization to the whole-transcript "The GeneChip® Mouse Gene 1.0 ST Array (Affymetrix)" and scanning of the arrays according to manufacturer's protocols (https://www.affymetrix.com). We performed a single microarray from each sample, since it is the widely accepted standard when using Affymetrix commercial microarrays that contain internal quality controls. Each independent tumor that arose upon integration in a specific CIS represents a biological replicate. Therefore, considering the HCCs from which RNA was available for the gene expression analyses, we could analyze 7 HCCs with integration in *Fign*, 3 HCCs with integration in *Braf*, 2 HCCs with integration in *Sos1* and 8 HCCs with integration in *Rtl1*.

Raw microarray data are preprocessed with RMA algorithm. When fold change expression in indicated, it refers to the average ± standard deviation for each group. We performed clustering analysis with unsupervised hierarchical methods with different distance (correlation and Euclidean) and linkage (average and centroid) by dCHIP software (http://www.biostat.harvard.edu/~cli/dchip_2010_01.exe); since we obtained overlapping results with different methods, a representative clusterization heatmap is showed.

We validated the gene expression data obtained by microarray analysis by Taqman RT-Q-PCR on representative samples and genes. We performed cDNA preparation using Mo-MLV reverse transcriptase and random hexamers primers (Invitrogen, Superscript III First-Strand Synthesis System for RT-PCR). We used cDNA as template for TaqMan® Gene Expression Assays specific for each gene (Applied Biosystems). Primers and probes for the detection of *Braf*, *Sos1* and *Fign* anneal at the 3′ portion of the gene, thus allowing the detection of both full-length and truncated transcripts. The Taqman gene expression assays that we used are: Mm01165837-m1 (for the detection of *Braf*); Mm02392620-s1 (for the detection of *Rtl1*); Mm00436730-m1 (for the detection of *Sos1*); Mm03048240-m1 (for the detection of *Sfrs4*) (Applied Biosystems). We performed amplification reactions on a 7900HT Real-time PCR thermal cycler. We calculated the relative expression level of each gene by the $^{\Delta\Delta}$Ct method[56], normalized to *Sfrs4* expression (housekeeping control gene), and represented it as fold change relative to the average of normal liver samples (calibrator). We used Real-time PCR Miner software (http://www.miner.ewindup.info)[57] to calculate the mean PCR amplification efficiency for each gene. We used the qBase software program (http://www.biogazelle.com) to measure the relative expression level for each gene[58]. In order to detect both full length and 5′-truncated transcripts generated by LV.ET.LTR integrations, we designed primers and probes at the 3′ of *Braf* and *Sos1* genes.

We performed analysis of the microarray data by Gene Set Enrichment Analysis (GSEA, http://www.broadinstitute.org/gsea/index.jsp)[59], comparing each murine HCC cohort versus pooled normal livers (from 3 *Cdkn2a*$^{-/-}$*Ifnar1*$^{-/-}$; 2 *Pten* liver-null; 3 wild type mice) or other HCCs with different integrations. GSEA overcomes several limitations of conventional GO-based pathway analysis that often fails to identify deregulated biological processes affecting sets of genes acting in concert. For example, critical pathways such as metabolic and transcriptional programs are characterized by modest increases or reductions in the expression of entire set of genes, more than large expression changes of individual genes belonging to the pathway. Such low-level gene deregulations are often highly relevant from a biological standpoint. Therefore, from an experimental perspective, GSEA presents two main methodological advantages. First, GSEA considers all of the genes in an experiment, not only those above an arbitrary cutoff in terms of fold-change or significance. Second, GSEA assesses the significance by permuting the class labels, which preserves gene-gene correlations providing a more accurate null model and attaching a meaningful statistical value to the results. Using GSEA, we could identify classes of genes specifically or commonly deregulated among different HCC groups. *Sos1* group could not be analyzed by GSEA because composed by only 2 samples (no permutation allowed). See also Supplementary Note 2.

For single probe analysis at CIS genes, we obtained probe-level intensities from The GeneChip® Mouse Gene 1.0 ST Array (Affymetrix) by performing the first two steps of RMA pipeline (RMA background correction and quantile normalization), thus excluding the summarization step. We calculated the fold changes in Figure 3c as ratio between the average expression levels in a specific HCC group Vs the average expression levels in pooled non-tumor livers for each single probe.

Western Blot

In order to test the efficiency of truncated proteins expression of the SINLV used for the validation experiment, we transduced HepG2 cells (human hepatocytic cell line) with SINLV.ET.trSOS1 at Multiplicity of Infection 10 and analyzed them two weeks after transduction. Western blot were performed on Hepg2 cells transduced with LV.ET.LTR (negative control) or SINLV.ET.trSOS1. We also analyzed representative HCCs induced in the cancer genes screening with LV.ET.LTR and HCCs induced by SINLVs in the validation experiment (see Supplementary Fig. 2c-d and Supplementary Table 1 and 3 for mouse and tumor IDs). We extracted total cellular proteins from cells and HCCs with RIPA buffer (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 5 mM EDTA, 1% Triton X-100) supplemented with proteases inhibitors cocktail (Sigma). We homogenized the samples in the lysis solution and incubated them at 4°C for 30 min. Cell lysates were cleared by centrifugation at 10,000 × g for 10 min at 4°C, and the supernatants were collected and assayed for protein concentration using Quick start Bradford dye reagent (BioRad). 30-60 micrograms of proteins were run on SDS-PAGE under reducing conditions. For immunoblotting, we transferred proteins to nylon membranes by iBlot dry blotting system (Invitrogen); the membranes were then blocked in 5% non-fat dry milk in a solution of TBS 1X Tween 0.1% and incubated with the specific primary antibody followed by peroxidase-conjugated secondary antibodies (anti-mouse or anti-rabbit IgG HRP-conjugated for secondary detection: #715-035-150 and #711-035-152, Jackson ImmunoResearch; dilution: 1/10,000). The signal was detected with horseradish peroxidase (HRP) chemiluminescent substrate (SuperSignal West Dura Chemiluminescent Substrate, Thermo scientific or ECL prime, Amersham) and exposure to autoradiography films (Amersham HyperfilmTM, GE Healthcare). We used the following primary antibodies : rabbit polyclonal anti-SOS1 (LS-C10294, LifeSpan BioSciences; 1:1,000) which is raised against an epitope at the C-terminus of the protein (corresponding to amino acids 1243-1258) and binds both human

and murine SOS1; mouse anti-Tubulin (T9026, Sigma; dilution 1/50,000) that binds both human and murine beta tubulin; and mouse anti-GAPDH (G9545, Sigma, 1:10,000) that binds efficiently human GAPDH and less efficiently murine GAPDH.

Gene expression on human samples

We utilized archival human samples following the rules of the Ethical Committee of Hospital of Saint Rafael (HSR). Each human specimen was evaluated in blinded fashion and independently by two pathologists (Francesca Sanvito, Claudio Doglioni) with expertise in human and mouse histopathology. The liver tumors were classified according to World Health Organization Classification of Tumors[48] and were graded according to the modified Edmonson-Steiner grading system[49]. The HCC collection harvested in our institution (HSR collection) is formed by 14 liver tumors (1 adenoma; 1, 9 and 3 HCCs of G1, G2 and G3 respectively) and 14 patient-matched normal liver samples.

We isolated total RNA from human HCC and patient-matched normal liver with the miRNeasy Mini Kit (Qiagen). We performed cDNA preparation using Mo-MLV reverse transcriptase and random hexamers primers (Invitrogen, Superscript III First-Strand Synthesis System for RT-PCR). We used cDNA as template for TaqMan® Gene Expression Assays specific for each gene (Applied Biosystems). Primers and probes for the detection of *BRAF*, *SOS1* and *FIGN* detect the 3′ portion of the gene. The Taqman gene expression assays that we used are: Hs00269944-m1 (for the detection of *BRAF*); Hs00250679-s1 (for the detection of *FIGN*); Hs00893134_m1 (for the detection of *SOS1*); Hs00194538-m1 (for the detection of *SFRS4*) (Applied Biosystems).

We performed amplification reactions on a 7900HT Real-time PCR thermal cycler. We calculated the relative expression level of each gene by the $^{\Delta\Delta}$Ct method[56], normalized to *SFRS4* expression (housekeeping control gene), and represented it as fold change relative to the normal liver (calibrator). We used Real-time PCR Miner software (http://www.miner.ewindup.info)[57] to calculate the mean PCR amplification efficiency for each gene. We used the qBase software program (http://www.biogazelle.com) to measure the relative expression level for each gene[58].

Data-mining of gene expression and CGH data

We analyzed a dataset of 75 human samples (Mount Sinai School of Medicine – MSSM collection; GEO GSE6764[25]) including normal liver (n=10), cirrhosis ($n = 13$), low grade dysplasia ($n = 10$), high grade dysplasia ($n = 7$), very early HCC ($n = 8$), early HCC ($n = 10$), advanced HCC ($n = 7$) and very advanced HCC ($n = 10$), for a total of 40 non-tumoral samples and 35 HCCs. These samples were referred in the manuscript as MSSM sample colletion. These patients were mainly affected by hepatitis C virus infection[25].

By SAM (Significance Analysis of Microarray) approach, we analyzed our murine microarray data described above and identified gene expression signatures specifically associated to the different cancer genes (*Braf*, *Fign* and *Rtl1*) with a False Discovery Rate < 0.01. By Gene Set Enrichment Analysis (GSEA), we compared and contrasted the probesets deregulated in the different disease groups in the human samples with the murine HCC subgroups. By this approach, we evaluated how the genes specifically deregulated in murine CIS-specific groups of HCC are deregulated in different classes of human samples in this dataset[25]. As for the single gene analysis, we calculated the fold change in Figure 5a-c versus the average of normal liver samples. See also Supplementary Note 3.

We performed clustering analysis with unsupervised hierarchical methods with different distance (correlation and Euclidean) and linkage (average and centroid) by dCHIP software

Europe PMC Funders Author Manuscripts

(www.dchip.org); since similar results were obtained with different methods, a representative clusterization is showed (correlation distance and centroid linkage).

We retrieved additional gene expression or CGH data on murine and human samples by data-mining using OncoDB.HCC database (http://oncodb.hcc.ibms.sinica.edu.tw, see also Supplementary Note 4), Oncomine web resource (www.oncomine.org) and Gene Expression Omnibus web resource (http://www.ncbi.nlm.nih.gov/geo/).

Correlation between gene expression and clinical outcome of patients affected by HCCs

We could access to the clinical data of 70 patients affected by hepatitis B virus-induced HCCs[27]. The gene expression data were publicly available by Gene Expression Omnibus web resource (GEO GSE15765[27]). Since Affymetrix Human Genome U133A 2.0 Array was used, only 14,500 well-characterized human genes were detected; unfortunately, some genes of interest were lacking on the array, such as *FIGN* and *RTL1*. We divided patients in two groups according to the expression level of *SOS1*, *BRAF* (above or below the average expression level of each gene in the entire collection of HCCs) and we compared the survival curve and disease-free survival curve between the two groups by GraphPad Prism applying both Log-rank (Mantel-Cox) Test and Gehan-Breslow-Wilcoxon Test. For instance, as for *SOS1*, we divided patients in two groups: the ones carrying HCCs that displayed *SOS1* expression levels above *SOS1* average expression in the HCC collection, and the ones carrying HCCs that displayed *SOS1* expression levels below average *SOS1* expression, n=40 and n=30 respectively. Analysis of survival curves showed that patients bearing HCC with high expression of *SOS1* have a significantly increased overall survival compared to patients bearing HCC with low *SOS1* levels (by both Gehan-Breslow-Wilcoxon Test and Log-rank Mantel-Cox Test).

We used the cancer gene-specific gene expression signatures identified in the LVV-induced HCCs in mice by SAM (FDR < 0.01) to perform hierarchical clustering of the 70 HCCs from the microarray dataset. We performed clustering analysis with unsupervised hierarchical methods with different distance (correlation and Euclidean) and linkage (average and centroid) by dCHIP software (www.dchip.org); since similar results were obtained with different methods, a representative clusterization was considered and showed (correlation distance and centroid linkage). By this approach, we generated a hierarchical clustering tree and identified two main sample clusters. Then we compared the survival curves and disease-free survival curves between the two clusters by GraphPad Prism applying both Log-rank (Mantel-Cox) Test and Gehan-Breslow-Wilcoxon Test. We also analyzed the main clusters of genes that were differentially expressed among the sample clusters by Ingenuity Pathway Analysis software (Ingenuity Systems, www.ingenuity.com) to identify the main Biological Functions associated to these genes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to M. Rocchi, A. Vino, S. Oldoni and M. Marini for technical help, G. Santambrogio for collaboration in human HCC collection at HSR, M. Volpin for help in animal handling, S. Annunziato and J. Sgualdino for help in molecular biology, M. A. Venneri and D. Biziato for suggestion regarding immunoflourescent staining, V. Neguembor, D. Cabianca and V. Casà for the suggestions regarding western blot, M. De Palma and D. Gabellini for critical reading of this manuscript.

We would like to acknowledge the PhD program in in Cellular and Molecular Biology, since Marco Ranzani conducted this study as partial fulfillment of his PhD in Molecular Medicine, Program in Cellular and Molecular Biology, San Raffaele University, Milan, Italy.

This work was supported by grants from the Association for International Cancer Research (AICR 09-0784 to E. Montini), Telethon Foundation (TGT11D1 to E. M.) European Union (Clinigene NoE LSHB-CT-2006-018933 to E. Montini), the Italian Ministries of Health (GR - 2007 - 684057 to E. Montini), European Union (PERSIST to L. Naldini), the Italian Ministries of Health (ONC-34/07 to L. Naldini), BRL (Basic Research Laboratory) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2011-0001564 to Y. J. Kim).

REFERENCES

1. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458:719–724. [PubMed: 19360079]
2. Kool J, Berns A. High-throughput insertional mutagenesis screens in mice to identify oncogenic networks. Nat Rev Cancer. 2009; 9:389–399. [PubMed: 19461666]
3. Montini E, et al. The genotoxic potential of retroviral vectors is strongly modulated by vector design and integration site selection in a mouse model of HSC gene therapy. J Clin Invest. 2009; 119:964–975. [PubMed: 19307726]
4. Naldini L, Blomer U, Gage FH, Trono D, Verma IM. Efficient transfer, integration, and sustained long-term expression of the transgene in adult rat brains injected with a lentiviral vector. Proc Natl Acad Sci U S A. 1996; 93:11382–11388. [PubMed: 8876144]
5. Croci C, et al. Cerebellar neurons and glial cells are transducible by lentiviral vectors without decrease of cerebellar functions. Dev Neurosci. 2006; 28:216–221. [PubMed: 16679768]
6. Dolcetta D, et al. Design and optimization of lentiviral vectors for transfer of GALC expression in Twitcher brain. J Gene Med. 2006; 8:962–971. [PubMed: 16732552]
7. Cefai D, et al. Multiply attenuated, self-inactivating lentiviral vectors efficiently transduce human coronary artery cells in vitro and rat arteries in vivo. J Mol Cell Cardiol. 2005; 38:333–344. [PubMed: 15698840]
8. Bonci D, et al. 'Advanced' generation lentiviruses as efficient vectors for cardiomyocyte gene transduction in vitro and in vivo. Gene Ther. 2003; 10:630–636. [PubMed: 12692591]
9. Buckley SM, et al. Lentiviral transduction of the murine lung provides efficient pseudotype and developmental stage-dependent cell-specific transgene expression. Gene Ther. 2008; 15:1167–1175. [PubMed: 18432275]
10. Di Nunzio F, et al. Correction of laminin-5 deficiency in human epidermal stem cells by transcriptionally targeted lentiviral vectors. Mol Ther. 2008; 16:1977–1985. [PubMed: 18813277]
11. Tedesco FS, et al. Transplantation of Genetically Corrected Human iPSC-Derived Progenitors in Mice with Limb-Girdle Muscular Dystrophy. Sci Transl Med. 2012; 4:140ra189. [PubMed: 22745439]
12. Follenzi A, Sabatino G, Lombardo A, Boccaccio C, Naldini L. Efficient gene delivery and targeted expression to hepatocytes in vivo by improved lentiviral vectors. Hum Gene Ther. 2002; 13:243–260. [PubMed: 11812281]
13. Serrano M, et al. Role of the INK4a locus in tumor suppression and cell mortality. Cell. 1996; 85:27–37. [PubMed: 8620534]
14. Brown BD, et al. In vivo administration of lentiviral vectors triggers a type I interferon response that restricts hepatocyte gene transfer and promotes vector clearance. Blood. 2007; 109:2797–2805. [PubMed: 17170119]
15. Tannapfel A, et al. INK4a-ARF alterations and p53 mutations in hepatocellular carcinomas. Oncogene. 2001; 20:7104–7109. [PubMed: 11704835]
16. Farazi PA, DePinho RA. Hepatocellular carcinoma pathogenesis: from genes to environment. Nat Rev Cancer. 2006; 6:674–687. [PubMed: 16929323]
17. Horie Y, et al. Hepatocyte-specific Pten deficiency results in steatohepatitis and hepatocellular carcinomas. J Clin Invest. 2004; 113:1774–1783. [PubMed: 15199412]
18. Hu TH, et al. Expression and prognostic role of tumor suppressor gene PTEN/MMAC1/TEP1 in hepatocellular carcinoma. Cancer. 2003; 97:1929–1940. [PubMed: 12673720]
19. Vigna E, et al. Efficient Tet-dependent expression of human factor IX in vivo by a new self-regulating lentiviral vector. Mol Ther. 2005; 11:763–775. [PubMed: 15851015]

20. Abel U, et al. Real-time definition of non-randomness in the distribution of genomic events. PLos One. 2007; 2:e570. [PubMed: 17593969]
21. Wu X, Luke BT, Burgess SM. Redefining the common insertion site. Virology. 2006; 344:292–295. [PubMed: 16271739]
22. Collier LS, Carlson CM, Ravimohan S, Dupuy AJ, Largaespada DA. Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. Nature. 2005; 436:272–276. [PubMed: 16015333]
23. Gureasko J, et al. Role of the histone domain in the autoinhibition and activation of the Ras activator Son of Sevenless. Proc Natl Acad Sci U S A. 2010; 107:3430–3435. [PubMed: 20133692]
24. Brown BD, et al. A microRNA-regulated lentiviral vector mediates stable correction of hemophilia B mice. Blood. 2007; 110:4144–4152. [PubMed: 17726165]
25. Wurmbach E, et al. Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. Hepatology. 2007; 45:938–947. [PubMed: 17393520]
26. Chen X, et al. Gene expression patterns in human liver cancers. Mol Biol Cell. 2002; 13:1929–1939. [PubMed: 12058060]
27. Woo HG, et al. Identification of a cholangiocarcinoma-like gene expression trait in hepatocellular carcinoma. Cancer Res. 2010; 70:3034–3041. [PubMed: 20395200]
28. Uren AG, et al. Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks. Cell. 2008; 133:727–741. [PubMed: 18485879]
29. Starr TK, et al. A Transposon-Based Genetic Screen in Mice Identifies Genes Altered in Colorectal Cancer. Science. 2009
30. Rad R, et al. PiggyBac transposon mutagenesis: a tool for cancer gene discovery in mice. Science. 2010; 330:1104–1107. [PubMed: 20947725]
31. Keng VW, et al. A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma. Nature biotechnology. 2009; 27:264–274.
32. Colombino M, et al. BRAF and PIK3CA genes are somatically mutated in hepatocellular carcinoma among patients from South Italy. Cell Death Dis. 2012; 3:e259. [PubMed: 22258409]
33. Ciampi R, et al. Oncogenic AKAP9-BRAF fusion is a novel mechanism of MAPK pathway activation in thyroid cancer. J Clin Invest. 2005; 115:94–101. [PubMed: 15630448]
34. Palanisamy N, et al. Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat Med. 2010; 16:793–798. [PubMed: 20526349]
35. Guichard C, et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet. 2012; 44:694–698. [PubMed: 22561517]
36. Schneider A, Mehmood T, Pannetier S, Hanauer A. Altered ERK/MAPK signaling in the hippocampus of the mrsk2_KO mouse model of Coffin-Lowry syndrome. J Neurochem. 2011; 119:447–459. [PubMed: 21838783]
37. Douville E, Downward J. EGF induced SOS phosphorylation in PC12 cells involves P90 RSK-2. Oncogene. 1997; 15:373–383. [PubMed: 9242373]
38. Llovet JM, et al. Sorafenib in advanced hepatocellular carcinoma. N Engl J Med. 2008; 359:378–390. [PubMed: 18650514]
39. Donsante A, et al. AAV vector integration sites in mouse hepatocellular carcinoma. Science. 2007; 317:477. [PubMed: 17656716]
40. Dupuy AJ, et al. A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice. Cancer research. 2009; 69:8150–8156. [PubMed: 19808965]
41. Montini E, et al. Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. Nature biotechnology. 2006; 24:687–696.
42. Themis M, et al. Oncogenesis following delivery of a nonprimate lentiviral gene therapy vector to fetal and neonatal mice. Mol Ther. 2005; 12:763–771. [PubMed: 16084128]
43. Follenzi A, Ailles LE, Bakovic S, Geuna M, Naldini L. Gene transfer by lentiviral vectors is limited by nuclear translocation and rescued by HIV-1 pol sequences. Nat Genet. 2000; 25:217–222. [PubMed: 10835641]

44. Kim WY, Sharpless NE. The regulation of INK4/ARF in cancer and aging. Cell. 2006; 127:265–275. [PubMed: 17055429]
45. Muller U, et al. Functional role of type I and type II interferons in antiviral defense. Science (New York, N.Y. 1994; 264:1918–1921. [PubMed: 8009221]
46. Clawson GA. Mechanisms of carbon tetrachloride hepatotoxicity. Pathol Immunopathol Res. 1989; 8:104–112. [PubMed: 2662164]
47. Weber LW, Boll M, Stampfl A. Hepatotoxicity and mechanism of action of haloalkanes: carbon tetrachloride as a toxicological model. Crit Rev Toxicol. 2003; 33:105–136. [PubMed: 12708612]
48. Bosman, FTC,F.; Hruban, RH.; Theise, ND., editors. WHO Classification of Tumors of the Digestive System. 4th edition. International Agency of Research on Cancer; Lyon: 2010.
49. Ishak, KG.; Goodman, ZD.; Stocker, JT. Atlas of Tumor Pathology: Tumors of the Liver and Intrahepatic Bile Ducts. Edn. 2001. Armed Forces Institute of Pathology; Washington, DC (USA): 2001.
50. Hadjantonakis AK, Gertsenstein M, Ikawa M, Okabe M, Nagy A. Generating green fluorescent mice by germline transmission of green fluorescent ES cells. Mechanisms of development. 1998; 76:79–90. [PubMed: 9867352]
51. Schmidt M, et al. A model for the detection of clonality in marked hematopoietic stem cells. Annals of the New York Academy of Sciences. 2001; 938:146–155. discussion 155-146. [PubMed: 11458502]
52. Ott MG, et al. Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. Nat Med. 2006; 12:401–409. [PubMed: 16582916]
53. Schmidt M, et al. Clonal evidence for the transduction of CD34+ cells with lymphomyeloid differentiation potential and self-renewal capacity in the SCID-X1 gene therapy trial. Blood. 2005; 105:2699–2706. [PubMed: 15585650]
54. Schmidt M, et al. Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. Hum Gene Ther. 2001; 12:743–749. [PubMed: 11339891]
55. Schmidt M, et al. High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nat Methods. 2007; 4:1051–1057. [PubMed: 18049469]
56. Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic acids research. 2001; 29:e45. [PubMed: 11328886]
57. Zhao S, Fernald RD. Comprehensive algorithm for quantitative real-time polymerase chain reaction. J Comput Biol. 2005; 12:1047–1064. [PubMed: 16241897]
58. Hellemans J, Mortier G, De Paepe A, Speleman F, Vandesompele J. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome biology. 2007; 8:R19. [PubMed: 17291332]
59. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102:15545–15550. [PubMed: 16199517]

Europe PMC Funders Author Manuscripts

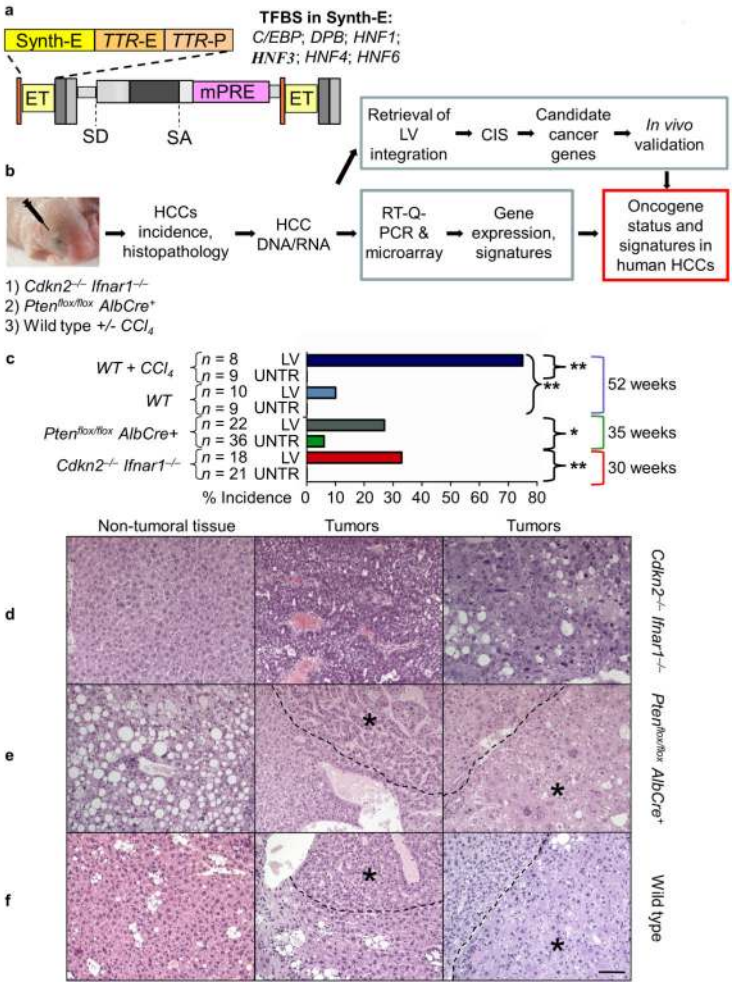
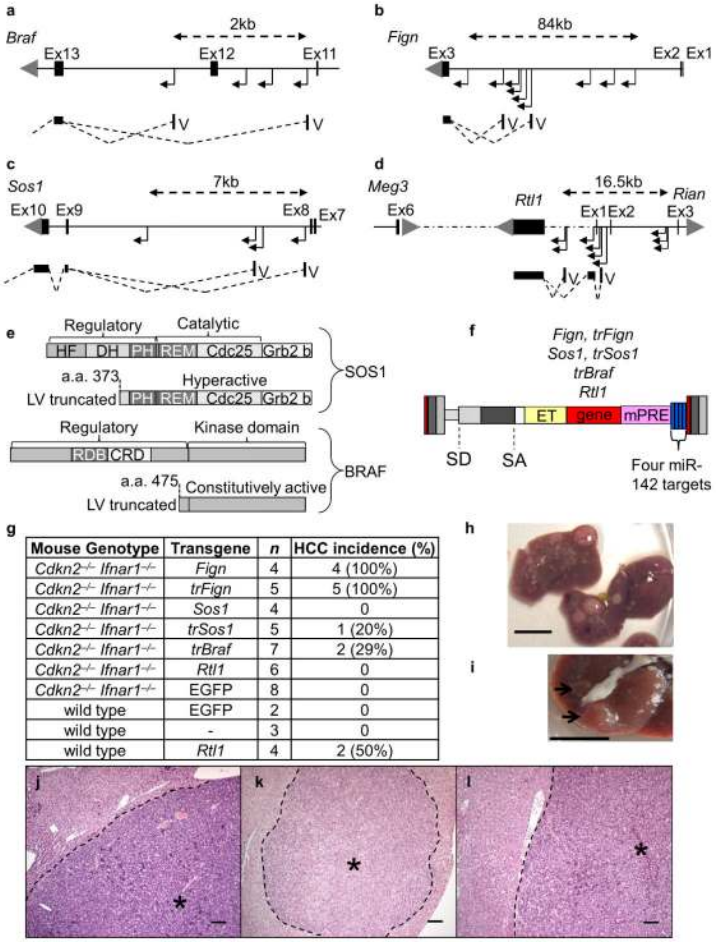


Figure 1. Lentiviral vector-mediated induction of HCC

(**a**) Schematic of LV.ET.LTR vector (see also Supplementary Note 1). The Enhanced Transthyretin enhancer-promoter sequence (ET)[19] was cloned in the Long Terminal Repeat (LTR). ETr contains a synthetic enhancer (Synth-E) bearing transcription factor binding sites (TFBS) highly-active in hepatocytes (indicated), the transthyretin enhancer (TTR-E) and transthyretin promoter (TTR-P); SD = splice donor site; SA = splice acceptor site; mPRE = Woodchuck Post Transcriptional Regulatory Element, mutated sequence. (**b**) Experimental outline for LVV-mediated insertional mutagenesis. (**c**) Liver tumor incidence (%) in different experimental groups (P-values by two tailed Fisher's exact test). *n* = number of mice; LVV indicates the LV.ET.LTR-transduced group, UNTR = the non-transduced age-matched control groups. Mice were euthanized if sick or at the final time point indicated on the right. (**d-f**) Hematoxylin and eosin stained sections of livers and HCC masses from: (**d**) *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mouse model, (**e**) *Pten*$^{flox/flox}$ *AlbCre*$^{+}$ mouse model, and (**f**) Wild type mouse model (left and middle panel WT with CCl_4, right panel WT without CCl_4).Left panels show tumor-free liver parenchyma from a 30 weeks (d), 35 weeks (e) and one year old WT mice treated with CCl_4 (f). Middle and right panels show tumor masses of different grades. (*) indicates the tumor area. Otherwise, only the tumor tissue is shown. Scale bar = 100 μm.

Europe PMC Funders Author Manuscripts



Mouse Genotype	Transgene	*n*	HCC incidence (%)
Cdkn2$^{-/-}$ *Ifnar1*$^{-/-}$	*Fign*	4	4 (100%)
Cdkn2$^{-/-}$ *Ifnar1*$^{-/-}$	*trFign*	5	5 (100%)
Cdkn2$^{-/-}$ *Ifnar1*$^{-/-}$	*Sos1*	4	0
Cdkn2$^{-/-}$ *Ifnar1*$^{-/-}$	*trSos1*	5	1 (20%)
Cdkn2$^{-/-}$ *Ifnar1*$^{-/-}$	*trBraf*	7	2 (29%)
Cdkn2$^{-/-}$ *Ifnar1*$^{-/-}$	*Rtl1*	6	0
Cdkn2$^{-/-}$ *Ifnar1*$^{-/-}$	EGFP	8	0
wild type	EGFP	2	0
wild type	-	3	0
wild type	*Rtl1*	4	2 (50%)

Figure 2. Identification and validation of liver cancer genes

LVV integrations in HCCs at CIS targeted different genes: (**a**) *Braf* (**b**) *Fign*, (**c**) *Sos1*, (**d**) *Rtl1* within the *Dlk1-Dio3* region. Dashed lines: intergenic chromosomal regions; solid lines and boxes: introns and exons of genes in the region, respectively. Grey triangles: transcript orientation; bended arrows: integration position and vector orientation. Below are represented the aberrant transcripts generated by LVV integration. V: vector-derived exon containing a portion of the LVV LTR and leader sequence up to the LVV splice donor; boxes: genomic exons; dashed lines: splicing events. (**e**) Representative functional domains of BRAF and SOS1 proteins and schematic of truncated proteins generated by LVV integration. Aminoacid number at the predicted truncation is indicated. (**f**) Vector design for the liver gene transfer of candidate cancer genes, based on LVV with SIN LTRs. (**g**) Liver tumor incidence in different experimental groups administered with SINLVs that express candidate cancer genes. Tr = truncated ORF; *n* = number of mice. (**h, i**) Liver of a 64-days-old *Cdkn2*$^{-/-}$ *Ifnar1*$^{-/-}$ mouse expressing truncated *Fign* (**h),** and of a 349-days-old wild type mouse expressing *Rtl1* and treated with CCl_4 (i).Arrows indicate 2 HCCs. Scale bar = 1 cm. (**j-l**) Representative hematoxylin and eosin stained sections of HCC from mice treated with the SINLV overexpressing full length *Fign* (**j**), truncated *Fign* (**k**) or truncated *Sos1* (**l**). * indicates tumor area. Scale bar= 100 μm. See also Supplementary Figure 2 and Supplementary Tables 2-3.

Europe PMC Funders Author Manuscripts

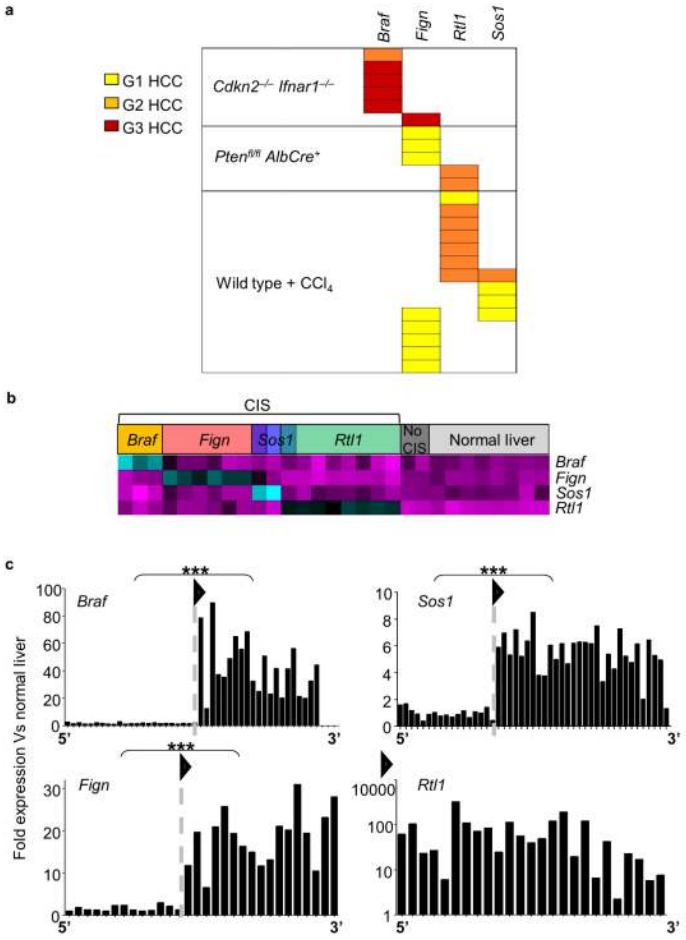


Figure 3. LVV integrations at CIS upregulate the targeted genes

(**a**) Each tumor bearing integrations targeting a CIS is represented as a square with color according to the grade.(**b**) The heatmap shows expression levels of LVV induced HCCs and non-tumor liver from experimental mice. Magenta indicates low expression and turquoise indicates high expression. (**c**) Plotted is the fold change of expression (versus expression in normal livers) of HCCs bearing the indicated CIS integration. Each plot shows every probe of the microarray probeset for the CIS gene (from 5′ to 3′ of each transcript). The black triangle and dashed grey line indicate LVV integration landing inside the transcript (*Braf*, *Sos1* and *Fign*).

Europe PMC Funders Author Manuscripts

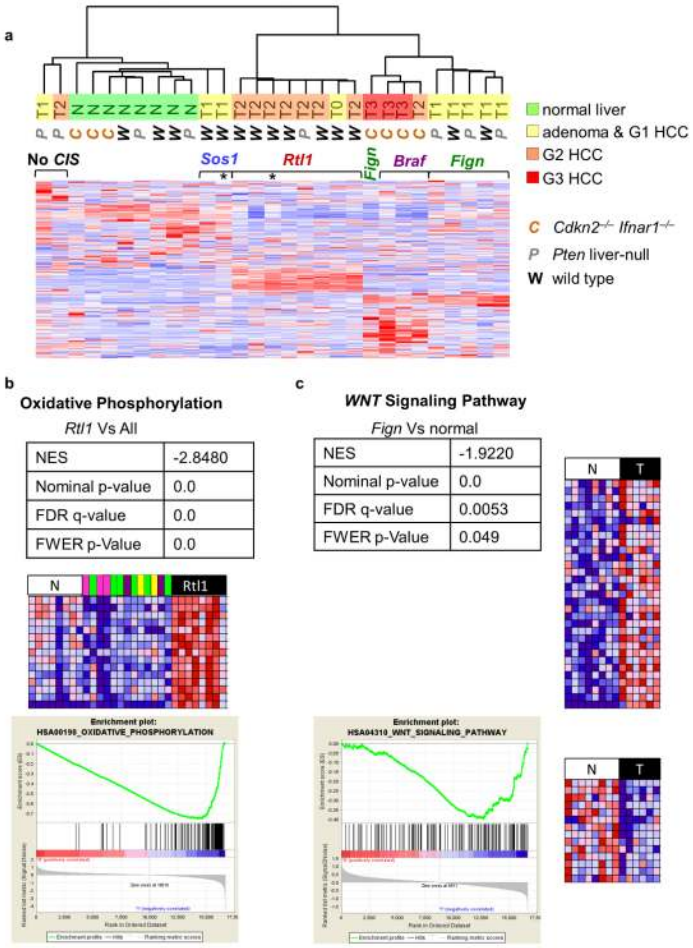


Figure 4. Transcriptome deregulations in LVV-induced HCCs

(**a**) Heat map and dendrogram showing hierarchical unsupervised clustering analysis of LVV-induced HCCs and nontumoral livers from experimental mice. CIS genes hit in the HCCs are indicated above the heat map. No CISs are HCCs without integrations at CIS genes. Asterisks indicate HCCs with additional integrations at CISs. Expression levels in the heat maps are color coded from blue (low) to red (high). (**b**) Expression profile of *Rtl1* HCCs as compared to pooled HCCs (three with *Braf* integration outlined in pink, six with *Fign* integration in green, two with *Sos1* integration in violet and two without integration at CIS in yellow) and nontumoral livers from the different genetic backgrounds (N, in white) by GSEA. A heat map representation of the 15 most overexpressed genes of oxidative phosphorylation between *Rtl1* HCCs and other samples is shown (from blue, low expression, to red, high expression). GSEA statistics: NES, normalized enrichment score; FDR q value: false discovery rate; FWER P value, family-wise error rate. Bottom, enrichment plot showing the overrepresentation at the top and bottom of the ranked gene set. (**c**) Expression profile of Fign HCCs as compared to eight nontumoral livers. Heat map representations of the most upregulated (top) and downregulated (bottom) genes of the WNT signaling pathway in *Fign* tumors versus expression in nontumoral livers are shown. Bottom, enrichment plot as in b. (See also Supplementary Note 2, Supplementary Fig. 3 and Supplementary Table 4.)

Europe PMC Funders Author Manuscripts

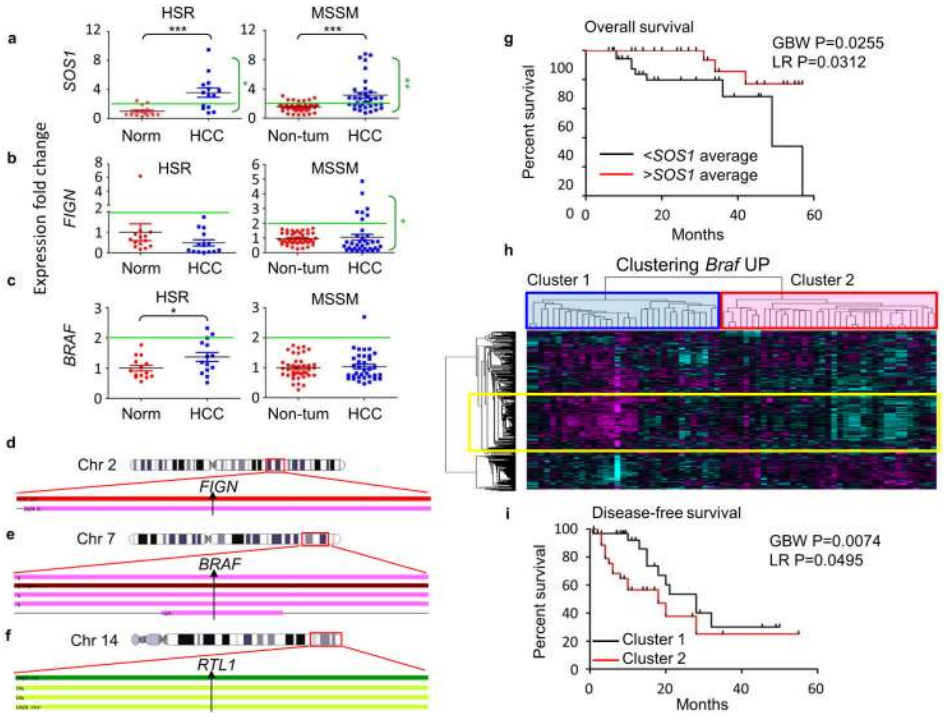


Figure 5. The newly identified liver cancer genes are implicated in human hepatocarcinogenesis

(**a–c**) Expression fold changes for *SOS1* (a), **FIGN** (b) and *BRAF*(c) for nontumoral liver (non-tum) and HCCs from the HSR and MSSM collections. The MSSM collection25 was analyzed by Affymetrix microarray, whereas the HSR collection was analyzed by RT-qPCR. Black lines, mean; colored whiskers, s.d. Black P value by unpaired t-test; green P value by two-tailed Fisher's exact test; *P < 0.05; **P < 0.01; ***P < 0.001 (see also Supplementary Note 3). (**d-f**) Comparative genomic hybridization data of the genomic regions encompassing FIGN, BRAF and RTL1 were obtained by consulting the OncoDB.HCC database (Supplementary Note 4). Graphs show a chromosomal region of ±10 megabase pairs centered at the gene of interest (arrow). Bars: pink, copy-number gain; red, minimal overlap region of copy-number gain; brown, amplification; pale green, copy-number loss; dark green, minimal overlap region of copy-number loss. (**g**) Survival curves for patients with HCC27 with high or low expression of SOS1 (Online Methods). GBW, Gehan-Breslow-Wilcoxon test; LR, log-rank Mantel-Cox test. (**h**) Clustering of human HCCs27 was performed considering the human orthologs of the upregulated (UP) genes from the murine Braf signature (Online Methods). Unsupervised clustering analysis identified two main HCC clusters (blue and red boxes). The yellow box marks genes highly expressed in the cluster with poorer prognosis and that mainly have IPA (Ingenuity Systems pathway analysis software) biological functions of Cell Cycle, DNA Replication and Cancer (Supplementary Table 4a). Magenta, low expression; cyan, high expression. (i) Disease-free survival of the patients with HCC belonging to the clusters identified in h.

Europe PMC Funders Author Manuscripts