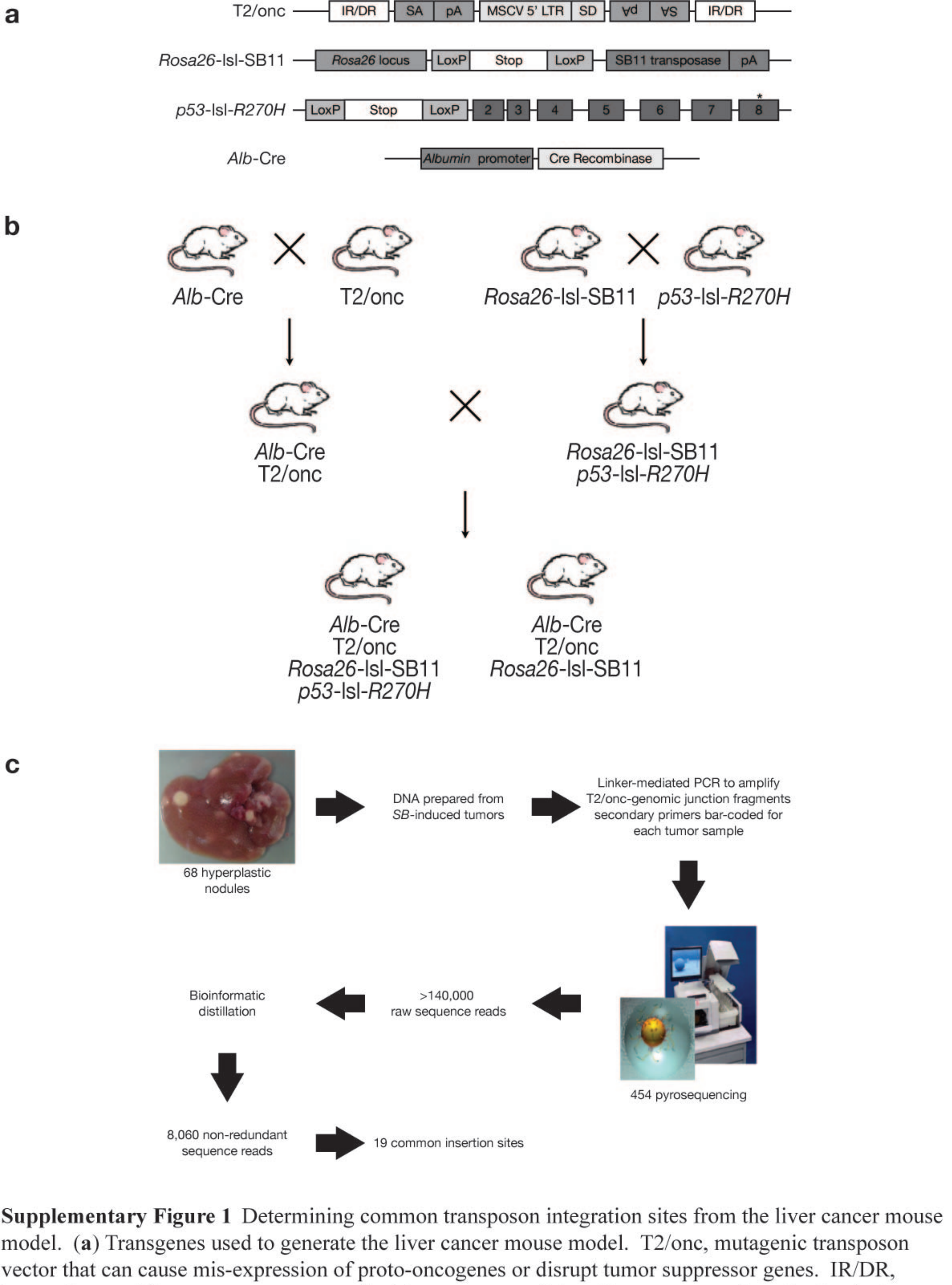
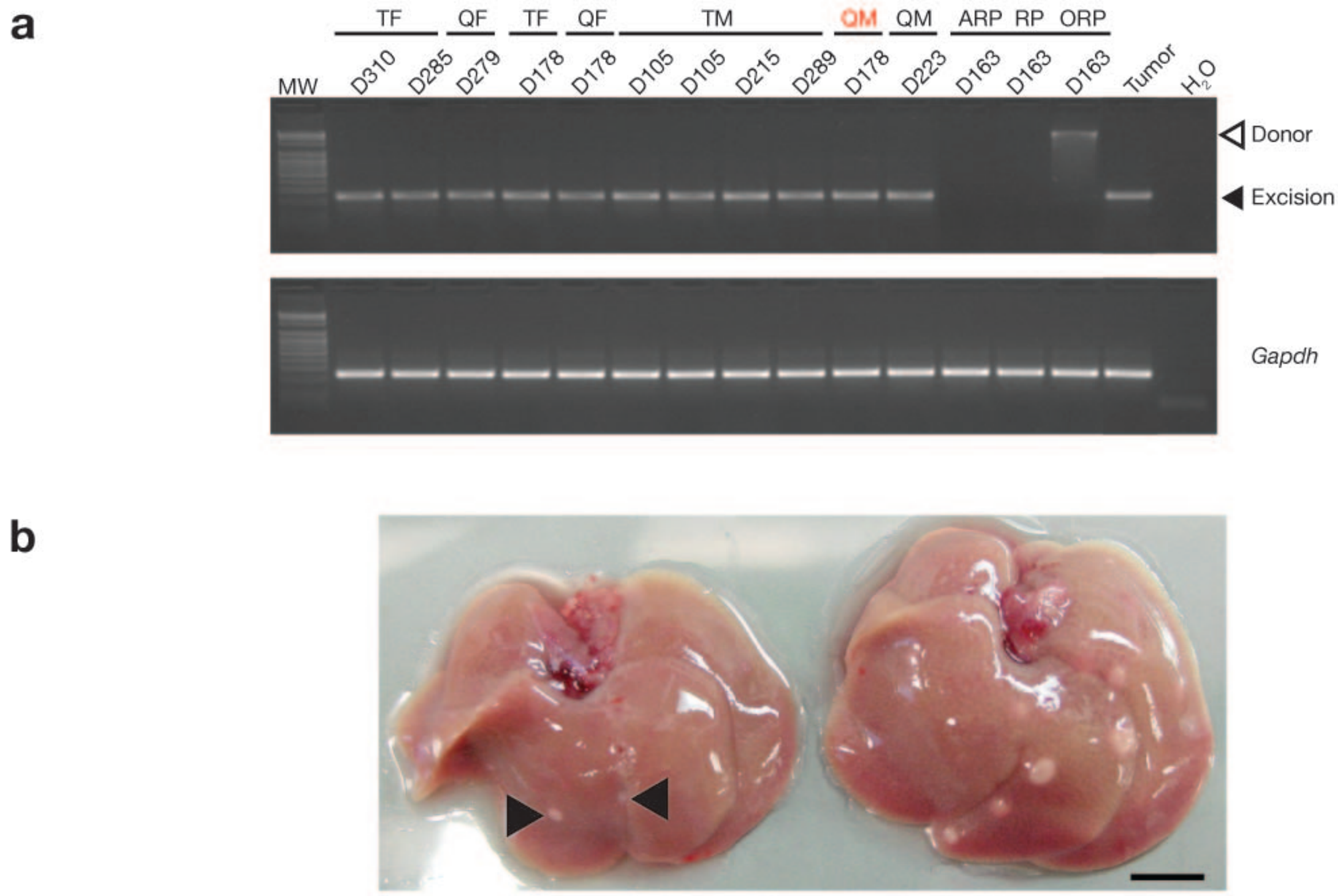
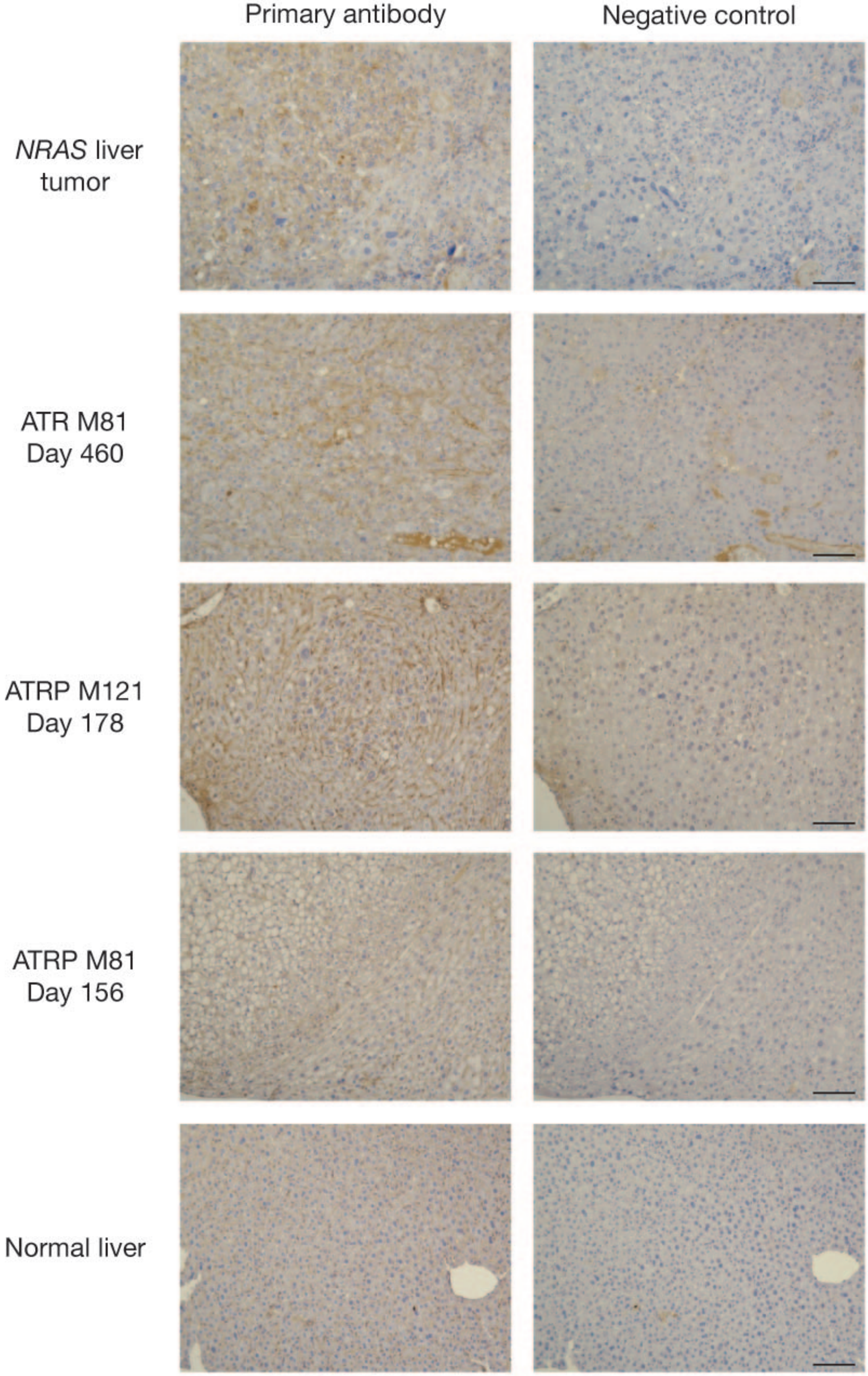


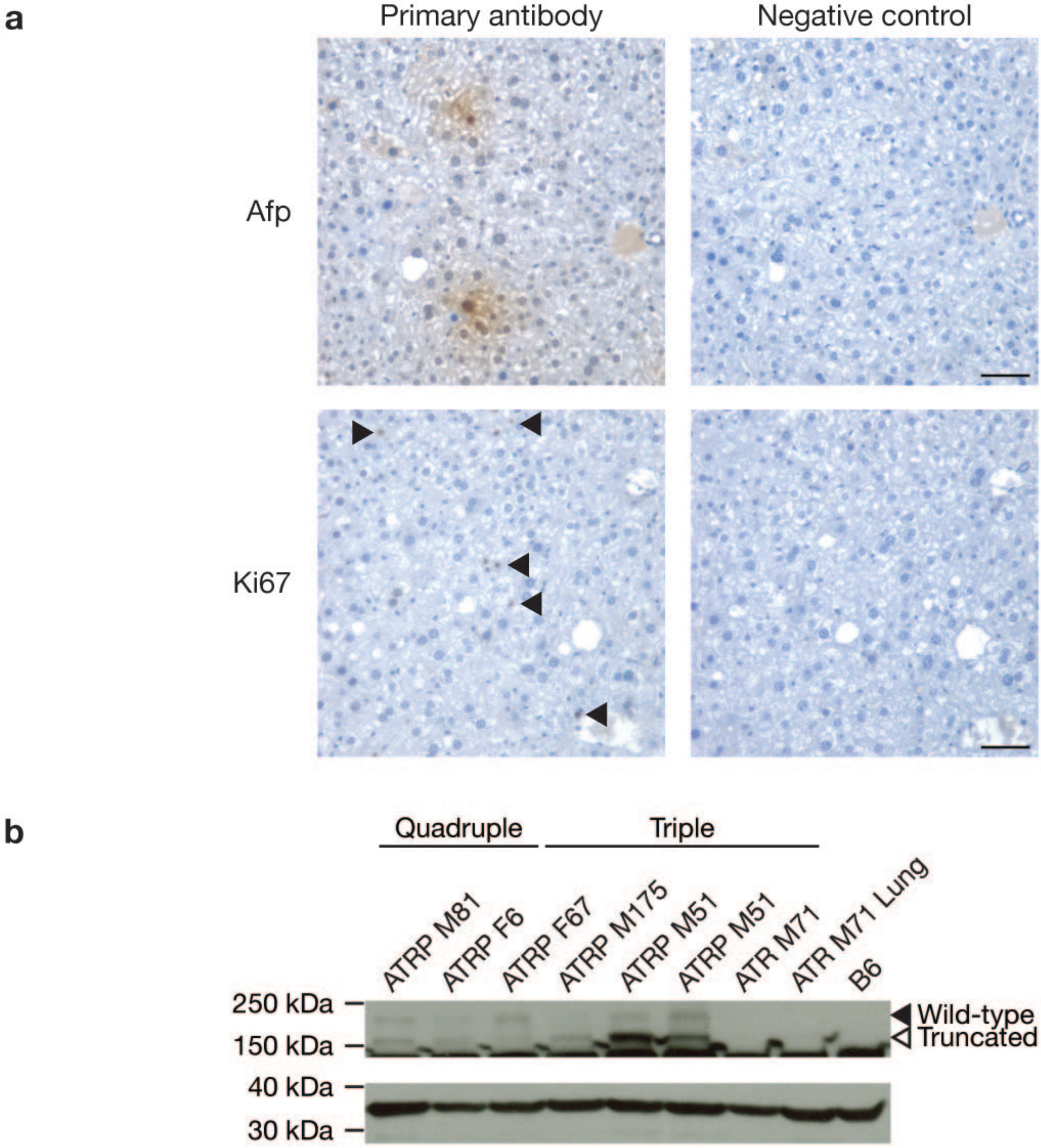
Supplementary Figure 1 Determining common transposon integration sites from the liver cancer mouse model. (**a**) Transgenes used to generate the liver cancer mouse model. T2/onc, mutagenic transposon vector that can cause mis-expression of proto-oncogenes or disrupt tumor suppressor genes. IR/DR, inverted repeat/direct repeat transposon flanking sequences; SA, splice acceptor; pA, polyadenalytion signal; MSCV 5' LTR, 5'-LTR of the murine stem cell virus; SD, splice donor. *Rosa26*-lsl-SB11, *SB* transposase (SB11) carrying a floxed-stop (lsl) cassette engineered into the mouse *Rosa26* locus. *p53*-lsl-*R270H*, conditional dominant negative *p53* transgene (asterisk indicates location of *R270H* mutation). *Alb*-Cre, *Albumin* promoter driving Cre recombinase for the removal of floxed-stop cassettes and activation of both *SB* transposase and dominant negative p53 R270H protein in hepatocytes. (**b**) Breeding strategy for generating experimental animals. Transgenic mice carrying the 4 individual transgenes were bred to obtain doubly transgenic animals. These doubly transgenic mice were subsequently bred to obtain either triple (non-predisposed genetic background) or quadruple (predisposed genetic background) experimental animal. Control animals carrying various combinations of transgenes were also generated but not shown in the diagram. (**c**) Flowchart for high-throughput barcode-assisted amplification procedure used to obtain transposon common insertion sites.



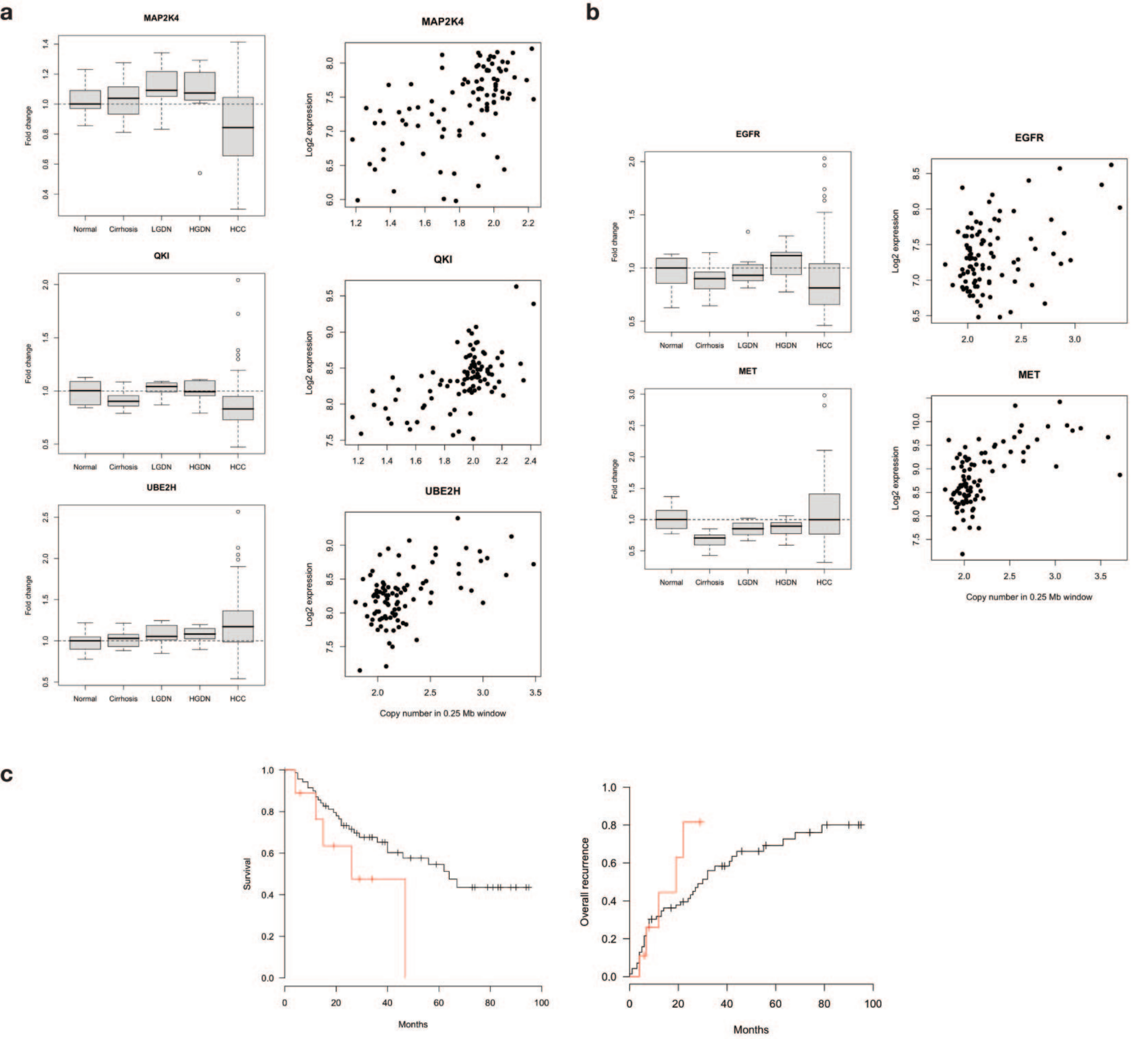
Supplementary Figure 2 Analyses of preneoplastic liver nodules isolated from experimental animals. (**a**) Excision PCR analyses demonstrating evidence of transposon (T2/onc) excision in livers taken from non-tumor producing experimental and control livers. TF, triple-transgenic female livers; QF, quadruple-transgenic female livers; TM, triple-transgenic male livers; QM, quadruple-transgenic male livers (red and black, tumor- and non-tumor producing, respectively); ARP, transgenic animal containing *Alb*-Cre, *Rosa26*-lsl-SB11 and *p53*-lsl-*R270H* transgenes; RP, transgenic animal containing *Rosa26*-lsl-SB11 and *p53*-lsl-*R270H* transgenes; ORP, transgenic animal containing the T2/onc, *Rosa26*-lsl-SB11 and *p53*-lsl-*R270H* transgenes; Tumor, genomic DNA isolated from a liver neoplastic nodule; H_2O, double-distilled water negative control; D, indicates the age of the animal in days; Donor, 2.4 kb PCR amplicon; Excision, 233 bp PCR amplicon; MW, 100-bp molecular standard; *Gapdh*, demonstrate equal genomic DNA template loading (100 ng) used for PCR reaction. (**b**) Tumorigenic livers extracted from 160-day old triple- (left) and quadruple- (right) experimental male transgenic littermates showing accelerated tumor formation in the latter. Arrowheads, denote smaller preneoplastic nodules in the triple-transgenic littermate; scale bar, 0.5 cm. (**c**) Semi-quantitative analysis of RT-PCR products from **Figure 3d**. The ImageJ software was used to quantify the band intensity of the RT-PCR amplicons for *Afp*, *Opn*, *Egfr* and *ErbB2*. Arbitrary units shown relative to *ß-actin* levels. Advanced tumors, n=3; Early tumors, n=13; HCC, n=1; Normal liver, n=1. Values are the mean ± SD.



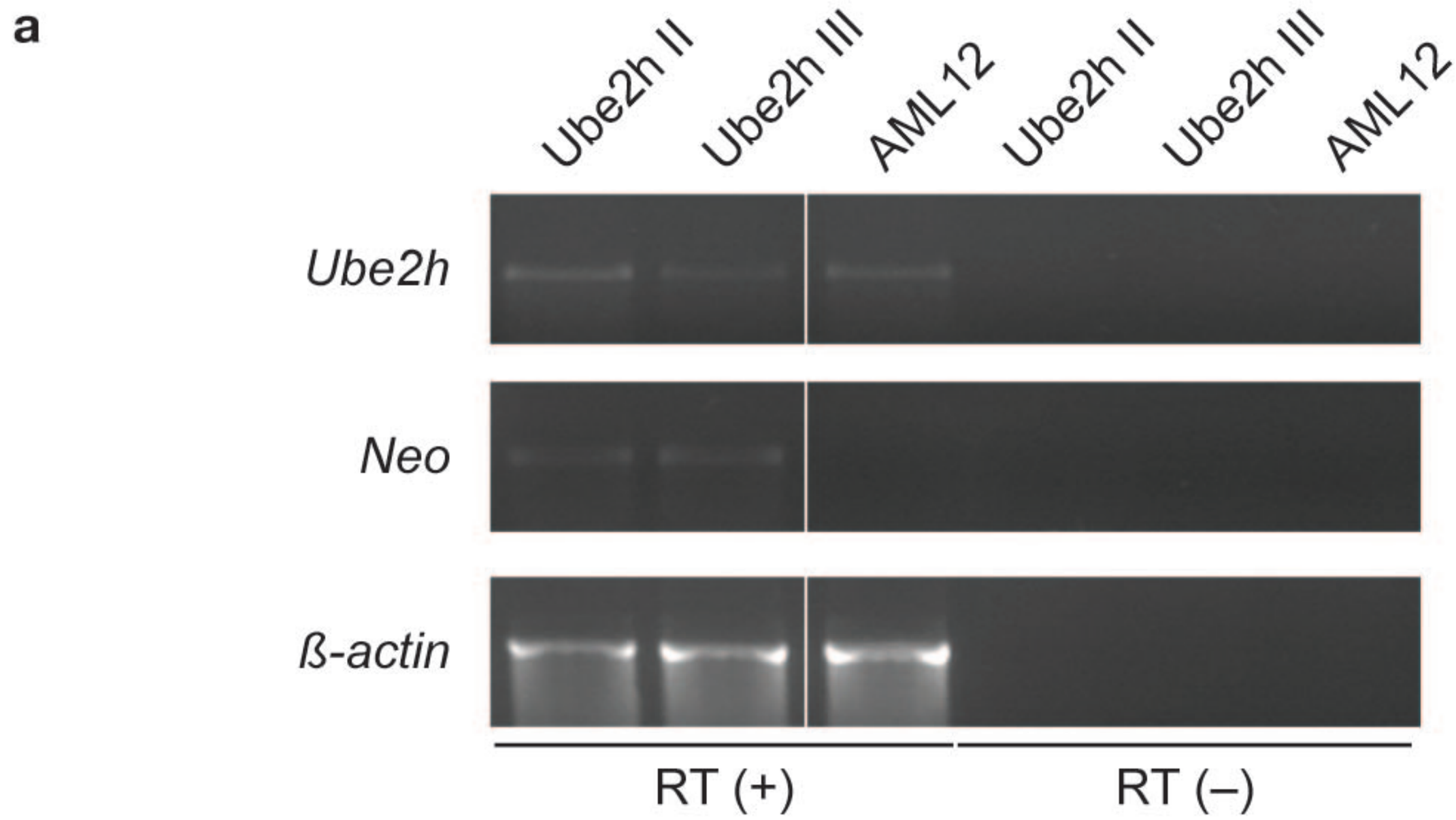
Supplementary Figure 3 Immunohistochemical (IHC) analyses of liver nodules at various stages of tumorigenesis showing positive reaction to ß-catenin. Normal liver, normal C57BL/6 liver; ATRP M81, preneoplastic nodules from a 156-day quadruple transgenic male mouse; ATRP M121, preneoplastic nodules from a 178-day quadruple transgenic male mouse; ATR M81, preneoplastic nodules from a 460-day triple transgenic male mouse; *NRAS* liver tumor, HCC control taken from a tumorigenic liver over-expressing *NRAS G12V* oncogene. Negative control, IHC of liver sections not treated with the primary antibody; Primary antibody, IHC of serial liver sections treated with the indicated primary antibody; scale bars, 100 μm.

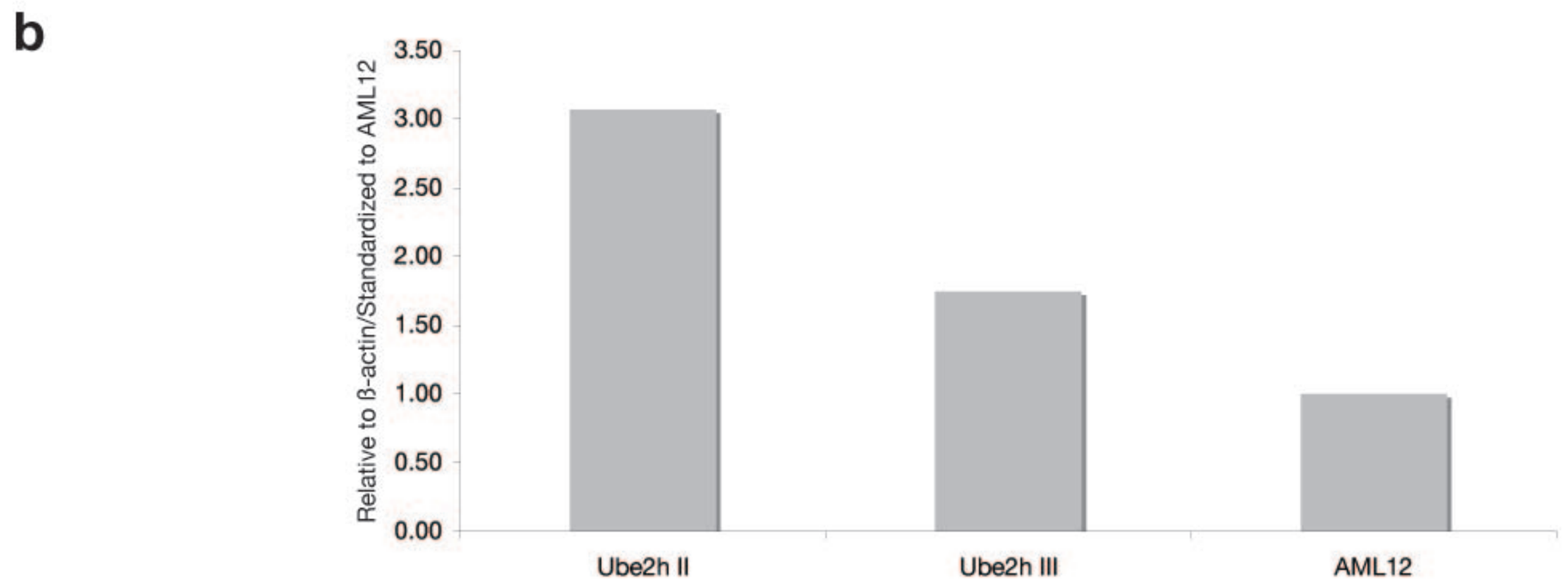


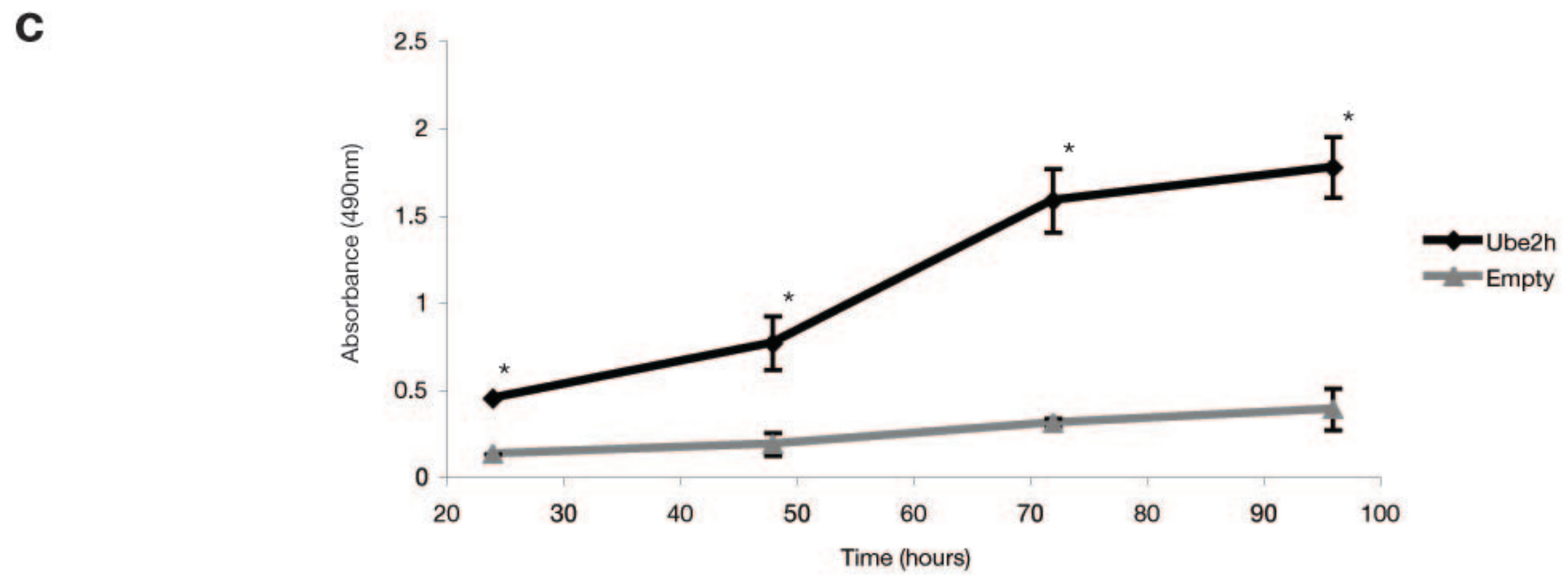
Supplementary Figure 4 Analyses of liver samples from female experimental animals and Western blot analyses for truncated Egfr. (**a**) Immunohistochemical (IHC) analyses of a 344-day old non-tumor producing quadruple-transgenic female liver showing positive reaction to Afp and Ki67 (arrowheads). These sections were also IHC positive for SB and Alb (data not shown). Negative control, IHC of liver sections not treated with the primary antibody; Primary antibody, IHC of serial liver sections treated with the indicated primary antibody; scale bars, 100 μm. (**b**) Western blot analysis for the truncated Egfr protein. Top panel, Using a phospho-Egf receptor (Tyr845) antibody, the truncated Egfr was detected at around the 150 kDa size (open arrowhead) and the wild-type Egfr can be weakly seen in some of the samples at around the 170 kDA size (arrowhead). Quadruple experimental animals: ATRP M81 preneoplastic liver sample, 156-days; ATRP F6 non-tumor producing liver, 279-days; ATRP F67 non-tumor producing liver, 344-days. Triple experimental animals: ATRP M175 preneoplastic liver sample, 375-days; ATRP M51 individual preneoplastic liver samples, 330-days; ATR M71 HCC sample and lung metastasis, 440-days. Truncated Egfr was detected in majority of samples, faintly in the lung metastasis but not in B6 and ATR M71 liver samples. Bottom panel, GAPDH monoclonal antibody was used to demonstrate protein loading for the Western blot. Lung, lung metastasis; B6, protein isolated from C57BL/6 liver.



Supplementary Figure 5 Integrative genomic analysis of CIS candidate genes in human hepatitis C virus (HCV)-related hepatocellular carcinoma (HCC). (**a**) Gene expression of 3 candidate genes in human HCC: *MAP2K4*, *QKI* and *UBE2H*. (**b**) Gene expression and its correlation with DNA copy numbers of *EGFR* and *MET* in human HCC. Left panels (**a**,**b**), changes in gene expression in the whole spectrum of human HCC. Normal, normal liver; Cirrhosis, cirrhotic tissue; LGDN, low-grade dysplastic nodules; HGDN, high-grade dysplastic nodules; HCC, hepatocellular carcinoma. Right panels (**a**,**b**), correlation between DNA copy numbers and gene expression for each candidate gene (log2 expression values). (**c**) Association between *UBE2H* expression and overall survival of HCV-induced HCC patients. Non-significant trend towards poorer survival associated with high *UBE2H* expression (red line) compared with low expression levels (black line) in HCV-induced HCC patients.







Supplementary Figure 6 Validating the effect of over-expressing *Ube2h* in AML12 cell line by cell proliferation assay. (**a**) RT-PCR of AML12 cells stably transfected with the *Ube2h* expression vector. Three different cell populations of *Ube2h* transfected cells from separate transfection experiments were generated. Representative RT-PCR of 2 transfected cell populations are presented. AML12, normal untransfected AML12 cells; RT (+), first strand cDNA synthesis with reverse transcriptase added; RT (–), first strand cDNA synthesis without reverse transcriptase. Presence of *Neomycin* (*Neo*) resistance gene was confirmed in *Ube2h* stably transfected cell populations. (**b**) Semi-quantitative RT-PCR using ImageJ was used to confirm the over-expression levels of *Ube2h* in transfected cells. Values are relative to *ß-actin* levels and standardized to untransfected AML12. (**c**) Representative graph showing the proliferative effect of over-expressing *Ube2h* in AML12 cells. Cells were initially seeded at 10,000 cells and cell proliferation rate determined using the CellTiter 96 Aqueous One Solution Cell Proliferation Assay at the indicated time points. Ube2h, AML12 transfected with the *Ube2h* expression vector; Empty, AML12 transfected with the pcDNA empty vector. Values are the representative mean ± SD of experiments done in triplicate, *$p<0.01$.

Supplementary Table 1 Frequency of liver tumor nodules in experimental mice

Experimental animal	Genotype	Sex	Age (days)	Number of visible tumors
ATR M11	Triple	Male	105	0
ATR M15	Triple	Male	105	0
ATRP M66	Triple	Male	160	3 (3)
ATR M36	Triple	Male	215	0
ATRP M134	Triple	Male	223	30
ATR M31	Triple	Male	250	3
ATR M41	Triple	Male	289	0
ATRP M175	Triple	Male	375	Massive hepatic adenomas
ATR M71	Triple	Male	440	HCC with lung metastases
ATR M81	Triple	Male	460	HCC with lung metastases
ATRP M2	Quadruple	Male	101	0
ATRP M81	Quadruple	Male	156	25 (25)
ATRP M94	Quadruple	Male	159	8 (8)
ATRP M65	Quadruple	Male	160	21 (21)
ATRP M121	Quadruple	Male	178	13 (11)
ATRP M131	Quadruple	Male	223	0
ATRP M232	Quadruple	Male	432	HCC (3) with lung metastases (32)
ATRP F121	Triple	Female	178	0
ATR F13	Triple	Female	285	0
ATR F1	Triple	Female	310	0
ATR F42	Triple	Female	342	0
ATR F106	Triple	Female	512	Several tiny nodules seen
ATRP F31	Triple	Female	575	2
ATRP F122	Quadruple	Female	178	0
ATRP F112	Quadruple	Female	278	0
ATRP F6	Quadruple	Female	279	0
ATRP F67	Quadruple	Female	344	0
ATRP F226	Quadruple	Female	432	1

Triple, animals carrying *Albumin*-Cre (*Alb*-Cre), T2/onc and *Rosa26*-lsl-SB11 transgenes; Quadruple, animals carrying *Alb*-Cre, T2/onc, *Rosa26*-lsl-SB11 and *p53*-lsl-*R270H* transgenes. Numbers in parentheses indicate actual samples used for determining insertion sites by pyrosequencing.

Supplementary Table 2 Ingenuity pathway analysis of CIS gene list

Process Annotation	*p*-value	Genes	No.
Autophosphorylation	4.61E-09	*EGFR*, *MET*, *TAOK3*, *TRPM7*, *VRK2*	5
Regression of tumor	8.09E-06	*EGFR*, *HIF1A*, *MET*	3
Moiety attachment of protein	1.52E-05	*EGFR*, *MET*, *TAOK3*, *TRPM7*, *UBE2H*, *VRK2*	6
Phosphorylation of protein	3.10E05	*EGFR*, *MET*, *TAOK3*, *TRPM7*, *VRK2*	5
Apoptosis of liver cells	6.60E-05	*EGFR*, *MAP2K4*, *MET*	3
Development of organ	8.02E-05	*EGFR*, *HIF1A*, *MAP2K4*, *MET*, *NFIB*, *PAK4*	6
Apoptosis of fibroblast cell lines	8.86E-05	*EGFR*, *HIF1A*, *MAP2K4*, *QKI*	4
Formation of tubules	1.08E-04	*HIF1A*, *MET*	2
Apoptosis of cell lines	1.28E-04	*EGFR*, *HIF1A*, *MAP2K4*, *MET*, *PAK4*, *QKI*, *TAOK3*	7

Process annotation, significant biological functions as determined by ingenuity pathway analysis. Genes and No., refers to the genes and number of genes from the CIS gene list that contribute to the predicted biological function, respectively.

Supplementary Table 3 Comparison between CIS genes with human HCC array CGH data analysis

Mouse gene name	Prediction	Human homologue	GeneID	Gain	Loss	Chr
Egfr	Truncate C-term	*EGFR*	1956	15	1	7
Sfi1	Disrupt	*SFI1*	9814	8	20	22
Zbtb20	Truncate C-term	*ZBTB20*	26137	15	8	3
Nfib	Truncate C-term	*NFIB*	4781	6	24	9
Taok3	Disrupt	*TAOK3*	51347	9	13	12
Slc25a13	Unknown	*SLC25A13*	10165	18	1	7
Qk	Truncate N-term	*QKI*	9444	14	25	6
Rnf13	Unknown	*RNF13*	11342	16	5	3
Met	Disrupt	*MET*	4233	22	1	7
March1	Unknown	*MARCH1*	55016	1	35	4
Psd3	Unknown	*PSD3*	740792	1	49	8
Map2k4	Unknown	*MAP2K4*	6416	4	39	17
Trpm7	Disrupt	*TRPM7*	54822	8	10	15
Ube2h	Unknown	*UBE2H*	7328	20	2	7
Vrk2	Truncate C-term	*VRK2*	7444	9	3	2
Hif1a	Truncate C-term	*HIF1A*	3091	8	16	14
Pak4	Truncate N-term	*PAK4*	10298	17	9	19

Prediction, probable outcome as a result of transposon insertions into CIS genes; Truncate C-term, probable C-terminus truncation resulting in gain-of-function activity; Truncate N-term, probable N-terminus truncation resulting in gain-of-function activity; Disrupt, probable gene disruption resulting in loss-of-function; Unknown, prediction not possible due to varying locations of transposon insertion sites and/or orientation. GeneID, NCBI gene identification number for the human homologue. Gain, value indicates the number of samples that had a gain in copy numbers (n=100). Loss, value indicates the number of samples that had a loss in copy numbers (n=100). Chr, chromosomal position of the human homologue gene.

Supplementary Methods

Detection of T2/onc excision

The following primers were used to detect the presence of T2/onc excision, which provides evidence of a transposition event. Primers used were forward 5'-TGTGCTGCAAGGCGATTA-3' and reverse 5'-ACCATGATTACGCCAAGC-3'. Evidence of excision is seen when an amplicon of 225 bp is seen and the lack of the excision amplicon or detection of the donor concatemer (2.4 kb), indicates no excision has occurred. To confirm the integrity of the genomic DNA template used for excision PCR, *Gapdh* PCR genotyping was also performed using the forward 5'-GGAGCCAAACGGGTCATCATCTC-3' and reverse 5'-GAGGGGCCATCCACAGTCTTCT-3' under the same conditions as previously described, with the expected amplicon of 233 bp.

Western blot analysis

Protein was isolated from tissue samples using a Norgen RNA/DNA/Protein Purification Kit (Norgen Biotek Corporation) as recommended by the manufacturer. Thirty micrograms of protein was loaded onto premade 7.5% acrylamide gels (Biorad) and transferred onto a nylon membrane using standard protocols. Membrane was blocked with 5% bovine serum albumin (BSA) in tris-buffered saline/0.1% Tween 20 (TBST) for 1 hr at room temperature, followed by overnight incubation with phospho-Egf receptor (Try845) antibody (1:1,000 dilution, Cell Signaling Technology) at 4°C with agitation. After primary antibody incubation, membrane was washed thoroughly with TBST followed by secondary antibody incubation at room temperature for 1 hr. An anti-rabbit IgG horseradish peroxidase conjugated antibody prepared in 5% BSA/TBST (1:10,000 dilution, Vector Laboratories) was used as the secondary antibody. After secondary antibody incubation, membrane was washed thoroughly with TBST and then developed using the SuperSignal West Pico Chemiluminescent Substrate (Pierce Thermo Scientific) as recommended by the manufacturer. GAPDH monoclonal antibody (1:1,000 dilution, Cell Signaling Technology) was used to demonstrate protein loading for the Western

blot. Secondary antibody and detection was done as previously described for phospho-Egf receptor (Try845) antibody.

Pyrosequencing processing steps

Library identification

Pyrosequencing reads were orientationally amplified and primed such that the 10-bp library-identifying barcode always appeared in the beginning of the sequence in the sense orientation. Sequence quality was outstanding even up to the first base. Thus, we scanned via a custom perl script positions 1-12 of all reads for the presence of the library barcode allowing 0 or 1 mismatches. We typically found perfect matches to a single barcode at positions 1-11, with matches at 2-12 occurring rarely. Zero to one mismatch hits were essentially absent from anywhere else in the read sequences. We used this information to successfully assign 98% of all reads to a library barcode. Due to careful selection of the barcodes prior to the experiment (such that all barcodes differed by at least 2 bp), we verified that indeed no sequence matched two or more barcodes in this region.

We did not attempt to assemble reads into contigs for several reasons: (1) the read quality was outstanding, matching the consensus with greater than 99.9% accuracy when assembly was performed, (2) the contigs tended not to tile at all, since the reads were all primed at the same location and are of similar length, conferring little advantage to using contigs and (3) assembly introduced chimeric artifacts, particularly in cases where two relatively closely-spaced insertion sequences appeared on opposite strands.

Identification and removal of IR/DR and linker constructs

We applied EMBOSS Vectorstrip[1] with custom designed modifications for pipeline application and assessed the best mismatch parameters to use. We sequentially attempted to match both construct elements (31 bp IR/DR and 21 bp linker) in sense and antisense orientations with 3 successively less stringent parameter sets: (10%, 15%, and 20% mismatches, respectively). For each read, the highest stringency level was retained that identified and trimmed the most construct elements. Hence, most reads were processed at the 10% mismatch level due to the high sequence quality level. Rare reads with

unexpected construct element patterns (e.g. multiple IR/DRs or linkers, or those with inverted orientation) were inspected manually. Insert sequences with fewer than 16 bp after IR/DR and linker removal were not processed further.

Mapping of insertion sequences to the mouse genome

In matching inserts to the mouse genome (NCBI Build 37) we used BLASTN (DeCypher's TeraBLASTN, Active Motif, http://timelogic.com), requiring query sequences to align within 1 bp of the start of right-IR/DR sequenced inserts or within 1 bp of the end of left-IR/DR sequenced reads (i.e. within 1 bp of the *SB* insertion site with both types of reads). Additionally, the query was required to match with at least 95% identity (90% and 85% thresholds were tested, but failed to yield sufficiently higher percentages of newly mapable insertions to warrant lowering the matching stringency). Note that because we are most interested in the IR/DR position, we were very careful to ensure that the query matched within 1 bp of the IR/DR insertion site, but we did not require the 3' end of the query to match (in case cloning artifacts had altered that end of the sequence). If secondary genome hits were found that were at least 95% as long as the first match, their count was recorded, and the insertion location was considered ambiguous. However, if all secondary hits appeared within 15,000 bp of the primary hit on the same chromosome, we considered the insertion to be uniquely mapable to that locus.

Coalescing redundant insertions

We removed redundant sequences that arose from the same tumor library and mapped to the same TA dinucleotide insertion site in the genome. This resulted in a smaller non-redundant (nr) set of insertions.

Artifact removal

Numerous insertion sequences were deemed unreliable or artifactual and hence removed: (1) those that did not map to a bona fide TA dinucleotide immediately after the IR/DR (rare), (2) those that mapped to the same chromosome as the donor concatemer, and hence might represent local hops of the *SB* transposon and (3) those that mapped to the

vicinity of the En2 gene (chr5: 28420000.. 28500000) which was included in the original cloning construct.

Identifying common integration sites (CISs) among the mapable non-redundant insertions

Unambiguously mapped non-redundant insertions were assigned to clusters if the local density of insertions in a given window size exceeded that which would be expected by chance, as determined by exact Monte Carlo simulation (see below). For 8,060 insertions, the significance thresholds obtained are ≥ 4 insertions with 20,000 bp, ≥ 5 insertions with 65,000 bp or ≥ 6 insertions with 130,000 bp.

Statistical significance of CISs

The assumption of standard poisson statistics, that potential insertion sites are randomly distributed throughout the genome, is not strictly correct since (1) TA dinucleotides are naturally clustered in genomes, and (2) numerous unfinished regions in the mouse genome are "off-limits" since they are long tracts of Ns (e.g. the initial telomeric region of every chromosome except Y is padded with 3 million consecutive Ns). Both of these factors lead standard analytical approaches to underestimate the size and number of clusters that would actually be encountered by simply picking randomly chosen real TA sites. In other words, by ignoring the natural clustering of TA sites in the genome, the number of false positive CISs that will be predicted is systematically increased. The magnitude of deviation gets larger as more and more insertion sites are scattered about the genome, as one would intuitively expect. Hence, we wrote a program to exactly compute the expected number of CISs of a given size in a specified window across all the chromosomes that one would encounter by chance via Monte Carlo simulation. The observed number of unambiguous mapable non-redundant insertions was used for each chromosome separately as input. For example if chromosome 1 and 2 had 2,100 and 1,420 insertions, respectively, then we simply randomly distributed 2,100 insertions among the real TA dinucleotide sites on mouse chromosome 1 and another 1,420 among the TA sites of chromosome 2. Once the total count of insertions was randomly distributed among the real TA sites across the whole genome, a tally of the number of

CISs of size ≥ 3, ≥ 4, ..., ≥ 15 was recorded within windows of 10,000 bp, 20,000 bp, ... 150,000 bp. This process was repeated 100 times, and the average counts over those 100 iterations were computed. Four independent simulations of 100 iterations each were performed, yielding standard error bars between simulations of less than 1%, indicating sufficient convergence. The values obtained can be interpreted as Expect values (E-values), as they indicate the expected number of CISs of a given number of insertions that would be observed within a given window size merely by chance. We chose an E-value threshold E<1 for all experiments. Thus, if one observes 18 CISs of ≥ 5 insertions within 65,000 bp, and not even a single one was expected, this is highly significant. We compared the thresholds obtained by this method to the ones obtained using standard Poisson statistics (with the assumption of random insertion in the genome) and found this method to be uniformly more stringent (i.e. yielding fewer false-positives when applied). To avoid the biases of gene size, the CIS analyses we performed are carried out on fixed sized windows of the genome, irrespective of whether any genes are in or near the windows. The sizes of the windows examined are determined purely by statistical concerns (i.e. based on the total number of insertions we compute via random statistical simulations, the size of the largest window for which one would not expect to see a chance cluster of 2 insertions. Then we find the largest window for which no observations of 3 insertions should be found, then 4, etc.). Only after the fixed-window CISs are determined do we check which nearby genes might be affected. Since in this study we only considered windows smaller than 130 kB, we may be biased against discovering very large genes that have disruptive insertions evenly spaced throughout the gene (since there may be no windows smaller than 130 kB that have a high enough concentration of insertions to rise above our statistical thresholds). So the fact that we identify disruptions in *Egfr*, *March1* and other large genes is very significant.

Annotation of reference sequences and CISs

We created two primary annotation files: one outlining details of each unique insertion, and one describing each CIS. These files provide information on the chromosomal mapping position of each insertion or CIS, redundancy information on each insertion, and characteristics on the nearest EnsEMBL gene that flanks the insertion. EnsEMBL

mappings were identified by a custom perl script that utilizes the published Application Programmer Interface (EnsEMBL API)[2]. We have also provided a prediction of the effect of the insertions on the nearest gene based on an algorithm that was designed based on the ability of the T2/onc transposon to either over-express downstream ORFs (via the MSCV LTR) or disrupt ORFs (via the two splice acceptors and bidirectional polyA signal or simply by landing within an exon).

By analyzing multiple insertions at the same location (CIS) it should be possible to predict the effect the transposon is having on the nearby gene based on the direction of the ORF and the direction and location of the transposon insertion. The rules that follow are based on the following three assumptions:

1) The MSCV LTR most likely acts in a directional manner, driving expression of downstream ORFS, but not driving expression of upstream ORFS. For reasons of convenience we will ignore the possibility that LTR can have an enhancer effect on upstream ORFS.
2) The splice acceptors/polyA signal will disrupt splicing when the transposon is located within an intron in either direction.
3) The transposon will disrupt transcription of a gene when it lands in an exon.

The "Predicted Effect on Gene" is based on the following decision rules:

1) "Drive-intact": >=75% of insertions in CIS are 1-20,000 bp upstream of a gene AND in same orientation.
2) "Enhance-intact": Rule 1 is False: >= 75% of insertions in the CIS are 1-20,000 bp upstream "OR" (exclusive or) >= 75% of insertions are 1-20,000 bp downstream.
   (This rule does not require a directionality bias and is based on the idea that the LTR can have an enhancer effect even if it is in the wrong orientation and/or is downstream).
3) "Drive-N term-truncate": >= 75% of insertions in the same orientation as the gene and are in a common exon or intron "OR" in two adjacent exons or introns.
   (This rule predicts the insertions are driving the creation of an aberrant protein

with the N-terminus truncated).

4) "Drive-C term-truncate": Rule 3 is FALSE: >= 75% of insertions are in a common intron or exon.
(This rule assumes the insertions are producing a C-term truncation via the bidirectional splice acceptors).

5) "Disrupt": Rules 1, 2, 3 and 4 are FALSE: between 25-75% of insertions are in the opposite orientation and no single location of the nearest gene (upstream, individual exon, individual intron, downstream) contains more than 50% of the insertions.
(The second part of this rule attempts to ensure that there is no location bias to the insertions, which might indicate something other than a simple disruption).

6) "Unknown": Rules 1 to 5 are false.

Sequence information management

To facilitate the management of all sequence information, a mysql relational database was constructed to store (1) genotypic and phenotypic information on all mice and tumors from which the insertion sequences were derived, (2) meta-information on the sequencing runs themselves, (3) raw pyrosequence read sequences, (4) construct element matching characteristics, (5) final processed insertion sequences, (6) mapping information for each processed insert sequence to the mouse genome, (7) clustering assignments of inserts into CISs and (8) annotation information on all mapped inserts and CISs. SQL queries were performed to facilitate the merging of distinct liver tumor data sets and the annotation process.

Generating phylogenic tree

To generate a formal and mathematically rigorous phylogenetic tree describing the relationship between the primary HCC tumors and the lung metastases, all insertions found within two or more tumors from ATRP M232 were used as input to the phylogenetic calculation program “Pars”, a part of the PHYLogeny Inference Package (PHYLIP-3.68). “Pars” is a general parsimony program that carries out the Wagner

parsimony method with multiple states[3]. Wagner parsimony allows changes among all states and the criterion is to find the tree that requires the minimum number of changes.

Combined SNP-array and gene expression data

A total of 132 human samples including the whole spectrum of human hepatocarcinogenesis were analyzed: normal liver (*n*=10), cirrhotic tissue (*n*=13), low-grade dysplastic nodules (*n*=10), high-grade dysplastic nodules (*n*=8), and hepatocellular carcinoma (*n*=91). DNA and RNA were extracted from human tissue, SNP-array technology (StyI chip of the 250K Human Mapping Array set from Affymetrix), and gene expression microarray methodology (U133 Human Chip Plus 2.0 from Affymetrix) are extensively described elsewhere[4,5]. Data analysis was conducted using the R software (http://www.R-project.org).

Selection criteria for appealing clinical correlation candidates

The selection of the most appealing candidates for clinical correlations was based in 4 criteria:

- Significant correlation between copy number and gene expression determined by a Pearson's/Spearman's coefficient higher than 0.5 and a *p*-value less than 0.01. *p*-values for Spearman's coefficient were adjusted according to Bonferroni's correction.
- Increase or decrease in copy number changes (in more than 10% of samples) in comparison with matched cirrhotic (upper and lower limit of copy number changes in cirrhotic are represented by dashed lines in the figure).
- Significant up/down-regulation of gene expression in HCC in comparison to normal liver (*t*-test with *p*-value less than 0.01)
- Concordance between gain/loss in DNA changes and up/down-regulation.

Association between candidate genes and clinico-pathological variables in patients

A cohort of 82 HCV-related HCC patients treated with liver resection was analyzed. Clinical characteristics of the patients included in the study are also available as previously described[4]. Clinical variables included in the analysis were: tumor size (U-

Mann Whitney), number of nodules (Fisher's Exact Test), presence of vascular invasion (Fisher's Exact Test), degree of differentiation (Fisher's Exact Test), BCLC stage (Fisher's Exact Test), AFP levels (U-Mann Whitney) as well as time to death and time to recurrence (Kaplan-Meier curves and log-rank test). We arbitrarily selected a cut-off that included those patients with top 10% gene downregulation (for *MAP2K4* and *QKI*) or upregulation (for *UBE2H*), since we failed to find an optimal cut-off after training ROC curves.

Supplementary Methods Reference

1. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277 (2000).

2. Stabenau, A. *et al.* The Ensembl core software libraries. *Genome Res.* **14**, 929-933 (2004).

3. Felsenstein, J. Mathematics vs. Evolution: Mathematical Evolutionary Theory. *Science* **246**, 941-942 (1989).

4. Wurmbach, E. *et al.* Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. *Hepatology* **45**, 938-947 (2007).

5. Chiang, D.Y. *et al.* Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. *Cancer Res.* **68**, 6779-6788 (2008).