Genomics 103 (2014) 189–203

Contents lists available at ScienceDirect

Genomics

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

Decoding complex patterns of genomic rearrangement in hepatocellular carcinoma[☆]

Julio Fernandez-Banet [a], Nikki P. Lee [b], Kin Tak Chan [b], Huan Gao [c], Xiao Liu [c,d], Wing-Kin Sung [e,f], Winnie Tan [b], Sheung Tat Fan [b], Ronnie T. Poon [b], Shiyong Li [c], Keith Ching [a], Paul A. Rejto [a], Mao Mao [a,g], Zhengyan Kan [a,*]

[a] Pfizer Oncology, San Diego, CA, USA
[b] Department of Surgery, University of Hong Kong, Hong Kong, China
[c] BGI-Shenzhen, Shenzhen, China
[d] Department of Biology, University of Copenhagen, Copenhagen, Denmark
[e] School of Computing, National University of Singapore, Singapore
[f] Computational and Systems Biology, Genome Institute of Singapore, Singapore
[g] Asian Cancer Research Group, Inc., Wilmington, DE, USA

ARTICLE INFO

Article history:
Received 12 October 2013
Accepted 11 January 2014
Available online 21 January 2014

Keywords:
Genomic rearrangement
Hepatocellular carcinoma
Whole-genome sequencing
Copy number variation

ABSTRACT

Elucidating the molecular basis of hepatocellular carcinoma (HCC) is crucial to developing targeted diagnostics and therapies for this deadly disease. The landscape of somatic genomic rearrangements (GRs), which can lead to oncogenic gene fusions, remains poorly characterized in HCC. We have predicted 4314 GRs including large-scale insertions, deletions, inversions and translocations based on the whole-genome sequencing data for 88 primary HCC tumor/non-tumor tissues. We identified chromothripsis in 5 HCC genomes (5.7%) recurrently affecting chromosomal arms 1q and 8q. Albumin (*ALB*) was found to harbor GRs, deactivating mutations and deletions in 10% of cohort. Integrative analysis identified a pattern of paired intra-chromosomal translocations flanking focal amplifications and asymmetrical patterns of copy number variation flanking breakpoints of translocations. Furthermore, we predicted 260 gene fusions which frequently result in aberrant over-expression of the 3′ genes in tumors and validated 18 gene fusions, including recurrent fusion (2/88) of *ABCB11* and *LRP2*.

1. Introduction

Hepatocellular carcinoma (HCC) is the major histological subtype of liver cancer, the third leading cause of cancer mortality worldwide with high prevalence in Asia and sub-Saharan Africa. Hepatitis B virus (HBV) infection is believed to cause the majority of HCCs while other etiological factors include hepatitis C viral (HCV) infection, alcoholism and aflatoxin B1 exposure [1]. Characterization of the molecular pathogenesis of HCC could have a major impact on the diagnosis and treatment of this disease with few effective therapies [2,3]. Significant progress has been made to uncover genetic aberrations in HCCs [4], including identification of mutations in p53 (*TP53*) and β-catenin (*CTNNB1*), amplifications of *MYC*, *FGF19* and cyclin D1 (*CCND1*), over-expression of ErbB and cMet receptors, and HBV integrations into the *TERT* and *KMT2B* gene loci. Recent next generation sequencing studies [5] have further implicated chromatin remodeling pathway genes such as *ARID2*, *ARID1A* and *KMT2C* as potential drivers of HCC carcinogenesis. However, the genomic landscape of somatic genomic rearrangement (GR) remains poorly characterized in HCC. Somatic genomic rearrangement is known to induce oncogenic gene fusions such as *TMPRSS2-ETS* in prostate cancer [6] and *EML4-ALK* in non-small cell lung cancer [7]. The advent of whole-genome and transcriptome sequencing provides opportunities to comprehensively characterize large-scale and complex genomic variations at single base-pair resolution [8,9]. Here we report a comprehensive study of somatic genomic rearrangements and gene fusions in HCC based on whole genome sequencing (WGS) of a cohort of 88 tumors and matched normal samples.

2. Materials and methods

2.1. Whole-genome sequencing

Liver tumor and matched adjacent non-tumor tissues were collected with written informed consents from 88 Chinese HCC patients who received surgical treatments at Hong Kong Queen Mary Hospital as previously described [10]. Approval for the use of clinical specimens for research was obtained from Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB). The vast majority (92%, $n = 81$) of patients in

[*] Corresponding author at: 10770 Science Center Drive, San Diego, CA 92121, USA. Fax: +1 8585264214.
E-mail address: zhengyan.kan@pfizer.com (Z. Kan).

0888-7543/$ – see front matter © 2014 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2014.01.003

this cohort were HBV carriers suffering from chronic hepatitis B or cirrhosis. WGS libraries of two different insert sizes (170-bp and 800-bp) were constructed from each sample and sequenced in 2:3 ratio on the Hiseq 2000 sequencers according to manufacturer's instructions (Illumina) [10]. The average depth of base pair coverage was 36X except for three tumor/normal pairs sequenced at 100X coverage. 90-bp paired-end reads were aligned to the hg19 reference genome, and somatic SNVs were called by SOAPsnv [11]. We used SegSeq [12] to identify copy number segments. Somatic mutation and CNV predictions [13], HBV integration site analysis [10] and gene expression profiling [14] were previously described.

2.2. Somatic GR detection and filtering

We developed a somatic GR detection and annotation pipeline consisting of four major steps (Supplementary Fig. 1). (1) Raw WGS reads were aligned to the reference human genome (hg19) by the Burrows-Wheeler Aligner (BWA) [15]. (2) The alignment outputs were screened, and soft-clipped sequences were extracted and analyzed using CREST [16], run in tumor and non-tumor samples independently. (3) GR calls were filtered as germline events if there was an exact match in the coordinates of the breakpoints with GRs identified in the matched or any other non-tumor samples. We also filtered GRs with at least one breakpoint matching a known germline event reported in DGV [17], GRs with breakpoint located within <1 kb from a gap region in genome assembly and GRs with breakpoint falling into a repeat-masked region. A GR event was called somatic only if it passes above filtering criteria and there is sufficient read coverage (≥3) at the genomic region corresponding to each GR breakpoint in the matched non-tumor sample. (4) RefSeq transcript dataset [18] was used to annotate the remaining somatic candidates. For each gene, a reference transcript was defined as the transcript having the longest protein-coding sequence. GR breakpoints were annotated based on locations relative to the reference transcript of the affected gene as “intronic”, “exonic”, “intergenic” or “promoter”, <1 kb upstream of transcription start site.

2.3. Gene fusion annotation

GR events can fuse together sequences from disparate gene loci to form gene fusions. To define a candidate gene fusion, we required transcriptional directions of the partner genes to agree in fused sequences (Supplementary Fig. 2). A gene fusion event was classified as “coding” if both breakpoints reside within the coding regions of affected genes, “UTR” if one or both breakpoints are located in the UTR or “promoter” if breakpoint at the 5′ or 3′ gene is located within the promoter region. Frame conservation status was evaluated for “coding” gene fusions. A fusion was classified as “frame-shift” if it alters the translation frame of the 3′ partner gene based on reference transcript for each of the fusion genes. “Frame-shift” fusion sequences were translated into the frame that maximally conserves the protein sequence of the 3′ gene. Alternative Methionine residues that could represent new translation initiation sites were identified. The protein domain composition in the fusion product sequence was analyzed using NCBI's conserved domain database and search tool [19].

2.4. RNA-seq experiment

Total RNA isolated with TRIzol reagent was treated with RNase-free DNaseI(New England BioLabs) at 37 °C for 10 min. The Dynabeads mRNA Purification Kit (Life Technologies) was used to isolate mRNA from the total RNA samples. The mRNA was chemically fragmented by divalent cations and converted into single-stranded cDNA using random hexamer primers and SuperscriptIIreverse transcriptase (Life Technologies). The second strand was generated to create double-stranded cDNA using RNase H (Enzymatics) and DNA polymeraseI. The cDNA product was purified by Ampure beads XP (Beckman). After converting the overhangs into blunt ends using T4 DNA polymerase and Klenow DNA polymerase, an “A” base was added to the 3′ end of the DNA fragments by the polymerase activity of Klenow fragment. Sequencing adapters were subsequently ligated to the cDNA fragment ends using T4 DNA Ligase (Enzymatics). Fragments of ~200 bps were selected by Ampure beads XP (Beckman) and enriched by 12 cycles of PCR. PCR products were sequenced by Hiseq 2000 (Illumina) according to manufacturer's instructions.

2.5. RNA-seq data analysis

Reads that contain adapter sequences, ≥10% unknown bases or ≥ 50% low quality bases (quality score ≤5) were removed before analysis. Filtered reads are mapped to reference genome (hg19) using SOAP2 [11] (http://soap.genomics.org.cn/). For the 90-bp reads, ≤5 mismatches are allowed in the alignment. The gene expression level is calculated using the RPKM method [20]:

$$RPKM = \frac{\frac{10^6 C}{NL}}{10^3}$$

C: number of reads uniquely aligned to gene of interest; *N*: number of reads uniquely aligned to all genes; *L*: gene length in bps. To assess RNA-seq support for gene fusion predictions, we performed read coverage analysis on gene fusions identified in 9 samples with RNA-seq data available. Each exon of the fusion gene was divided into 100-bp windows, and the RPKM values for each window were calculated. For cases where an exon was split by a GR breakpoint, 100-bp windows were derived for each exon segment independently. The RNA-seq read coverage flanking the GR breakpoint as well as coverage in tumors vs. matched non-tumors was compared to identify anomalous expression patterns indicative of gene fusion.

2.6. GR simulation

We repeated the following process for 1000 iterations to generate simulated GR events using the set of 4314 somatic GRs as seed, keeping the number of events per sample constant for each iteration. First, chromosome and coordinates of the observed GR breakpoints were randomized. For inter-chromosomal events, both breakpoints are randomized. For intra-chromosomal events, only one of the breakpoints was randomized, and the other breakpoint was kept in the same distance as observed with a correction applied to fit within the chromosome. The mitochondrial (MT) chromosome was excluded from this step. Simulated GRs were annotated in the same way as described previously.

2.7. Integrative analysis of GR and CNV patterns

Predicted CNV segments were filtered of segments shorter than 500 bps. We define “breakpoint juxtaposition” as an event where a GR breakpoint falls within 100 bps of the start or end coordinates of a CNV segment. The percentage of GR breakpoints were calculated for each GR type and shown in Fig. 5a. Both CNV segments upstream and downstream of the GR breakpoint were counted to calculate the relative distribution of copy gain/loss statuses for CNVs juxtaposed with a specific GR type.

The copy number profile of the 200-kb region flanking translocation breakpoints on both sides was derived from read coverage (*cov*) of tumor and matched non-tumor samples. The “mpileup” utility from the samtools package [21] was used to fetch the coverage in 100-bp windows, and the copy number (CN) for each window is calculated as the following.

$$CN = 2^{*}\left(\frac{cov_{tumor}}{cov_{non-tumor}}\right)$$

Table 1
Genes with significant GR prevalence.

Gene	Gene_description	Chrom	Length	# DEL	# DUP	# ITX	# CTX	# FUSION	# GR	*P*-value	FDR
ALB	Albumin	4	17158	1	1	0	1	1	3	2.92E−05	0.020976545
CEBPB	CCAAT/enhancer binding protein (C/EBP), beta	20	1837	1	0	1	0	0	2	3.01E−05	0.020976545
CACNG8	calcium channel, voltage-dependent, gamma subunit 8	19	27180	1	1	1	0	1	3	0.0001048	0.04873851
MCL1	myeloid cell leukemia sequence 1 (BCL2-related)	1	5188	1	0	0	1	1	2	0.0002064	0.050505452
CYP1A1	cytochrome P450, family 1, subfamily A, polypeptide 1	15	5995	1	1	0	0	2	2	0.00022	0.050505452
EMILIN3	elastin microfibril interfacer 3	20	6893	1	1	0	0	0	2	0.0002896	0.050505452
HAPLN4	hyaluronan and proteoglycan link protein 4	19	7145	1	0	1	0	0	2	0.0002896	0.050505452
C11orf87	chromosome 11 open reading frame 87	11	7048	0	0	1	1	0	2	0.0002896	0.050505452
MYADML2	myeloid-associated differentiation marker-like 2	17	7589	0	1	1	0	0	2	0.0003765	0.057541797
MERTK	c-mer proto-oncogene tyrosine kinase	2	130755	2	2	0	0	1	4	0.0004125	0.057541797
AXIN1	axin 1	16	65237	2	0	0	1	1	3	0.0007382	0.093613504
HRASLS2	HRAS-like suppressor 2	11	10614	0	1	0	1	1	2	0.0008319	0.096705353
ATG9B	ATG9 autophagy related 9 homolog B (S. cerevisiae)	7	12290	1	1	0	0	0	2	0.000912	0.09786322

Significantly affected genes based on the number of GR events affecting gene coding regions. Shown in table are the numbers of tumors harboring GR events in different categories. # FUSION: the number of tumors where a gene is fused with another gene as a result of GR. # GR: the total number of tumors where a gene is affected by various GR events.

The copy number value for each window is then smoothed by averaging over 10 consecutive windows and then normalized by dividing by the mean CN across the 200-kb region.

For DUP and DEL events called by CREST, the read sequences flanking the breakpoint must both align to the reference genome in the same strand orientation. Reads supporting ITX events always align to the opposite strands (+/− or −/+) whereas the strand orientations for reads supporting CTX events may be the same (+/+, −/−) or opposite (+/−, −/+). For DUP, DEL and CTX breakpoints with the same strand orientations, the 5′ breakpoint is simply shown at the left side and the 3′ breakpoint at the right side. The windows are numbered in the 5′ to 3′ direction. For CTX breakpoints with opposite strand orientations, we reverse the window ordering for the breakpoint on the negative strand (Supplementary Fig. 3a). For ITX breakpoints, if the strand orientations are (−/+) then we reverse the order of the windows around the 5′ breakpoint; if the strand orientations are (+/−) then we reverse the order of the windows around the 3′ breakpoint (Supplementary Fig. 3b).

2.8. Analysis of gene-level GR prevalence

GR prevalence (*GR*) at gene level is defined as the number of tumors where GR breakpoints directly affect the gene coding region. As expected, we observed that gene-level GR prevalence is strongly correlated with genomic length of a gene (data not shown). To assess the statistical significance of recurrently affected genes taking gene length into consideration, we compared observed prevalence with background rates derived using the simulated GR sets. First, the observed and simulated

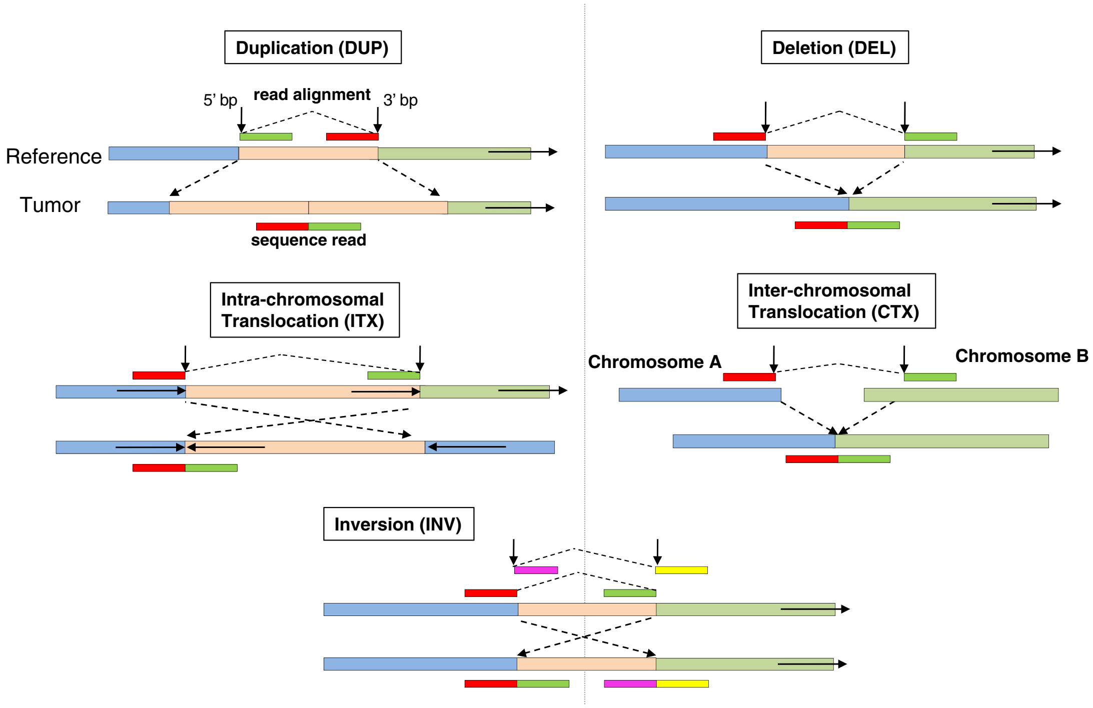

Fig. 1. Genomic rearrangement patterns. Five types of genomic rearrangement patterns are identified by CREST based on genomic alignment of soft-clipped reads. Duplications (DUP) correspond to insertions as defined by CREST.

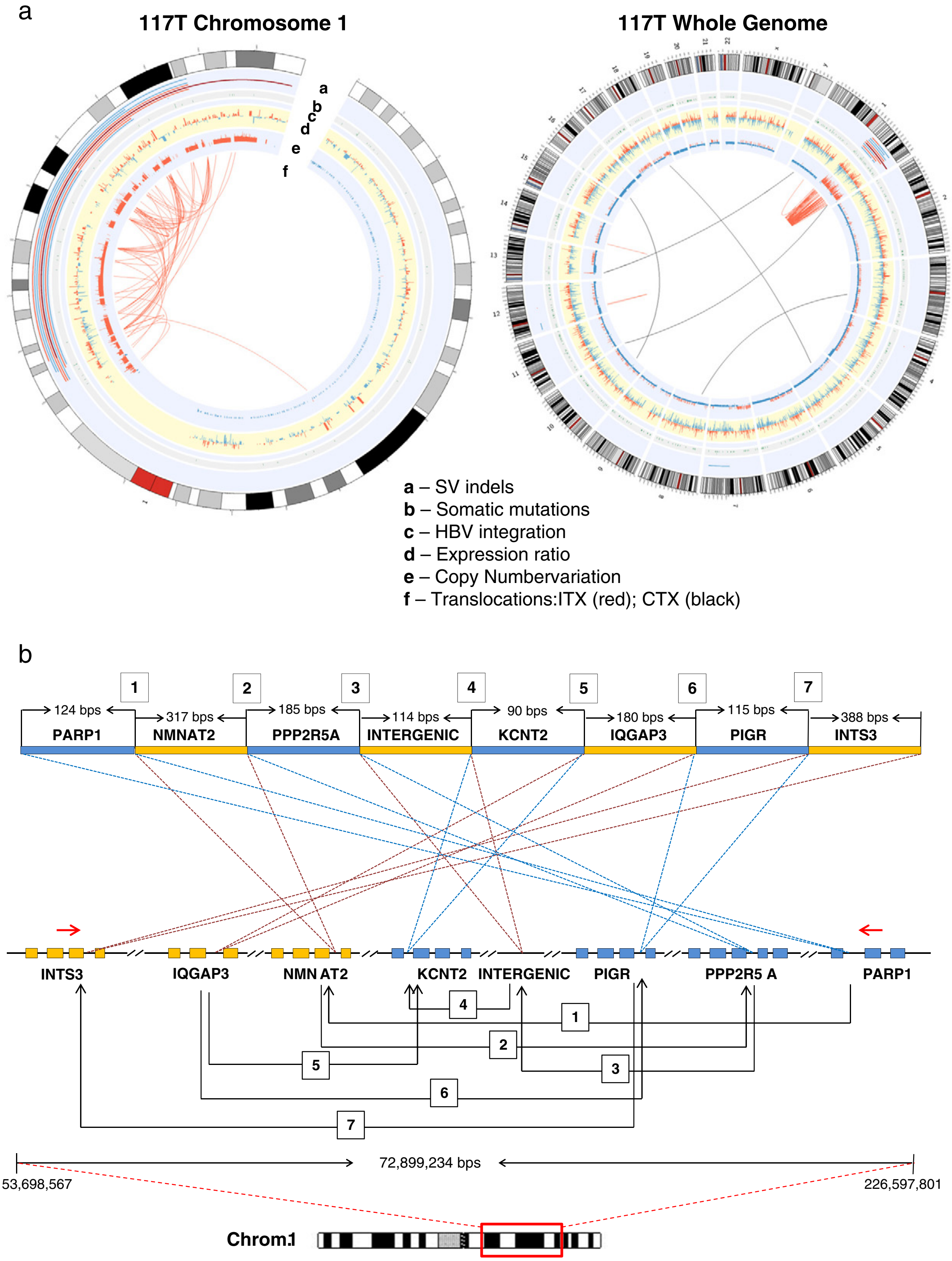

Fig. 2. Chromothripsis in HCC genomes. (a) Circos plots illustrating the chromosome and genome-wide patterns of alterations for 117 T, one of the tumors affected by chromothripsis. Chromosome ideograms are shown in the outer-most ring of the circos plot. The inner rings show indel events where blue color represents insertions and red color represents deletions, somatic mutations, copy number variations where blue represents deletions and red represents amplification, gene expression change where purple represents up-regulation and green represents down-regulation and HBV integration sites. In the ring center, black lines indicate inter-chromosomal translocation, and red lines indicate intra-chromosomal translocations. (b) Chromothripsis creates a mosaic of genomic fragments. Red arrows indicate positions of 5′ and 3′ PCR primers. Colored bar above annotates the locus origins and lengths of various segments comprised in the sequence produced by PCR-sequencing. The PCR-enclosed region is zoomed out to show the seven separate gene loci involved in rearrangement. Numbers in boxes indicate inferred translocation events. Dotted lines link genomic fragments to mapped loci in the reference genome, and arrowed lines indicate inferred patterns of translocations.

GR prevalence for all protein-coding genes were calculated for individual iterations of simulated GR data sets. Genes were then divided into 50 groups based on genomic length. The length range for each group was determined to allow ~500 genes in each group except for three groups with longer genomic lengths (>250 kb) where there are fewer genes. For each gene with observed prevalence of k and each iteration i, we counted the number of genes x with simulated GR prevalence $\geq k$ in the same length group with n genes. Summarizing the calculated proportions across 1000 iterations, we calculated p-value using the following formula.

$$p(sv \geq k|G) = \frac{\sum_{i=1}^{1000} \frac{x_i}{n_i}}{1000}$$

To estimate the false discovery rate (FDR), the p-values were adjusted using the Benjamini–Hochberg method, and significantly affected genes were selected if FDR $\leq$10%.

To investigate biochemical pathways differentially expressed as a result of *ALB* genomic alterations, we compared gene expression profiles in 9 *ALB*-altered vs. 79 un-altered cases using GSEA v2.07 [22] and the MSigDB canonical pathway gene sets [23].

2.9. Integrative analysis of GR and gene expression

Expression array data on 88 HCC tumor/non-tumor samples was previously published [14]. Gene-level expression value was derived by taking the mean of RMA normalized probe intensities. Expression change associated with each GR event is calculated as log2 fold-change of gene expression in the affected tumor relative to expression in the matched non-tumor. GR events are compared in groups classified by GR types, gene fusion types or chromothripsis statuses. A gene or gene fusion is affected by chromothripsis if the tumor genome exhibits the chromothripsis pattern in the chromosomal arm containing the gene affected.

2.10. Analysis of sample-level GR prevalence

GR prevalence at the sample-level is calculated as the total number of GRs identified in individual tumors. Tumor characteristics were derived from mutation and CNV predictions based on WGS data [13]. CIN score was calculated as follows:

$$\mathrm{CIN} = 100\left(\sum_{i=1}^{n}|\mathrm{C_i}-2|\mathrm{L_i}/2\sum_{i=1}^{n}\mathrm{L_i}\right)$$

C: copy number value of a copy number segment; L: length of a copy number segment; n: number of copy number segments. Statistical significance (p-value) of the association was determined by one-way ANOVA or Cochran–Armitage trend test (# categories >2), whichever is smaller.

2.11. Experimental validation of gene fusions

To validate predicted gene fusions in genomic DNA, we carried out PCR to capture the variation region flanking the breakpoints and performed Sanger sequencing. PCR primer pairs with expected lengths ranging from 150 bps to 1500 bps were designed, and the longer span was intended to avoid potentially complex variations near the breakpoints (Supplementary Table 1). After PCR was carried out on a GeneAmp PCR System 9700 thermal cycler (Life Technologies), products were recovered using MinElute PCR Purification Kit (QIAGEN). If there was more than one band due to unspecific amplification, gel cutting was performed. Subsequently, we sequenced all the products by Applied Biosystems 3730× DNA analyzers (Life Technologies).

To validate predicted gene fusions in RNA, we performed reverse transcription polymerase chain reaction (RT-PCR) using fusion-specific primer pairs flanking the fusion sites (Supplementary Table 2). RNA was extracted from frozen tumors and adjacent non-tumor tissues from 5 HCC cases (i.e. 11 T, 22 T, 23 T, 43 T and 81 T). Reverse transcription and PCR were then performed as described [10] using paired

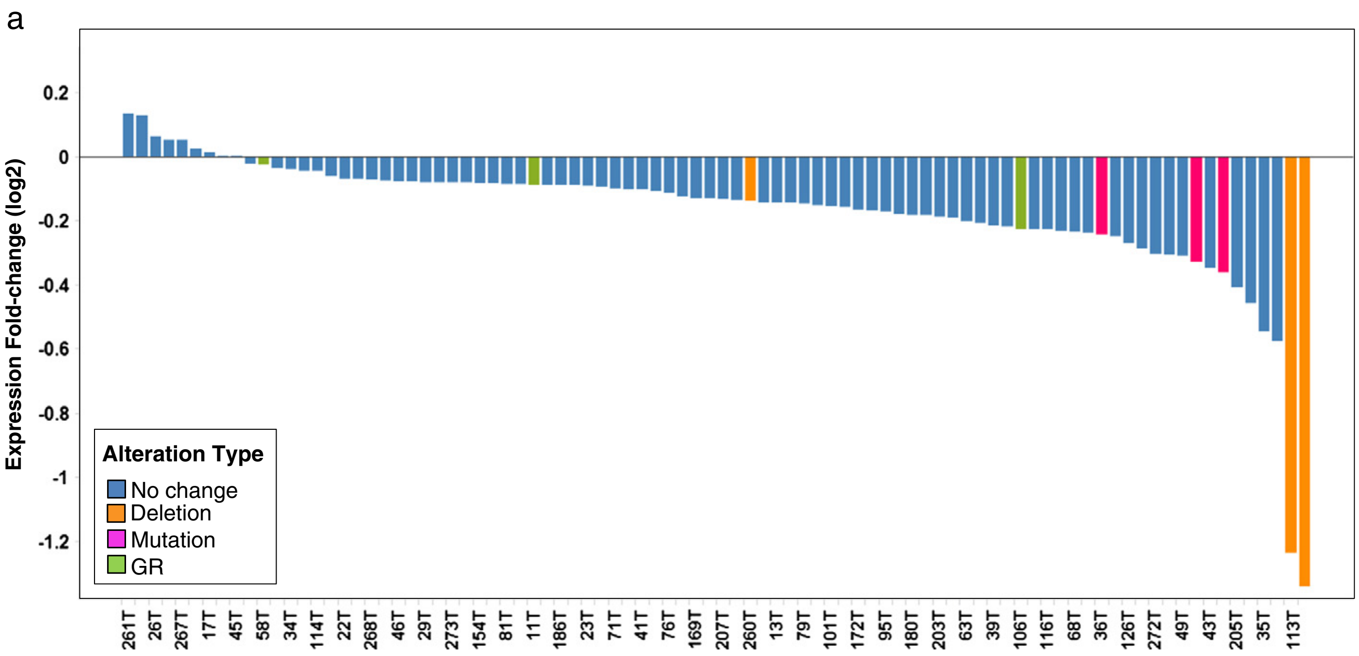

Fig. 3. *ALB* alterations may disrupt albumin production. (a) Bar chart shows the tumor/non-tumor expression change across tumors colored by *ALB* alteration statuses. (b) GSEA enrichment plots for 6 canonical pathways significantly down-regulated in *ALB*-altered vs. unaltered cases. Genes are ranked based on significance of differential expression.

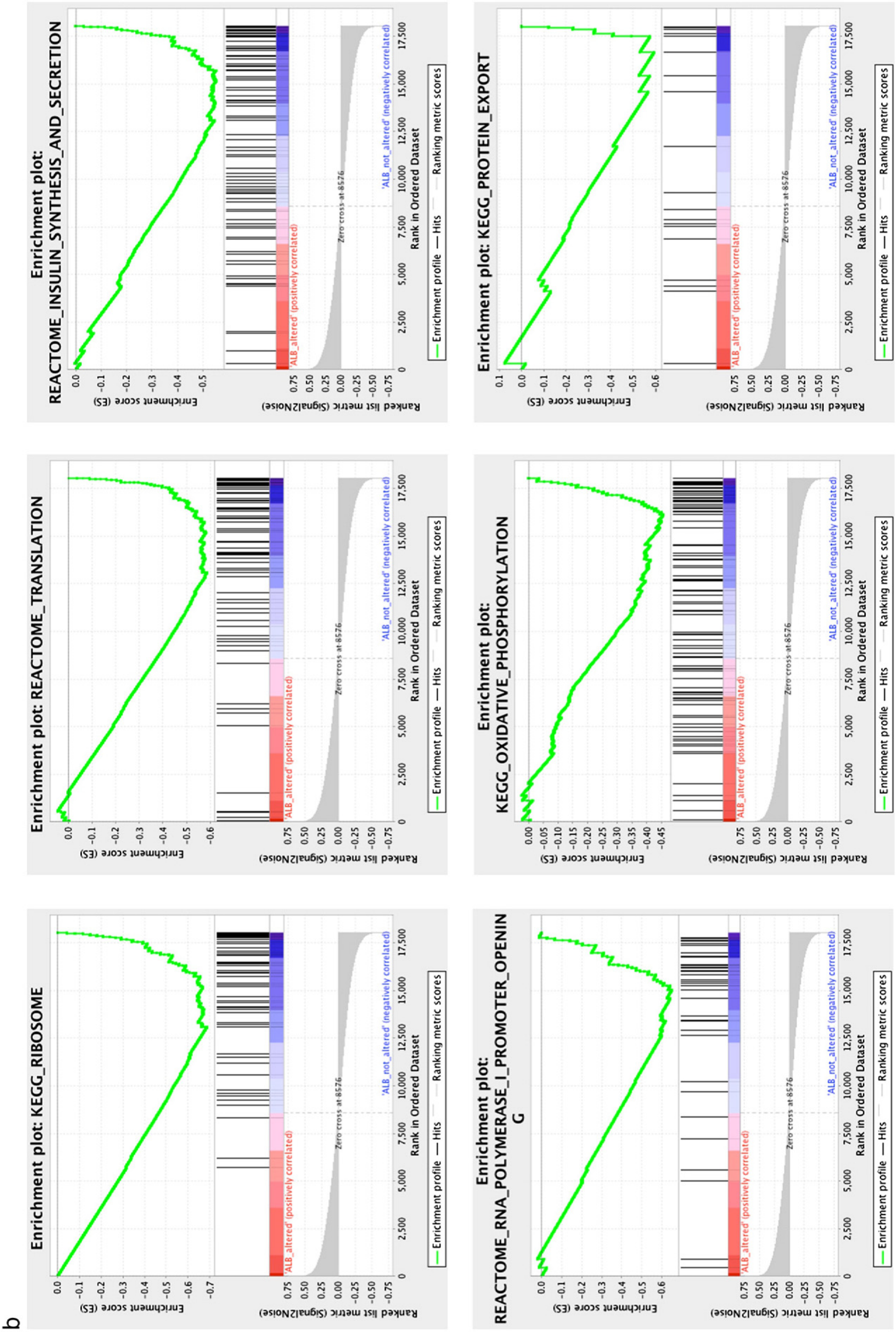

Fig. 3 (*continued*).

primers in Table 1. PCR was run for 30–35 cycles under the following condition: 94 °C for 30 s, 58–60 °C for 30 s and 72 °C for 30 s. The resulting PCR products were resolved on 1.8% agarose gel to reveal the presence of fusion genes. For all PCR products, DNA sequencing was performed to validate whether those sequences correspond to predicted sequences of the fusion genes.

3. Results and discussion

3.1. Overview and distribution analysis

We had performed whole genome sequencing on 88 HCC tumors and matched adjacent non-tumor samples, achieving mean coverage

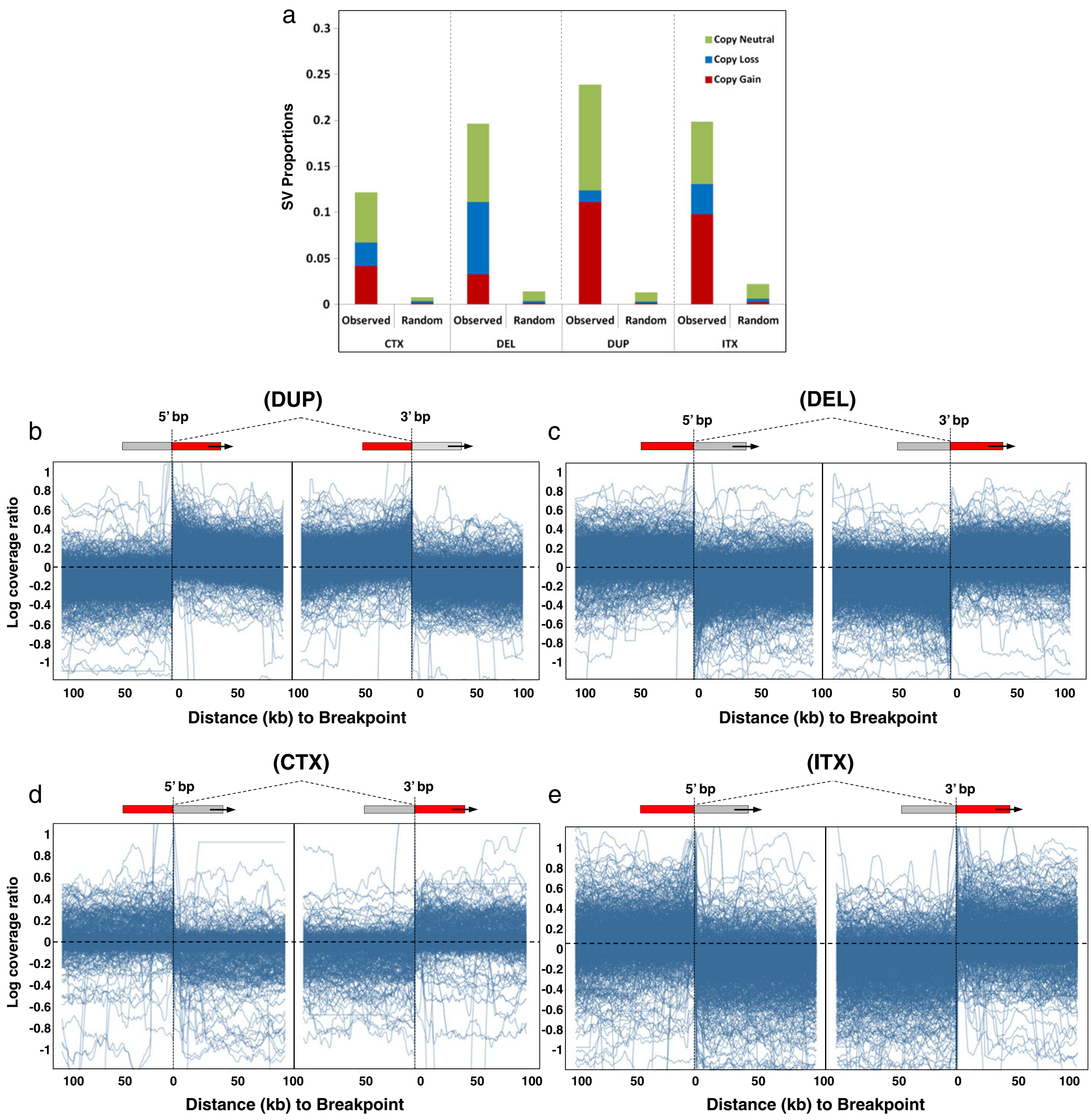

Fig. 4. Integrative analysis of GR and CNV patterns. (a) Bar chart comparing the fractions of observed GRs from different GR types having breakpoints located ≤100 bps of CNV segments with simulated GRs. Colors indicate CNV statuses for the overlapped CNV segments. CNV profiles in 200-kb genomic regions flanking 5′ and 3′ breakpoints of duplications (b), deletions (c), intra-chromosomal translocations (d) and inter-chromosomal translocations (e) are shown as line plots. Y-axis represents the CNV profile measured as the log10 ratio of WGS coverage in tumor vs. non-tumor calculated in 1-kb windows (see Methods). Each line represents CNV profile of a single GR event. Shown on top are hypothetical gene fusions resulting from four GR patterns where fusion junctions coincide with the breakpoints and fused regions are highlighted in red. For translocations, breakpoint sides and window directions were determined based on strand orientation (see Methods). CNV profiles for unbalanced (f) and balanced (g) translocations are shown as bar charts with standard errors indicated. bp: breakpoint.

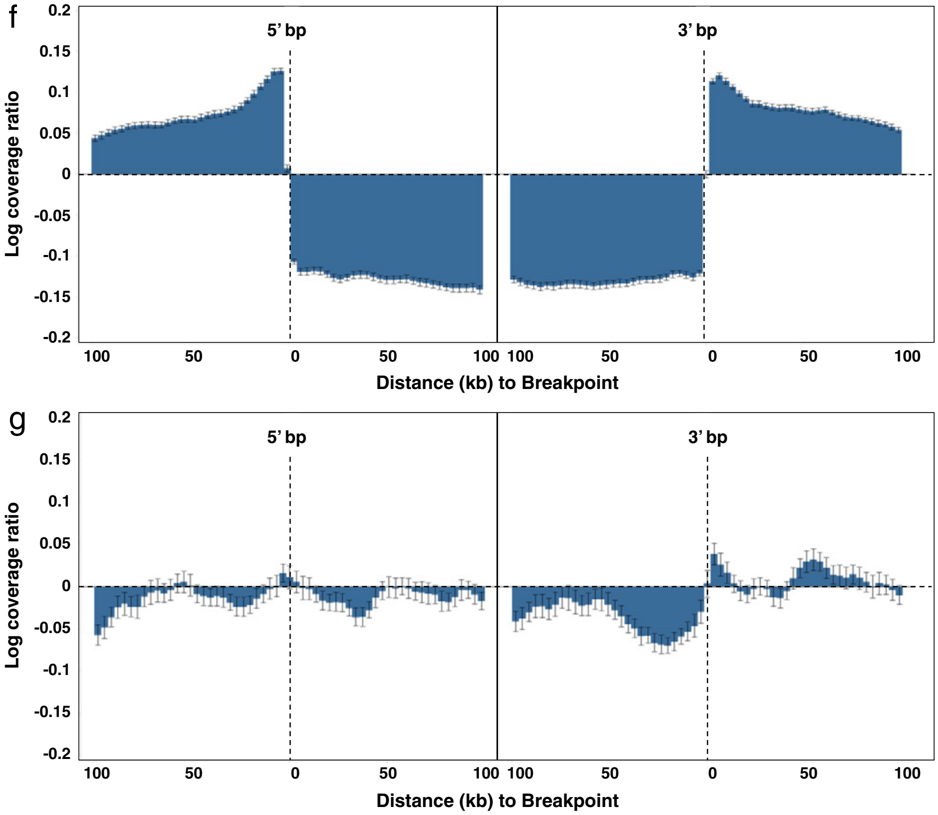

Fig. 4 (*continued*).

of 36X per base pair for 85 tumor/non-tumor pairs and 100X for 3 pairs. Analyses of somatic mutations, copy number changes and HBV integration events have been described elsewhere [10]. In this study, we developed a pipeline based on CREST [16] to detect and annotate somatic GR patterns at single-nucleotide resolution (Supplementary Fig. 1). We predicted a total of 4314 somatic GR events including 1293 duplications (DUP), 1566 deletions (DEL), 892 intra-chromosomal translocation (ITX), 554 inter-chromosomal translocation (CTX) and 9 inversion (INV) events (Fig. 1, Supplementary Table 3). The required support evidence for ITX consists of read alignments to the reference genome that exhibit a “fold-back inversion” pattern [24] believed to implicate breakage-fusion-bridge (BFB) [25] as a potential causal mechanism (Fig. 1). We also identified 260 gene fusion events in 58 tumors with an average of 4.5 events per tumor (Supplementary Table 4). There are 154 coding fusion events where both breakpoints reside in protein-coding regions and 106 UTR fusions where one or both breakpoints reside in UTRs. There is an additional 9 promoter fusions where one of the fusion breakpoints is located within a promoter region. Using PCR-sequencing, we experimentally validated 12 of the 16 in-frame coding fusions (75%) resulting from intra- and inter-chromosomal translocations (Supplementary Table 5, Supplementary Fig. 4).

The majority of GRs (58.7%; 2,532/4,314) affect coding regions of 2031 genes with breakpoints disrupting either exons or introns. Simulated GRs over 1000 iterations were generated by randomizing GR breakpoints while maintaining the same number of GRs and sample associations (see Methods). We observed that 33% (2852/8628) of somatic GR breakpoints are located in introns and 2.3% (200/8628) in exons, compared to an average of 26.4 ± 0.46% and 1.7 ± 0.14% for simulated GR breakpoints (Supplementary Fig. 5). The significant enrichment of GR breakpoints in gene coding regions suggests that active transcription may increase the accessibility of chromatin structures to genomic rearrangement, consistent with earlier report of significant correlation between open chromatin marks and sites of somatic rearrangement in breast and prostate cancer cells [26]. In addition, 11.3% (260/2,295) of gene-affecting rearrangements result in gene fusions involving 428 distinct genes. On average, only 7.6 ± 0.55% of simulated GR events result in gene fusions, indicating a statistically significant selection bias for GRs to induce gene fusions (binomial test: $p \leq 1e-6$). Nearly half (45.8%; 119/260) of gene fusions result from duplications whereas the remainder results from ITX (20.4%), deletion (18.5%) and CTX (15.4%). Given that duplications only account for 30% (1,293/4,314) of GR events, there appears to be an enrichment bias for duplication-induced gene fusions (one-tailed binomial test: $p < 1e-6$).

3.2. Chromothripsis in HCC genomes

Chromothripsis refers to a single cellular crisis where a chromosome is shattered and reassembled by DNA repair mechanism, resulting in a large number of rearrangements clustered in a chromosomal region [27]. Defining a chromothripsis event if there are >30 rearrangements and >10 translocations in a single chromosomal arm [28], we identified 5 HCC tumors (5.7%) harboring chromothripsis that affect chromosome 1q (117T, 206T), chromosome 8q (39T, 64T) and chromosome 5p (172T) (Fig. 2a, Supplementary Fig. 6). PCR-sequencing of genomic DNA from tumor 117T revealed a mosaic patchwork of eight genomic “shards”, ranging in lengths from 90 to 388 bps, derived from 8 distinct loci spanning a 72.9-Mb region on chromosome 1q (Fig. 2b). This finding matched a reported hallmark of chromothripsis where many shattered DNA fragments from a circumscribed genomic region are stitched together in a haphazard fashion. Further, we saw that chromosome 1q arm from 117T harbors a large number of intra-chromosomal translocation events that colocalize with copy number amplifications in the same region (Fig. 2a). In fact, all three chromothripsis-affected chromosomal arms harbor frequent amplifications in both chromothripsis-affected and unaffected tumors (Supplementary Fig. 7). The highest frequencies of chromosomal arm-level amplifications across the cohort were exhibited by chromosome 1q (46.5%), chromosome 8q (43.2%) and chromosome 5p (17%) (Supplementary Fig. 8). Hence, chromosomal instability and chromothripsis appeared to have converged to impact the same chromosomal regions in HCC. While chromothripsis may be a passenger effect of arm-level amplification, it could contribute to HCC oncogenesis by creating gene fusions and activating aberrant gene expressions. Four of the five tumors affected by chromothripsis harbor *TP53* alterations (3 mutated, 1 deleted), consistent with recent reports linking *TP53* mutations to chromothripsis in pediatric medulloblastoma and acute myeloid leukemia [29].

3.3. Significantly affected genes

To identify cancer genes selected for functional alterations by GRs, we assessed statistical significance of GR prevalence affecting individual genes using simulated GRs and normalizing against genomic length, a variable strongly correlated with GR prevalence (see Methods). Table 1 shows 13 genes with significant GR prevalence (FDR ≤ 10%), including known cancer genes such as *CEBPB*, *MCL1* and *AXIN1*. The most significantly affected gene is *ALB* which encodes serum albumin, the most abundant plasma protein synthesized exclusively by hepatocytes.

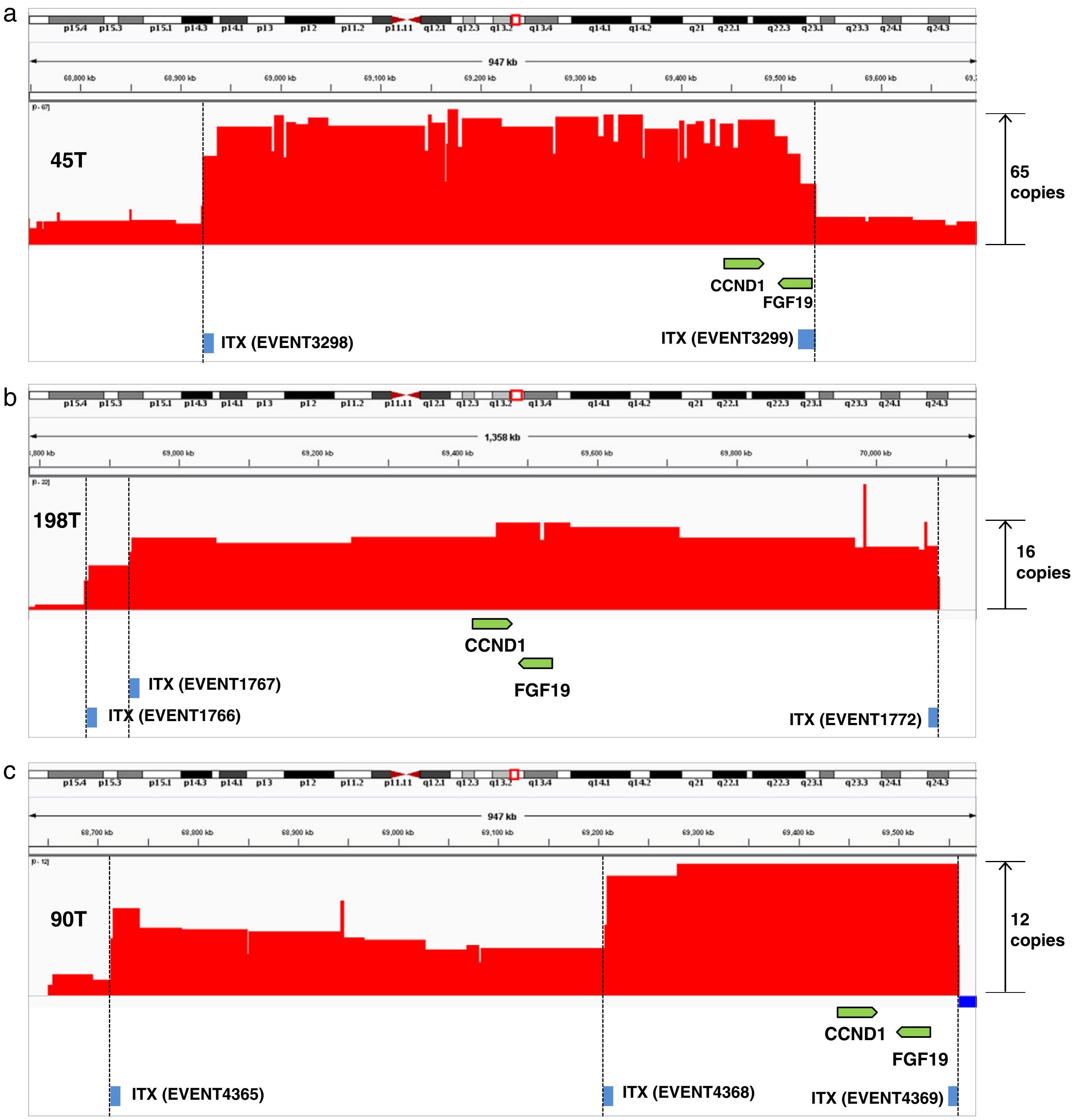

Fig. 5. GR pattern characteristic of focal amplification. IGV snapshots and drawings illustrating characteristic GR patterns for high-level, focal amplifications of *CCND1*/*FGF19* locus in 45 T (a), 198 T (b) and 90 T (c). (d) Diagram illustrating the hypothesis that two successive translocations give rise to double-minute mediated focal amplification. Copy number profiles in (d) are hand-drawn.

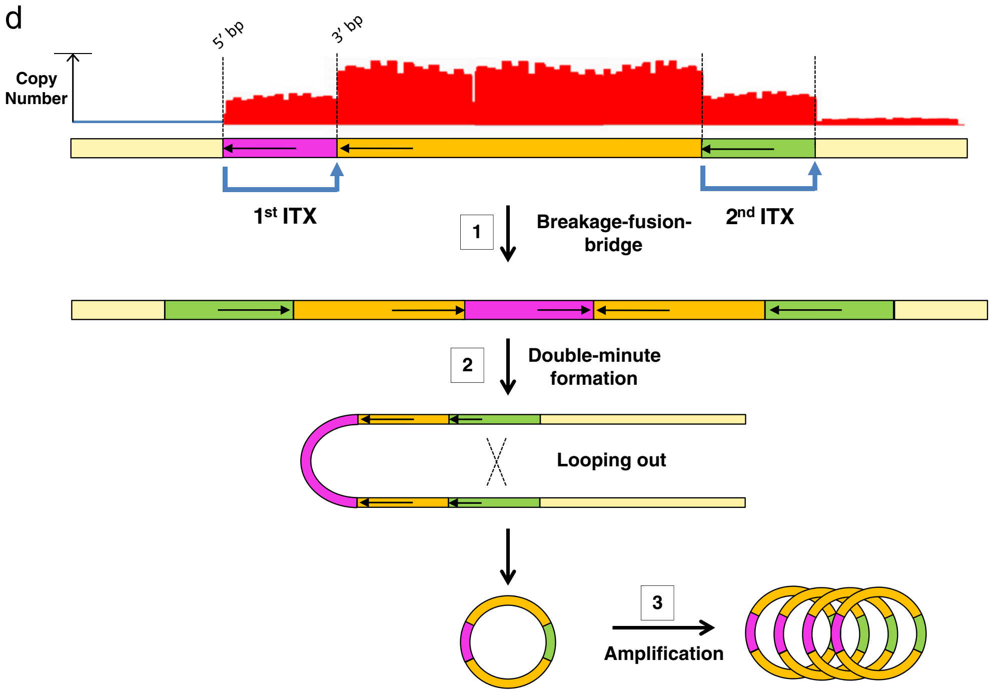

Fig. 5 (*continued*).

In addition, *ALB* harbors loss-of-function mutations and deletions in 6 additional tumors, totaling a genomic alteration frequency of 10.2% (9/88). Recurrent deletions and mutations in *ALB* have also been identified by genome sequencing in 4.8% (7/147) of another HBV-associated HCC cohort [30]. Moreover, *ALB* gene expression appears to be down-regulated in affected tumors (Fig. 3a). *ALB*-altered tumors exhibit shorter overall survival of 39.4 ± 31.3 months vs. 60.5 ± 34.2 months for the remaining cases, and worse progression-free survival of 21 ± 21.7 vs. 43.4 ± 38.4 months. GSEA analysis [22] further reveals that the most significantly down-regulated pathways in *ALB*-altered vs. unaltered tumors function in protein translation, protein transport, gene expression and mitochondrial ATP synthesis (Fig. 3b; Supplementary Table 6). These evidences taken together suggest that genomic alterations of *ALB* have functional impact, perhaps to disrupt cellular production of albumin and reallocate cellular resources for oncogenic activities.

3.4. Integrative analysis of GR and CNV patterns

Study of somatic CNV in cancer genomes provides a powerful approach to identify key genes with causal roles in oncogenesis [31]. Juxtaposition of genomic breakpoint positions between CNVs and somatic rearrangement such as translocations [32] and deletions [33] have been reported. Defining "breakpoint juxtaposition" as close proximity (≤100 bps apart) in genomic coordinates between a GR breakpoint and a CNV boundary, we found that 19.9% of observed GRs have juxtaposed breakpoints compared to an average of 1.5 ± 0.12% from the simulated GR data, a 14-fold enrichment (Fig. 4a). While there are roughly equal proportions of copy gain (8.3%) and copy loss regions (8.6%), we observe a strong bias for duplications to be juxtaposed with copy gain regions and deletions to be juxtaposed with copy loss regions (Fig. 4a). This pattern is intuitive as it reflects the causal relationships between large-scale insertion/deletion and copy gains/losses. What is unexpected is the frequent juxtaposition of inter- and intra-chromosomal translocations with CNVs, in particular the copy gain regions.

To further explore associations between rearrangement and copy number changes, we computed copy number profiles for genomic regions flanking the breakpoints of different GR types (Figs. 4b–e). We were surprised to see a distinctive pattern for translocations where the 5′ side of the 5′ breakpoint and the 3′ side of the 3′ breakpoint exhibit copy gains and the 3′ side of the 5′ breakpoint and the 5′ of the 3′ breakpoint exhibit copy losses (Figs. 4d–e). Further, this pattern appears to be specific to unbalanced translocations (Figs. 4f–g). Previously, integrative analysis of gene fusions identified from transcriptome with copy number data has suggested a "breakpoint principle" whereby fused sequences resulting from gene fusion would increase in copies whereas regions excluded from fusion would lose copies [34]. As gene fusions are caused by different types of genomic rearrangements, our findings of asymmetrical CNV patterns flanking GR breakpoints extend the breakpoint principle as a characteristic for all rearrangements, not only those involved in gene fusions (Figs. 4b–e). CNV patterns flanking ITX breakpoints might be attributable to breakage-fusion-bridge (BFB) cycles [25], expected to result in a duplicated region with increased copy number at one side of the translocation, a translocation region neutral in copy number and a deleted region on the opposite side. While it is less obvious how inter-chromosomal translocations could create a distinctive CNV pattern, one possible explanation is that double-stranded breakage associated with translocation could have promoted formation of partial or extra-chromosomal structures that result in the observed copy number differences.

The focal amplicon on chromosome 11q11.3 containing *CCND1* and *FGF19* has been shown to be an oncogenic driver of HCC [35]. In our cohort, we identified high-level, focal amplification of the *CCND1*/*FGF19* locus in 3 tumors with copy numbers ranging from 12 to 65. We were intrigued to find that in all three cases (45T, 198T and 90T) the focally amplified regions are flanked by two intra-chromosomal translocations with breakpoints juxtaposed to amplicon boundaries (Figs. 5a–c). This pattern suggests that two successive translocations had created extra-chromosomal double-minutes containing the oncogenes that had undergone amplification (Fig. 5d). Two of the most prominent focal amplifications in this cohort, each having ~20 copies in amplitude, were located on chromosomal 19p in tumor 43T. Interestingly, two amplicons are separated by an 8.2-Mb ITX translocation with breakpoints juxtaposed to the 5′ boundary of the 17-kb amplicon upstream and the 5′ boundary of the 929-kb amplicon downstream (Supplementary Fig. 9). At one end, the 17-kb amplicon overlaps with the 3′ terminal exon of the *INSR* gene and its 3′ breakpoint is juxtaposed with the experimentally validated CTX translocation between *INSR* and *LGR5* on chromosome 12. The *LGR5* locus is copy number neutral. On the other end, the 929-kb amplicon is breakpoint juxtaposed to a second ITX at the 3′ side. Since focal amplification at such high levels is unlikely to occur twice within a short range, the 8.2-Mb translocation probably occurred first to fuse the 17-kb region to the 929-kb region in inverted orientation (Supplementary Fig. 9a). A double-minute containing the fused 946-kb region is likely to have formed following the 3′ ITX, the *LGR5-INSR* translocation and undergoes subsequent amplification.

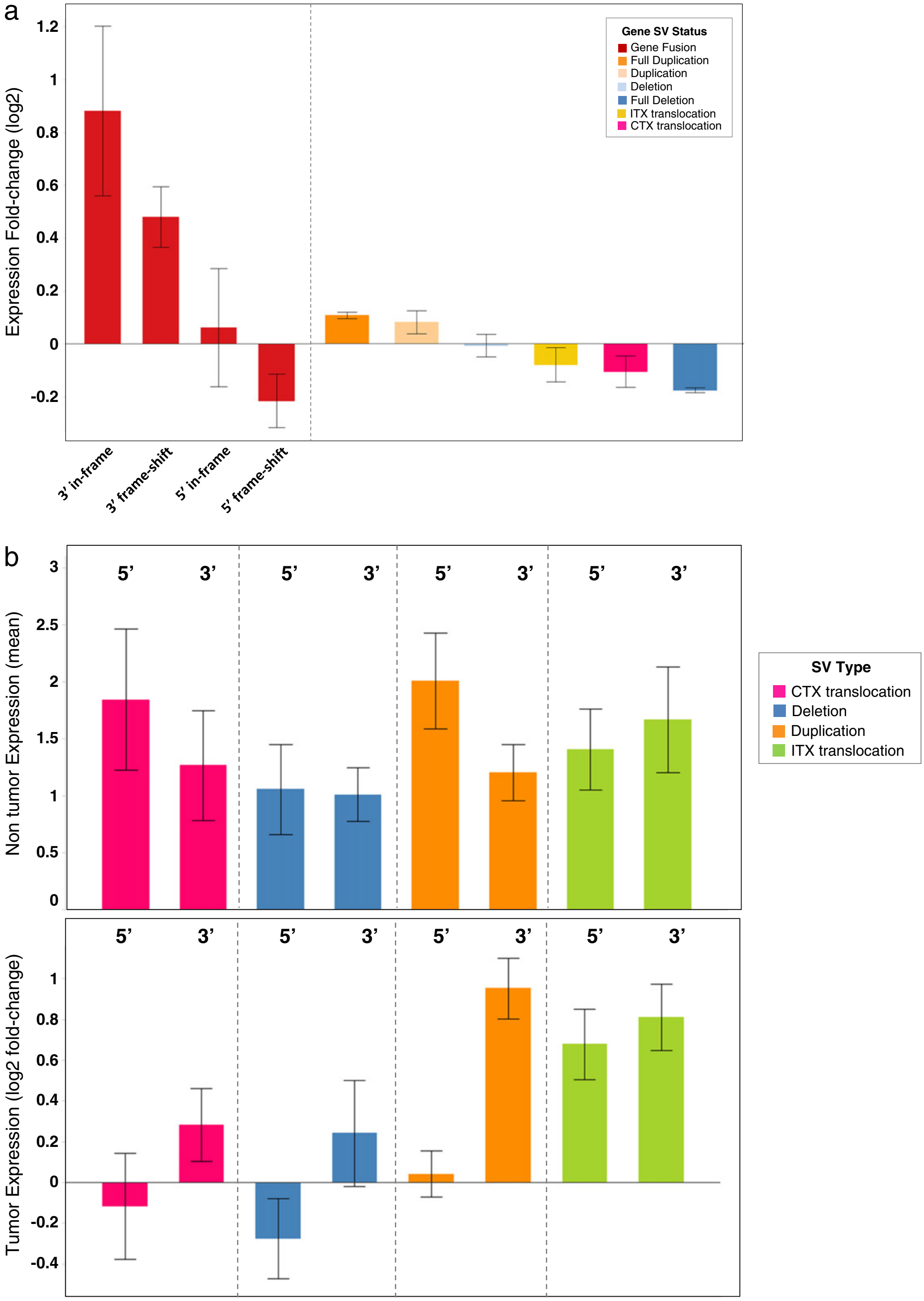

Fig. 6. Gene fusion results in tumor over-expression. (a) Bar chart comparing expression of genes affected by GRs by GR types. Y-axis represents mean expression change (log2 fold-change) in tumor relative to matched non-tumor. Each gene is classified into one of six types—duplication, deletion, ITX and CTX, full duplication and full deletion. Fusion genes are further classified based on frame status and position (5′ or 3′). (b) Bar charts showing non-tumor expression levels and tumor expression changes for 5′/3′ fusion genes classified by causal GR types. (c) Bar charts showing that tumor up-regulations of 3′ fusion genes are correlated with higher levels of 5′ gene expression in non-tumors relative to the 3′ genes. Y-axis measures the averaged expression difference (Δ) between pairs of 5′ and 3′ fusion genes in non-tumors.

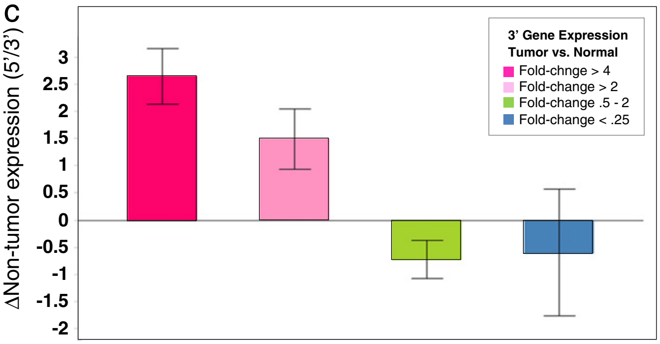

Fig. 6 (*continued*).

Amongst protein-coding genes located in the focally amplified region, *BRD4*, a bromodomain and extra-terminal (BET) family member implicated in oncogenesis [36], is highly over-expressed in 43 T and therefore a potential driver of the focal amplification (Supplementary Fig. 9b).

3.5. Gene fusion

Studies of oncogenic gene fusions have revealed that aberrant apposition of regulatory elements of one highly expressed gene to a proto-oncogene leads to outlier over-expression pattern and activation of the oncogene [37]. To assess the impact on gene expression by gene fusion and genomic rearrangement, we used the microarray expression data [14] to compare expression changes of GR-affected genes in tumors vs. matched non-tumors across GR types and gene fusion statuses (Fig. 6a). As expected, fully and partially duplicated genes exhibit higher expression than genes subject to full or partial deletions. Surprisingly, genes affected by intra- and inter-chromosomal translocations but not involved in gene fusions seem to be down-regulated as a whole. The expression array adopted a 3′ biased probe design as 78.9% of expression array probes (2643/3351) were mapped to the 3′ side of the breakpoints in the fusion genes. Hence, expression measured for the 3′ fusion genes tend to reflect expression of the fusion transcripts. We saw a striking up-regulation of 3′ fusion genes relative to other GR-affected genes (Fig. 6a). Moreover, in-frame fusions show higher expression than frame-shift fusions, possibly due to NMD of fusion transcripts with truncated open reading frames [38]. Further classifying fusion genes based on causal GR types, we see duplication inducing the highest tumor over-expression of 3′ fusion genes (Fig. 6b). In general, 5′ genes involved in fusions show higher expression levels in non-tumors than 3′ fusion genes. Further, higher levels of 5′ gene expression in non-tumors relative to 3′ partner genes correlate with greater up-regulation of 3′ partner genes in tumors (Fig. 6c). Hence, up-regulation of 3′ fusion genes in tumors due to translocation of stronger promoters from 5′ partner genes appears to be common, indicating that some of the gene fusions identified may be oncogenic events in HCC.

Gene fusion of *ABCB11-LRP2* was identified in two cases (11 T, 81 T), both resulting from duplication with one in-frame (11 T) and one out-of-frame (81 T). The in-frame fusion in 11 T fuses the 5′ portion of *ABCB11* (1–794 a.a.) to the 3′ portion of *LRP2* (1877–4732 a.a.) containing truncated extracellular domains (Fig. 7a). Normalized expression levels from RNA-seq (Fig. 7b) and expression array (Fig. 7c) both indicate that *LRP2* expressions in tumors harboring gene fusions are among the highest in the entire cohort. Similar outlier expression patterns were observed for multiple cases where gene fusion appears to induce tumor over-expression of 3′ genes (Supplementary Fig. 10). Whole transcriptome sequencing (RNA-seq) was performed on 9 of the tumor/non-tumor pairs. We computed RNA-seq coverage profiles for the 11 T tumor/non-tumor samples (see Methods) and found that *LRP2* is expressed only in the tumor. Moreover, there is aberrant expression of only the exons present in the fusion product, consistent with predicted location of the fusion breakpoint (Fig. 7a). Similar coverage profiles were seen for two other fusions found in 11 T—*TM4SF4-KCNIP3* and *EHBP1-NKD2* (Supplementary Fig. 11). We then performed RT-PCR followed by Sanger sequencing and validated *ABCB11-LRP2* fusions (Fig. 7d) along with 4 other gene fusions in tumor and matched non-tumor RNA samples: *TM4SF4-KCNIP3* (11 T), *EHBP1-NKD2* (11 T), *ABHD2-ACAN* (22 T) and *PCCA-HS6ST3* (23 T) (Supplementary Fig. 12; Supplementary Table 2).

None of the fusion genes are known oncogenes although several show emerging evidences of functional roles in carcinogenesis. *ABCB11* encodes an ABC transporter known as bile salt export pump (BSEP), responsible for the transport of cholate conjugates from hepatocytes to the bile. Germline loss-of-function mutations in *ABCB11* cause progressive familial intrahepatic cholestasis and elevate risk of HCC in childhood [39]. *LRP2* encodes megalin, a multiligand binding LDL-receptor normally expressed in epithelial cells of thyroid and kidney that functions to mediate endocytosis and transcytosis [40]. It has been shown that megalin is up-regulated in response to chemotherapy and oxidative stress in cancer cells [41–43]. Moreover, a recent study shows that clusterin, a megalin ligand, induces expression of megalin and activates survival through phosphatidylinositol 3-kinase/Akt pathway in prostate cancer cells [44]. In our HCC cohort, *LRP2* is over-expressed in tumors relative to non-tumors (Fig. 7c) and mutated in 5.7% (5/88) of cases [13]. *LRP2* is also frequently mutated in multiple cancers—16% of colon and rectum adenocarcinoma (34/212) [45], 19.7% of lung squamous cell carcinoma (35/178) [46] and 9.3% of bladder cancer (9/97) [47]. *HS6ST3* encodes heparan sulfate 6-O-sulfotransferase 3 that mediates 6-O sulfation of heparin sulfate (HS) required for interaction with a variety of growth factors and implicated in proliferation, invasion, migration and other diverse processes. HS sulfotransferases have been linked to tumorigenesis in prostate and pancreatic cancers [48,49]; therefore, tumor aberrant expression of *HS6ST3* induced by fusion may play an oncogenic role in HCC. *NKD2* encodes naked cuticle homolog 2 (Drosophila), a negative regulator of Wnt receptor signaling through interaction with Dishevelled family members [50]. *ACAN* encodes aggrecan, a member of the chondroitin sulfate proteoglycan (CSPG) family and an integral part of the extracellular matrix in cartilagenous tissue. Ectopic expression of aggrecan, along with other CSPGs, has been reported in rat HCC tissues and suspected to result from epithelial mesenchymal transition (EMT) [51]. *KCNIP3* encodes calsenilin, a member of the family of voltage-gated potassium (Kv) channel-interacting proteins. Calsenilin has been shown to regulate presenilin 1/γ-secretase-mediated N-cadherin ε-cleavage and β-catenin signaling [52]. Additional experimental characterization will be required to reveal the functional implication of *ABCB11-LRP2* and other fusions identified from our study.

4. Conclusions

Our study is the largest-scale whole-genome characterization of genomic rearrangements in HCC to date and provides a comprehensive set of somatic genomic rearrangement and gene fusion predictions, including *ABCB11-LRP2*, the first reported recurrent gene fusion in HCC. Our work revealed a complex landscape of genomic rearrangements in HCC, underscoring this deadly disease as a genomic disorder. Moreover, our finding that chromothripsis and chromosomal instability recurrently affect chromosomal arms 1q and 8q to create gene amplifications suggests that chromothripsis may contribute to hepatocarcinogenesis. Our finding of frequent loss-of-function alterations in *ALB* suggests a potential role in cancer for albumin not previously described where the wild-type protein production is disrupted to reallocate cellular resources for oncogenesis. Integrative analyses have revealed characteristic patterns of genomic rearrangement associated with focal amplifications and asymmetrical CNV patterns flanking breakpoints of genomic rearrangement, shedding light on the potential causal mechanisms.

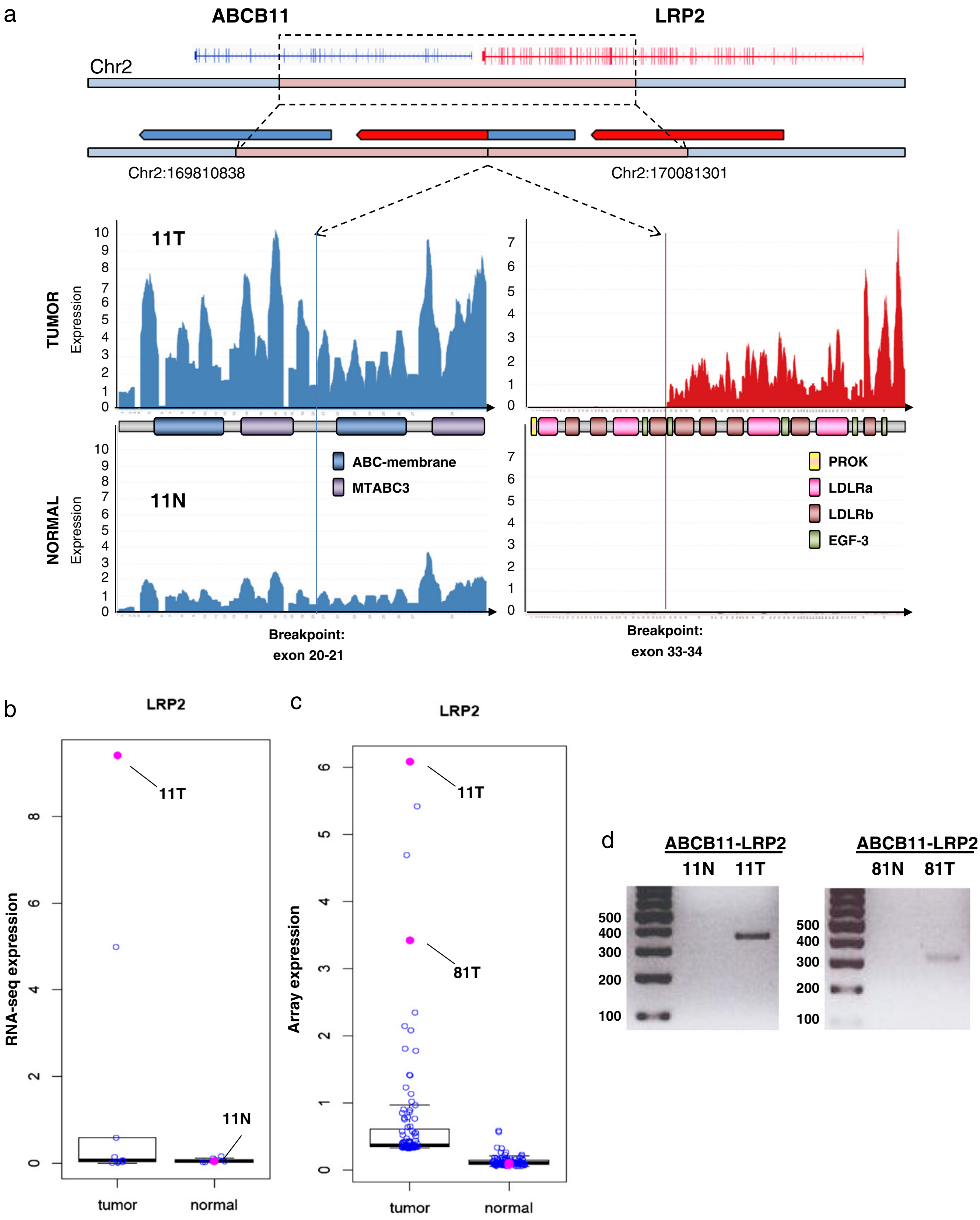

Fig. 7. Recurrent fusion of *ABCB11-LRP2*. (a) Detection of *ABCB11-LRP2* in-frame fusion in tumor 11 T and supported by transcript expression measured by RNA-seq. ABC-membrane: ABC transporter integral membrane type-1 fused domain. MTABC3: ATP-binding cassette. PROK: Prokaryotic membrane lipoprotein lipid attachment site. LDLRa: Low density lipoprotein receptor class A. LDLRb: Low density lipoprotein receptor class B. EGF-3: calcium binding EGF-like domains. Box plots comparing tumor and non-tumor expression levels of *LRP2* measured by RNA-seq (b) and microarray (c). Tumors harboring detected fusion and matched non-tumors are highlighted. (d) RT-PCR using fusion-specific primers was performed to validate predicted fusion genes in tumors 11 T and 81 T. T—tumor, N—matched adjacent non-tumor.

Genomic rearrangement is known to produce gene fusions that drive oncogenesis and often define molecular subtypes of cancers targeted for therapeutic intervention [53,54]. We found that gene fusions frequently result in up-regulation of 3′ genes in HCC and may therefore drive hepatocarcinogenesis through dysregulation of oncogenic expression. Our findings taken together suggest that genomic rearrangement is a primordial mechanism giving rise to copy number changes, gene fusions and aberrant gene expressions subsequently selected to promote carcinogenesis.

5. Database linking

The whole-genome sequencing data supporting the results of this article is available in the European Genome-phenome Archive (EGA) [accession: ERP001196]. Microarray data supporting the results of this article is available in Gene Expression Omnibus (GEO) [accession: GSE25097].

6. Authors' contributions

The study was initiated and designed by Z.K., P.R. and M.M. J.F. developed the GR detection and annotation pipeline. Data analysis was conducted by Z.K., J.F., H.G., S. L., K. C. and W.S. RNA-sequencing and genomic DNA validation were done by X.L. N.P.L., K.C. and W.T. performed RNA validation. S.T.F. and R.T.P. collected tumor specimen. Z.K. and J.F. wrote the manuscript. J.F., Z.K., N.P.L., X.L., G.H., K. C. and W.S. produced tables and figures.

Abbreviations

GR	genomic rearrangement
HCC	hepatocellular carcinoma
WGS	whole-genome sequencing
CNV	copy number variation
CIN	chromosomal instability
DEL	deletion
DUP	duplication
ITX	intra-chromosomal translocation
CTX	inter-chromosomal translocation
INV	inversion
NMD	nonsense mediated decay
kb	kilobase
bp	base pair
FDR	false discovery rate
SNV	single nucleotide variation
indel	insertion/deletion
HBV	hepatitis B virus
UTR	untranslated region

Acknowledgments

We thank Zhuolin Gong from BGI-Shenzhen for project coordination and Shibing Deng, Heather Estrella from Pfizer, Jian Wang from Eli Lilly for their contributions to statistical and expression data analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2014.01.003.

References

[1] A. Jemal, F. Bray, M.M. Center, J. Ferlay, E. Ward, D. Forman, Global cancer statistics, CA Cancer J. Clin. 61 (2011) 69–90.

[2] C. Neuveut, Y. Wei, M.A. Buendia, Mechanisms of HBV-related hepatocarcinogenesis, J. Hepatol. 52 (2010) 594–604.

[3] P.A. Farazi, R.A. DePinho, Hepatocellular carcinoma pathogenesis: from genes to environment, Nat. Rev. Cancer 6 (2006) 674–687.

[4] Z.G. Han, Functional genomic studies: insights into the pathogenesis of liver cancer, Annu. Rev. Genomics Hum. Genet. 13 (2012) 171–205.

[5] S. Li, M. Mao, Next generation sequencing reveals genetic landscape of hepatocellular carcinomas, Cancer Lett. 340 (2013) 247–253.

[6] S.A. Tomlins, D.R. Rhodes, S. Perner, S.M. Dhanasekaran, R. Mehra, X.W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J.E. Montie, R.B. Shah, K.J. Pienta, M.A. Rubin, A.M. Chinnaiyan, Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer, Science 310 (2005) 644–648.

[7] M. Soda, Y.L. Choi, M. Enomoto, S. Takada, Y. Yamashita, S. Ishikawa, S. Fujiwara, H. Watanabe, K. Kurashina, H. Hatanaka, M. Bando, S. Ohno, Y. Ishikawa, H. Aburatani, T. Niki, Y. Sohara, Y. Sugiyama, H. Mano, Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer, Nature 448 (2007) 561–566.

[8] M. Meyerson, S. Gabriel, G. Getz, Advances in understanding cancer genomes through second-generation sequencing, Nat. Rev. Genet. 11 (2010) 685–696.

[9] J.O. Korbel, A.E. Urban, J.P. Affourtit, B. Godwin, F. Grubert, J.F. Simons, P.M. Kim, D. Palejev, N.J. Carriero, L. Du, B.E. Taillon, Z. Chen, A. Tanzer, A.C. Saunders, J. Chi, F. Yang, N.P. Carter, M.E. Hurles, S.M. Weissman, T.T. Harkins, M.B. Gerstein, M. Egholm, M. Snyder, Paired-end mapping reveals extensive structural variation in the human genome, Science 318 (2007) 420–426.

[10] W.K. Sung, H. Zheng, S. Li, R. Chen, X. Liu, Y. Li, N.P. Lee, W.H. Lee, P.N. Ariyaratne, C. Tennakoon, F.H. Mulawadi, K.F. Wong, A.M. Liu, R.T. Poon, S.T. Fan, K.L. Chan, Z. Gong, Y. Hu, Z. Lin, G. Wang, Q. Zhang, T.D. Barber, W.C. Chou, A. Aggarwal, K. Hao, W. Zhou, C. Zhang, J. Hardwick, C. Buser, J. Xu, Z. Kan, H. Dai, M. Mao, C. Reinhard, J. Wang, J.M. Luk, Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma, Nat. Genet. 44 (2012) 765–769.

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

[12] D.Y. Chiang, G. Getz, D.B. Jaffe, M.J. O'Kelly, X. Zhao, S.L. Carter, C. Russ, C. Nusbaum, M. Meyerson, E.S. Lander, High-resolution mapping of copy-number alterations with massively parallel sequencing, Nat. Methods 6 (2009) 99–103.

[13] Z. Kan, H. Zheng, X. Liu, S. Li, T.D. Barber, Z. Gong, H. Gao, K. Hao, M.D. Willard, J. Xu, R. Hauptschein, P.A. Rejto, J. Fernandez, G. Wang, Q. Zhang, B. Wang, R. Chen, J. Wang, N.P. Lee, W. Zhou, Z. Lin, Z. Peng, K. Yi, S. Chen, L. Li, X. Fan, J. Yang, R. Ye, J. Ju, K. Wang, H. Estrella, S. Deng, P. Wei, M. Qiu, I.H. Wulur, J. Liu, M.E. Ehsani, C. Zhang, A. Loboda, W.K. Sung, A. Aggarwal, R.T. Poon, S.T. Fan, J. Wang, J. Hardwick, C. Reinhard, H. Dai, Y. Li, J.M. Luk, M. Mao, Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma, Genome Res. 23 (2013) 1422–1433.

[14] J.R. Lamb, C. Zhang, T. Xie, K. Wang, B. Zhang, K. Hao, E. Chudin, H.B. Fraser, J. Millstein, M. Ferguson, C. Suver, I. Ivanovska, M. Scott, U. Philippar, D. Bansal, Z. Zhang, J. Burchard, R. Smith, D. Greenawalt, M. Cleary, J. Derry, A. Loboda, J. Watters, R.T. Poon, S.T. Fan, C. Yeung, N.P. Lee, J. Guinney, C. Molony, V. Emilsson, C. Buser-Doepner, J. Zhu, S. Friend, M. Mao, P.M. Shaw, H. Dai, J.M. Luk, E.E. Schadt, Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting, PLoS One 6 (2011) e20090.

[15] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform, Bioinformatics 26 (2010) 589–595.

[16] J. Wang, C.G. Mullighan, J. Easton, S. Roberts, S.L. Heatley, J. Ma, M.C. Rusch, K. Chen, C.C. Harris, L. Ding, L. Holmfeldt, D. Payne-Turner, X. Fan, L. Wei, D. Zhao, J.C. Obenauer, C. Naeve, E.R. Mardis, R.K. Wilson, J.R. Downing, J. Zhang, CREST maps somatic structural variation in cancer genomes with base-pair resolution, Nat. Methods 8 (2011) 652–654.

[17] A.J. Iafrate, L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, C. Lee, Detection of large-scale variation in the human genome, Nat. Genet. 36 (2004) 949–951.

[18] K.D. Pruitt, T. Tatusova, G.R. Brown, D.R. Maglott, NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy, Nucleic Acids Res. 40 (2012) D130–D135.

[19] A. Marchler-Bauer, S. Lu, J.B. Anderson, F. Chitsaz, M.K. Derbyshire, C. DeWeese-Scott, J.H. Fong, L.Y. Geer, R.C. Geer, N.R. Gonzales, M. Gwadz, D.I. Hurwitz, J.D. Jackson, Z. Ke, C.J. Lanczycki, F. Lu, G.H. Marchler, M. Mullokandov, M.V. Omelchenko, C.L. Robertson, J.S. Song, N. Thanki, R.A. Yamashita, D. Zhang, N. Zhang, C. Zheng, S.H. Bryant, CDD: a Conserved Domain Database for the functional annotation of proteins, Nucleic Acids Res. 39 (2011) D225–D229.

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

[21] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, S. Genome, Project Data Processing, The Sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.

[22] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 15545–15550.

[23] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdottir, P. Tamayo, J.P. Mesirov, Molecular signatures database (MSigDB) 3.0, Bioinformatics 27 (2011) 1739–1740.

[24] P.J. Stephens, D.J. McBride, M.L. Lin, I. Varela, E.D. Pleasance, J.T. Simpson, L.A. Stebbings, C. Leroy, S. Edkins, L.J. Mudie, C.D. Greenman, M. Jia, C. Latimer, J.W. Teague, K.W. Lau, J. Burton, M.A. Quail, H. Swerdlow, C. Churcher, R. Natrajan, A.M. Sieuwerts, J.W. Martens, D.P. Silver, A. Langerod, H.E. Russnes, J.A. Foekens, J.S. Reis-Filho, L. van 't Veer, A.L. Richardson, A.L. Borresen-Dale, P.J. Campbell, P.A. Futreal, M.R. Stratton, Complex landscapes of somatic rearrangement in human breast cancer genomes, Nature 462 (2009) 1005–1010.

[25] H. Tanaka, M.C. Yao, Palindromic gene amplification–an evolutionarily conserved role for DNA inverted repeats in the genome, Nat. Rev. Cancer 9 (2009) 216–224.

[26] M.F. Berger, M.S. Lawrence, F. Demichelis, Y. Drier, K. Cibulskis, A.Y. Sivachenko, A. Sboner, R. Esgueva, D. Pflueger, C. Sougnez, R. Onofrio, S.L. Carter, K. Park, L. Habegger, L. Ambrogio, T. Fennell, M. Parkin, G. Saksena, D. Voet, A.H. Ramos, T.J. Pugh, J. Wilkinson, S. Fisher, W. Winckler, S. Mahan, K. Ardlie, J. Baldwin, J.W. Simons, N. Kitabayashi, T.Y. MacDonald, P.W. Kantoff, L. Chin, S.B. Gabriel, M.B. Gerstein, T.R. Golub, M. Meyerson, A. Tewari, E.S. Lander, G. Getz, M.A. Rubin, L.A. Garraway, The genomic complexity of primary human prostate cancer, Nature 470 (2011) 214–220.

[27] P.J. Stephens, C.D. Greenman, B. Fu, F. Yang, G.R. Bignell, L.J. Mudie, E.D. Pleasance, K.W. Lau, D. Beare, L.A. Stebbings, S. McLaren, M.L. Lin, D.J. McBride, I. Varela, S. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A.P. Butler, J.W. Teague, M.A. Quail, J. Burton, H. Swerdlow, N.P. Carter, L.A. Morsberger, C. Iacobuzio-Donahue, G.A. Follows, A.R. Green, A.M. Flanagan, M.R. Stratton, P.A. Futreal, P.J. Campbell, Massive genomic rearrangement acquired in a single catastrophic event during cancer development, Cell 144 (2011) 27–40.

[28] J.J. Molenaar, J. Koster, D.A. Zwijnenburg, P. van Sluis, L.J. Valentijn, I. van der Ploeg, M. Hamdi, J. van Nes, B.A. Westerman, J. van Arkel, M.E. Ebus, F. Haneveld, A. Lakeman, L. Schild, P. Molenaar, P. Stroeken, M.M. van Noesel, I. Ora, E.E. Santo, H.N. Caron, E.M. Westerhout, R. Versteeg, Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes, Nature 483 (2012) 589–593.

[29] T. Rausch, D.T. Jones, M. Zapatka, A.M. Stutz, T. Zichner, J. Weischenfeldt, N. Jager, M. Remke, D. Shih, P.A. Northcott, E. Pfaff, J. Tica, Q. Wang, L. Massimi, H. Witt, S. Bender, S. Pleier, H. Cin, C. Hawkins, C. Beck, A. von Deimling, V. Hans, B. Brors, R. Eils, W. Scheurlen, J. Blake, V. Benes, A.E. Kulozik, O. Witt, D. Martin, C. Zhang, R. Porat, D.M. Merino, J. Wasserman, N. Jabado, A. Fontebasso, L. Bullinger, F.G. Rucker, K. Dohner, H. Dohner, J. Koster, J.J. Molenaar, R. Versteeg, M. Kool, U. Tabori, D. Malkin, A. Korshunov, M.D. Taylor, P. Lichter, S.M. Pfister, J.O. Korbel, Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations, Cell 148 (2012) 59–71.

[30] A. Fujimoto, Y. Totoki, T. Abe, K.A. Boroevich, F. Hosoda, H.H. Nguyen, M. Aoki, N. Hosono, M. Kubo, F. Miya, Y. Arai, H. Takahashi, T. Shirakihara, M. Nagasaki, T. Shibuya, K. Nakano, K. Watanabe-Makino, H. Tanaka, H. Nakamura, J. Kusuda, H. Ojima, K. Shimada, T. Okusaka, M. Ueno, Y. Shigekawa, Y. Kawakami, K. Arihiro, H. Ohdan, K. Gotoh, O. Ishikawa, S. Ariizumi, M. Yamamoto, T. Yamada, K. Chayama, T. Kosuge, H. Yamaue, N. Kamatani, S. Miyano, H. Nakagama, Y. Nakamura, T. Tsunoda, T. Shibata, H. Nakagawa, Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators, Nat. Genet. 44 (2012) 760–764.

[31] R. Beroukhim, C.H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J.S. Boehm, J. Dobson, M. Urashima, K.T. Mc Henry, R.M. Pinchback, A.H. Ligon, Y.J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M.S. Lawrence, B.A. Weir, K.E. Tanaka, D.Y. Chiang, A.J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F.J. Kaye, H. Sasaki, J.E. Tepper, J.A. Fletcher, J. Tabernero, J. Baselga, M.S. Tsao, F. Demichelis, M.A. Rubin, P.A. Janne, M.J. Daly, C. Nucera, R.L. Levine, B.L. Ebert, S. Gabriel, A.K. Rustgi, C.R. Antonescu, M. Ladanyi, A. Letai, L.A. Garraway, M. Loda, D.G. Beer, L.D. True, A. Okamoto, S.L. Pomeroy, S. Singer, T.R. Golub, E.S. Lander, G. Getz, W.R. Sellers, M. Meyerson, The landscape of somatic copy-number alteration across human cancers, Nature 463 (2010) 899–905.

[32] H. Liu, A. Zilberstein, P. Pannier, F. Fleche, C. Arendt, C. Lengauer, C.S. Hahn, Evaluating translocation gene fusions by SNP array data, Cancer Informat. 11 (2012) 15–27.

[33] A. Abyzov, A.E. Urban, M. Snyder, M. Gerstein, CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing, Genome Res. 21 (2011) 974–984.

[34] X.S. Wang, J.R. Prensner, G. Chen, Q. Cao, B. Han, S.M. Dhanasekaran, R. Ponnala, X. Cao, S. Varambally, D.G. Thomas, T.J. Giordano, D.G. Beer, N. Palanisamy, M.A. Sartor, G.S. Omenn, A.M. Chinnaiyan, An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer, Nat. Biotechnol. 27 (2009) 1005–1011.

[35] E.T. Sawey, M. Chanrion, C. Cai, G. Wu, J. Zhang, L. Zender, A. Zhao, R.W. Busuttil, H. Yee, L. Stein, D.M. French, R.S. Finn, S.W. Lowe, S. Powers, Identification of a therapeutic strategy targeting amplified FGF19 in liver cancer by Oncogenomic screening, Cancer Cell 19 (2011) 347–358.

[36] P. Filippakopoulos, J. Qi, S. Picaud, Y. Shen, W.B. Smith, O. Fedorov, E.M. Morse, T. Keates, T.T. Hickman, I. Felletar, M. Philpott, S. Munro, M.R. McKeown, Y. Wang, A.L. Christie, N. West, M.J. Cameron, B. Schwartz, T.D. Heightman, N. La Thangue, C.A. French, O. Wiest, A.L. Kung, S. Knapp, J.E. Bradner, Selective inhibition of BET bromodomains, Nature 468 (2010) 1067–1073.

[37] C. Kumar-Sinha, S.A. Tomlins, A.M. Chinnaiyan, Recurrent gene fusions in prostate cancer, Nat. Rev. Cancer 8 (2008) 497–511.

[38] S. Kervestin, A. Jacobson, NMD: a multifaceted response to premature translational termination, Nat. Rev. Mol. Cell Biol. 13 (2012) 700–712.

[39] A.S. Knisely, S.S. Strautnieks, Y. Meier, B. Stieger, J.A. Byrne, B.C. Portmann, L.N. Bull, L. Pawlikowska, B. Bilezikci, F. Ozcay, A. Laszlo, L. Tiszlavicz, L. Moore, J. Raftos, H. Arnell, B. Fischler, A. Nemeth, N. Papadogiannakis, J. Cielecka-Kuszyk, I. Jankowska, J. Pawlowska, H. Melin-Aldana, K.M. Emerick, P.F. Whitington, G. Mieli-Vergani, R.J. Thompson, Hepatocellular carcinoma in ten children under five years of age with bile salt export pump deficiency, Hepatology 44 (2006) 478–486.

[40] G. Zheng, M. Marino, J. Zhao, R.T. McCluskey, Megalin (gp330): a putative endocytic receptor for thyroglobulin (Tg), Endocrinology 139 (1998) 1462–1465.

[41] T.M. Chlon, D.A. Taffany, J. Welsh, M.J. Rowling, Retinoids modulate expression of the endocytic partners megalin, cubilin, and disabled-2 and uptake of vitamin D-binding protein in human mammary cells, J. Nutr. 138 (2008) 1323–1328.

[42] H.Y. Xue, H.L. Wong, Targeting megalin to enhance delivery of anti-clusterin small-interfering RNA nanomedicine to chemo-treated breast cancer, Eur. J. Pharm. Biopharm. 81 (2012) 24–32.

[43] M.O. Pedersen, P.B. Hansen, S.L. Nielsen, M. Penkowa, Metallothionein-I + II and receptor megalin are altered in relation to oxidative stress in cerebral lymphomas, Leuk. Lymphoma 51 (2010) 314–328.

[44] H. Ammar, J.L. Closset, Clusterin activates survival through the phosphatidylinositol 3-kinase/Akt pathway, J. Biol. Chem. 283 (2008) 12851–12861.

[45] N. Cancer Genome Atlas, Comprehensive molecular characterization of human colon and rectal cancer, Nature 487 (2012) 330–337.

[46] N. Cancer Genome Atlas Research, P.S. Hammerman, D.N. Hayes, M.D. Wilkerson, N. Schultz, R. Bose, A. Chu, E.A. Collisson, L. Cope, C.J. Creighton, G. Getz, J.G. Herman, B.E. Johnson, R. Kucherlapati, M. Ladanyi, C.A. Maher, G. Robertson, C. Sander, R. Shen, R. Sinha, A. Sivachenko, R.K. Thomas, W.D. Travis, M.S. Tsao, J.N. Weinstein, D.A. Wigle, S.B. Baylin, R. Govindan, M. Meyerson, Comprehensive genomic characterization of squamous cell lung cancers, Nature 489 (2012) 519–525.

[47] Y. Gui, G. Guo, Y. Huang, X. Hu, A. Tang, S. Gao, R. Wu, C. Chen, X. Li, L. Zhou, M. He, Z. Li, X. Sun, W. Jia, J. Chen, S. Yang, F. Zhou, X. Zhao, S. Wan, R. Ye, C. Liang, Z. Liu, P. Huang, C. Liu, H. Jiang, Y. Wang, H. Zheng, L. Sun, X. Liu, Z. Jiang, D. Feng, J. Chen, S. Wu, J. Zou, Z. Zhang, R. Yang, J. Zhao, C. Xu, W. Yin, Z. Guan, J. Ye, H. Zhang, J. Li, K. Kristiansen, M.L. Nickerson, D. Theodorescu, Y. Li, X. Zhang, S. Li, J. Wang, H. Yang, J. Wang, Z. Cai, Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder, Nat. Genet. 43 (2011) 875–878.

[48] B.W. Ferguson, S. Datta, Role of heparan sulfate 2-o-sulfotransferase in prostate cancer cell proliferation, invasion, and growth factor signaling, Prostate Cancer 2011 (2011) 893208.

[49] K. Song, Q. Li, Y.B. Peng, J. Li, K. Ding, L.J. Chen, C.H. Shao, L.J. Zhang, P. Li, Silencing of hHS6ST2 inhibits progression of pancreatic cancer through inhibition of Notch signalling, Biochem. J. 436 (2011) 271–282.

[50] K.A. Wharton Jr., G. Zimmermann, R. Rousset, M.P. Scott, Vertebrate proteins related to Drosophila Naked Cuticle bind Dishevelled and antagonize Wnt signaling, Dev. Biol. 234 (2001) 93–106.

[51] X.L. Jia, S.Y. Li, S.S. Dang, Y.A. Cheng, X. Zhang, W.J. Wang, C.E. Hughes, B. Caterson, Increased expression of chondroitin sulphate proteoglycans in rat hepatocellular carcinoma tissues, World J. Gastroenterol. 18 (2012) 3962–3976.

[52] C. Jang, J.K. Choi, Y.J. Na, B. Jang, W. Wasco, J.D. Buxbaum, Y.S. Kim, E.K. Choi, Calsenilin regulates presenilin 1/gamma-secretase-mediated N-cadherin epsilon-cleavage and beta-catenin signaling, FASEB J. 25 (2011) 4174–4183.

[53] R.S. Mani, A.M. Chinnaiyan, Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences, Nat. Rev. Genet. 11 (2010) 819–829.

[54] W. Pao, K.E. Hutchinson, Chipping away at the lung cancer genome, Nat. Med. 18 (2012) 349–351.