doi: 10.1093/nar/gkaa1119

The landscape of long noncoding RNA-involved and tumor-specific fusions across various cancers

Mengbiao Guo[1,†], Zhen-Dong Xiao[2,†], Zhiming Dai[3,†], Ling Zhu[1,†], Hang Lei[2], Li-Ting Diao[2] and Yuanyan Xiong[1,*]

[1]Key Laboratory of Gene Engineering of the Ministry of Education, Institute of Healthy Aging Research, School of Life Sciences, Sun Yat-sen University, Guangzhou 510006, China, [2]The Biotherapy Center, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510630, China and [3]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China

Received May 26, 2020; Revised October 15, 2020; Editorial Decision October 27, 2020; Accepted November 03, 2020

ABSTRACT

The majority of the human genome encodes long noncoding RNA (lncRNA) genes, critical regulators of various cellular processes, which largely outnumber protein-coding genes. However, lncRNA-involved fusions have not been surveyed and characterized yet. Here, we present a systematic study of the lncRNA fusion landscape across cancer types and identify >30 000 high-confidence tumor-specific lncRNA fusions (using 8284 tumor and 6946 normal samples). Fusions positively correlated with DNA damage and cancer stemness and were specifically low in microsatellite instable (MSI)-High or virus-infected tumors. Moreover, fusions distribute differently among cancer molecular subtypes, but with shared enrichment in tumors that are microsatellite stable (MSS), with high somatic copy number alterations (SCNA), and with poor survival. Importantly, we find a potentially new mechanism, mediated by enhancer RNAs (eRNA), which generates secondary fusions that form densely connected fusion networks with many fusion hubs targeted by FDA-approved drugs. Finally, we experimentally validate functions of two tumor-promoting chimeric proteins derived from mRNA-lncRNA fusions, KDM4B–G039927 and EPS15L1–lncOR7C2–1. The EPS15L1 fusion protein may regulate (Gasdermin E) GSDME, critical in pyroptosis and anti-tumor immunity. Our study completes the fusion landscape in cancers, sheds light on fusion mechanisms, and enriches lncRNA functions in tumorigenesis and cancer progression.

INTRODUCTION

Fusions are products mostly resulting from DNA structural changes (1,2). Many fusions can be used as biomarkers and therapeutic targets in various cancers (3,4), such as BCR–ABL1 in leukemia, EWSR1–FLI1 in Ewing's sarcoma and TMPRSS2–ERG in prostate cancer. Thanks to the huge number of samples with RNA sequencing (RNA-seq) available from projects including The Cancer Genome Atlas (TCGA) (5) and Genotype-Tissue Expression (GTEx) (6), fusion events have been studied extensively in both cancer (1,7–11) and normal tissues (12). Integrative analysis of fusions with kinase fusions, druggability and other driver mutations was performed (10). A database of systematic functional annotations of fusions was constructed (9). The fusion knowledge database based on text-mining and manual curation has been updated recently (8).

However, all these efforts have been focused on mRNA–mRNA fusions, and most studies dedicated to either identification or collection of fusion events. Few fusions were reported for lncRNA, especially enhancer RNAs (eRNA), which largely outnumber the protein-coding genes and play important roles in cellular processes during development and diseases (13,14). Numerous efforts strived to annotate lncRNA functions (15) and study lncRNA functional mechanisms (16). Adding lncRNA-involved fusions would provide new insights into the whole picture of fusions, especially in cancers. Moreover, fusion connections with cancer molecular subtypes have not been reported before, which may have considerable contributions to our understanding of the underlying mechanisms of fusion generation and tumorigenesis.

Here we present the first atlas of lncRNA fusions, together with mRNA–mRNA fusions, across cancers. The validation rate of our lncRNA fusions was estimated to be as high as ∼80%. We study this fusion landscape in novel angles and provide interesting insights into fusions in cancer, including fusion connections with MSI, virus in-

[*]To whom correspondence should be addressed. Tel: +86 20 39943531; Fax: +86 20 39943778; Email: xyyan@mail.sysu.edu.cn
[†]The authors wish it to be known that, in their opinion, the first four authors should be regarded as Joint First Authors.

fection, molecular subtypes, SCNA, cancer stemness and patient survival. More importantly, we report a subset of secondary fusions mediated by primary eRNA fusions and experimentally validated the tumor-promoting functions of two mRNA-lncRNA fusions, KDM4B–G039927 and EPS15L1–lncOR7C2–1.

MATERIALS AND METHODS

Data collection

TCGA RNA-seq raw reads and associated clinical information were downloaded from the GDC portal (https://portal.gdc.cancer.gov/). GTEx RNA-seq raw reads and sample information were downloaded from the dbGaP database (phs000424.v6.p1). Virus infection status for cancers was obtained from (17). Tumor MSI genotyping information was obtained from (18). Cancer subtypes, stemness, DNA damage scores and immune signature classifications were downloaded from UCSC Xena (https://xenabrowser.net/). CPTAC-BRCA proteomics and phosphoproteomics datasets were downloaded from http://linkedomics.org. Interactions of eRNA-target and eRNA-drug were obtained from (19). Hi-C-based enhancer–promoter interactions were from (20). CRISPR-based enhancer-target were from (21). RNA-seq reads (SRR8615767) for MDA-MB-231 cells were downloaded from Gene Expression Omnibus (GEO). Our SKBR3 RNA-seq reads were deposited on GEO (GSE157986).

Build an integrative set of lncRNA annotations

GTF annotations of protein-coding and lncRNA genes for GENCODE v28 were downloaded from www.gencodegenes.org. Three more additional comprehensive lncRNA annotations were further obtained, including MiTranscriptome v2 (22), NONCODE v5 (23) and LNCipedia v5.2 (24), all as GTF files. All genomic coordinates were converted to hg38. Then, we merged the four GTF files and removed duplicated gene annotations using gffcompare (https://ccb.jhu.edu/software/stringtie/gffcompare.shtml). Specifically, any transcript annotated as 'protein_coding' in the MiTranscriptome database was removed. A final set of 251 692 genes including protein-coding ($n = 19\ 889$) and ncRNAs ($n = 231\ 803$, mostly lncRNAs) were used for genome indexing and then fusion identification (indexing this huge set of lncRNA annotations for STAR-Fusions was very time-consuming, taking several weeks on the server). We deliberately included as many lncRNA annotations as possible (more than those in the LncBook database (25)), which is not published before this project started) for fusion algorithms to better detect fusion events, and we applied strict filtering steps to clean the fusion results.

Identify and clean fusion events

Most computing work of fusion identification was performed on the Tianhe-2 supercomputer, supported by National Supercomputer Center in Guangzhou. Two highly reliable fusion algorithms (26), Arriba (https://github.com/suhrig/arriba, default parameters) and STAR-Fusion (26) (–FFPM 0.1, also used by (10)) were used to call fusion events. Both algorithms were applied to each sample from TCGA and GTEx. Then, fusion results from both algorithms were merged (using only fusions with 'high' confidence from Arriba and with FFPM ≥ 0.5 and 'YES_LDAS' from STAR-Fusion) separately for cancer and normal samples. When both algorithms identified the same fusion in a sample, the fusion from Arriba was retained. Next, cancer fusions overlapping any fusion found in normal samples (TCGA tumor matched normal or all GTEx normal samples) were removed. Then, we further required support of both types of fusion reads: split-reads ≥ 1 and spanning-reads ≥ 1. Moreover, although duplicated annotations were removed, there were still some transcripts that may share the same fusion breakpoint for one fusion event, and only the event with the most supporting reads was kept. The remaining fusions were used for downstream analysis.

NFPT calculation

For each cancer, NFPT was calculated as the total fusions divided by the total samples (normalization). This was done similarly for each cancer subtype and cancers with or without virus infection.

Analysis of MSI with fusions

For each MSI-prone cancer, COAD, ESCA and STAD, MSS and MSI-Low were grouped together as 'others' and compared with MSI-High, because we found MSS and MSI-Low samples did not differ in fusion abundances. Fusions (log-transformed) for 'MSI-High' and 'others' samples were compared by two-sided wilcox.test in R.

Survival analysis

For each survival analysis, a log-rank test by TCGAbiolinks (27) was used to compute the significance, and Cox multivariate regression was further used to confirm the signal and compute the hazard ratio, after adjusting for confounding variables including sex, age, stage, ploidy and tumor purity. In survival analysis of KDM4B and EPS15L1 fusions, both the clean (high-confidence only) and raw (including fusions with other confidence levels) fusions were examined, and raw fusions were found useful and improved the significance of survival difference.

Analysis of DNA damage score, cancer stemness and immune signatures with fusions

For each sample, fusions (log-transformed) were correlated with DNA damage score or cancer stemness using Pearson correlation and P-value was calculated by the cor.test function in R. For immune signatures, fusion numbers were normalized by the number of samples classified as each signature for each cancer, similar to NFPT calculation.

Analysis of virus infection with fusions

For each virus-associated cancer, fusions (log-transformed) for virus-positive and virus-negative samples were compared by one-sided t-test in R.

FGI network construction and analysis

LncRNA breakpoints from FPL fusions were used to examine potential overlap with eRNAs that had potential interaction with long-range target genes (19) and with enhancer elements that had physical contact with long-range target genes determined by Hi-C (20) or CRISPR-based (21) technologies. Then, the protein gene from FPL fusions and target genes of the eRNA from the same tissue were connected together and formed the FGI network. Both the clean and raw tumor fusions were used to check for detection of predicted secondary fusions in this FGI network, and raw fusions were found useful to enhance the support for secondary fusions, compared with only using high-confidence fusions. Cytoscape (28) was used for network analysis and visualization. For iMARGI (29) data re-analysis, we downloaded processed data files in BEDPE format containing interacting regions from GEO (GSE122690). The nearest gene was assigned to each region. Secondary FPP gene pairs were then compared with iMARGI gene pairs to calculate the overlap.

Drug target and drug responses analysis

Drugs targeting FGI hub genes were obtained from DGIdb (30) with information of FDA approval. For eRNA associated drug responses, genomic breakpoints of eRNA fusions were compared with eRNA regions from (19). Breakpoints located in eRNAs with drug associations were regarded as affecting eRNA–drug associations.

Proteomic and phosphoproteomic analysis

Identification of novel peptides from KDM4Bf and EPS15L1f was performed by PepQuery (31). Differential protein and phosphorylation levels were analyzed by R package limma, adjusted for confounding variables, including sex, age, stage, tumor purity and ploidy.

Plasmids, cell lines and cell growth assay

cDNA clones of KDM4B and EPS15L1 were purchased from JuJiao Biotehnology. Flag-tagged full and fusion ORFs of KDM4B and EPS15L1 were subcloned to PLVX-puro lentivirus vectors by Gibson Assemble. Primers for cloning were listed in Supplementary Table S8. All breast cancer cell lines were purchased from Cell Bank of Chinese Academy of Sciences at Shanghai. Cells were cultured in Dulbecco's Modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and maintained in CO_2 incubators (Thermo Scientific) at 37°C, 5% $CO_{2.}$ Lentiviruses were produced in HEK293T cells with the viral packaging constructs psPAX2 and pMD2G (Addgene). Stable cell lines were generated by infecting the cells with corresponding virus and selected by puromycin for 1 week. To measure cell growth, stable cells were seed in 12-well plated and were cultured for indicating days. Culture medium were refreshed every 2 days. Cells were stained with 0.1% Crystal Violet (Sigma) for 15 min at room temperature. Stained Crystal Violet was then extracted with 10% acetic acid. The intensity of the color was measured by a photospectrometry at OD570.

Quantitative real-time PCR

Total RNA was extracted from cells using Trizol (Invitrogen) and cDNA was prepared using Reverse Transcriptase Kit (Vazyme). Real-time PCR was performed using SYBR Green PCR kit (Vazyme), and was run on Roche LightCycler 480. Expression levels were normalized by GAPDH. Primers for real-time PCR were listed in Supplementary Table S8.

Western blot analysis

Cultured cells were lysed with RIPA buffer containing complete mini protease and phosphatase inhibitors (Roche). Western blots were obtained utilizing 20–40 μg of lysate protein. The following antibodies were used in this study: Vinculin (Sigma, V9264, 1:5000 dilution) and Flag M2 (Sigma, F1804, 1:2000 dilution).

Fusion validation

Total RNA of SKBR3 and MDA-MB-231 were extracted using TRIzol according to the manufacturer's instructions (Invitrogen, Thermo Scientific, U.S.A.). One microgram total RNA was used for cDNA synthesis. PCR products were gel purification and were cloned into T vectors by TA cloning. Insertions were sequenced by Sanger sequencing. Primers used for validation were provided in Supplementary Table S9. For RNA Sequencing of SKBR3 cells, RNA was extracted with TRIzol and sequenced following the manufacturer's protocol. RNA-seq data were processed as the TCGA RNA-seq data.

RESULTS

Widespread lncRNA fusions across cancer types

We obtained 48 545 high-confidence tumor-specific fusions (Figure 1A and Supplementary Table S1) after strict filtering (see 'Materials and Methods' section) of the initial 1 867 911 fusions in 8284 cancer samples (30 cancer types) from TCGA and 6946 normal samples from TCGA (n = 585) and GTEx (n = 6361) (Supplementary Table S2). The initial fusions were identified by combining three large lncRNA annotation databases MiTranscriptome (22), NONCODE (23) and LNCipedia (24)) with GENCODE (www.gencodegenes.org) as reference and applying two efficient fusion calling algorithms (Arriba and STAR-Fusion) (26) (see 'Materials and Methods' section). Most fusions were predicted by Arriba to originate from duplications or translocations (Figure 1B).

Our strategy enabled comprehensive identification of lncRNA-involved fusions across cancer types. Among all these tumor fusions, 14 673 (30.2%), 20 816 (42.9%), 5230 (10.8%) and 7826 (16.1%) events were mRNA–mRNA (FPP), mRNA–lncRNA (FPL), lncRNA–mRNA (FLP) and lncRNA–lncRNA (FLL) fusions, respectively (Figure 1C). In previous studies, lncRNA-involved fusions in TCGA only comprised ∼1.0% (Hu2018 (7)) or 4.3% (Gao2018, (10)) (Figure 1D). About half (56.9%) of our mRNA-mRNA fusions were previously reported in TCGA (7,10). We identified 72.8% of fusions from the Gao2018

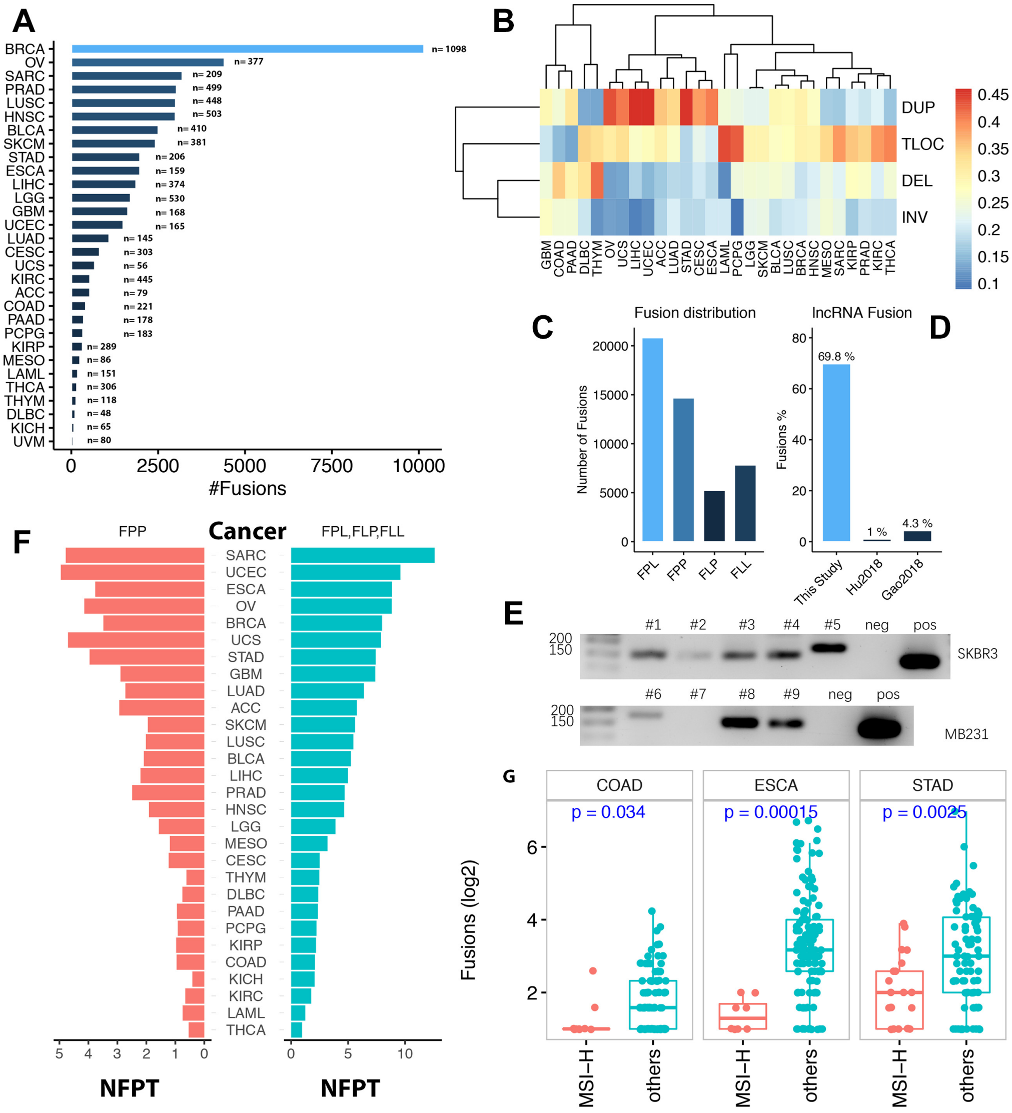

Figure 1. Characteristics of high-confidence fusions and the relationship between fusion and MSI. (**A**) Fusion distribution across cancer types. BRCA has the highest fusions partly because it has the largest number of tumor samples. Numbers for each bar represent numbers of tumor samples used in this study. (**B**) Fusions are predicted as results of different structural variations, mostly duplications (DUP) and translocations (TLOC), and less from deletions (DEL) and inversions (INV). Color scale stands for fusion fraction for each category. (**C**) Comparison of mRNA–mRNA fusions (FPP) and lncRNA-involved fusions (FPL, FLP, FLL). Most fusions are from FPL, followed by FPP. (**D**) Comparison of percentages of lncRNA fusions in three studies of TCGA cancer samples. (**E**) RT-PCR detection of selected fusions in SKBR3 and MB-231 cells. Neg and pos represent negative control and positive control, respectively. (#1, G082937–G082922, #2, G082937–NONHSAG051248; #3, RF02164–PCAT1; #4, ARHGAP10–NONHSAG112173; #5, G082937–POU5F1B; #6, NONHSAG073428–MICAL1; #7, G038808–G045118; #8, FOCAD–G084147; #9, FOCAD–NONHSAG114490). (**F**) Similar distribution across cancers of FPP and lncRNA fusions (FPL, FLP, FLL). MSI-prone cancers (UCEC, ESCA, OV, STAD) rank at the top. (**G**) MSI-High tumors (COAD, ESCA, and STAD) have significantly less fusions, consistently, than other tumor subtypes; two-sided Wilcoxon test.

study and the difference was probably mostly because of about 8% of tumor samples were not included here due to download issues or failed the fusion pipeline, or because of a much larger set of reference annotation were used. Interestingly, FPLs was much higher than FLPs, which is probably due to FPLs used protein-coding promoters and FLPs used lncRNA promoters, because it is known that generally lncRNAs had much lower expression than protein-coding genes. To support this hypothesis, we further showed that FPLs generally had more supporting reads than FLPs (Supplementary Figure S1).

Highly reliable lncRNA fusions

Some lncRNAs may overlap with protein-coding genes, which may result in false lncRNA fusions. We investigated the breakpoints distribution of lncRNA fusions and found that almost all lncRNA fusion breakpoints were outside of protein-coding genes, with a mean and median distance of 230 and 69 kb, respectively, to the nearest protein-coding genes (Supplementary Figure S2).

To further demonstrate the reliability of our lncRNA fusions, we deep-sequenced our SKBR3 cells using RNA-seq and also downloaded public RNA-seq data for MDA-MB-

231 cells. We then applied the same fusion detection process and selected nine of twelve lncRNA fusions (Supplementary Table S3, five from SKBR3 and four from MDA-MB-231) from these two cell lines for experiment using RT-PCR and subsequently Sanger sequencing. Expectedly, we successfully validated seven (∼80%) out of the nine fusions (3/3 FPLs, 1/2 FLPs, and 3/4 FLLs) (Figure 1E and Supplementary Figure S3). Therefore, our lncRNA fusions should be reliable and complement the previous protein-coding gene dominated fusion landscape.

Fusions were negatively associated with microsatellite instability

We found the distribution of FLPs per tumor was very similar to FPP distribution (Figure 1F), and microsatellite instable (MSI) cancers (for example, UCEC, ESCA, STAD and OV), rather than less MSI-prone cancers THCA and KIRC (18), seemed to have larger number of fusions per tumor (NFPT), indicating some connections between fusion and MSI. We then investigated the connections in COAD, ESCA and STAD, which have enough samples with both RNA-seq and TCGA MSI genotypes. Surprisingly, MSI-High tumors had the least fusions among these three cancers (Figure 1G), which may be explained by the lower demand of other oncogenic events, such as fusions, in MSI-High tumors. Further work is needed to uncover the reason why MSI-High COAD tumors, displaying frequent kinase fusions (32), had lower fusions than STAD and ESCA. Of note, we found that most fusions in UCEC, STAD, ESCA, and OV originated from duplications, while most fusions from THCA and KIRC, which had the lowest NFPTs, were from translocations (Figure 1C).

Interestingly, SARC showed the highest NFPTs (Figure 1F) and had connections with MSI, which was detected in a maximum of 44% sarcoma patients, although with controversies (33). It was reported 85–90% Ewing sarcomas had the EWS/FLI fusion (34), a subset of whose target genes harbor a microsatellite response element in their promoters (33). Meanwhile, the number of kinase fusions in SARC was among the lowest (10), in contrast to THCA and KIRP showing the highest kinase fusions (10) but with the lowest NFPTs (Figure 1F) and the fewest MSI loci (18).

Fusions were enriched in specific cancer subtypes with high copy number alterations and with poor prognosis

Fusions are very cancer-specific (10). Whether fusion distribution in subtypes within each cancer differs has not been explored. Inspired by the observed fusion characteristics in MSI cancers, we computed the NFPTs across subtypes for each cancer. Surprisingly, the subtype fusion profiles showed mostly cancer subtype-specific but also some subtype-shared characteristics (Figure 2). A summary of the connections between fusions and cancer subtypes was presented in Table 1.

In HNSC, classical tumors had the highest NFPTs, which showed the worst survival rate. These tumors were HPV-negative with highest copy number of EGFR, PIK3CA and TP63 (35). In SKCM, the Triple_WT subtype had much higher NFPTs than other subtypes. It is a heterogeneous subtype characterized by high somatic copy number alterations (SCNA) and complex structural rearrangements (36), consistent with our fusion results. In UCEC, the CN_High subtype showed extremely high NFPTs compared to other subtypes. It is microsatellite stable (MSS), with the highest TP53 mutation rate and with the worst survival rate (37).

In SARC, subtypes iCluster:2 and iCluster:3 showed very high NFPTs, possibly because they comprised of a large fraction of dedifferentiated liposarcoma samples—almost 100% for iCluster:2 and 50% for iCluster:3—that displayed the highest level of SCNAs across all TCGA cancer types (38). Interestingly, iCluster:4 had extremely low NFPTs, which may due to this subtype basically only need one fusion (SS18–SSX) that would disrupt epigenetic regulation (38).

Interestingly, brain tumors, LGG and GBM, showed very similar subtype NFPTs (Figure 2, red circles). G-CIMP-low tumors showed extremely high NFPTs than other subtypes and was characterized by high-frequency mutations and activation of cell cycle genes, CDKN2A/B, CCND2, CDK4 and RB1 (39).

Furthermore, digestive system cancers, including COAD, ESCA and STAD, also showed highly similar NFPT distribution (Figure 2, orange circles), highest in chromosome instable (CIN) tumors. Most CIN samples were MSS (40), which further corroborated our conclusion of the connection between MSS tumors and their fusions. Moreover, CIN samples displayed more focal SCNAs, especially in the upper gastrointestinal tract (meaning NFPTs in ESCA and STAD are much higher than in COAD), which is consistent with the fusion characteristics of the CN_High subtype from various other cancer types. This positive correlation may be resulted from the abundant extrachromosomal DNAs (ecDNA) in cancer (41). Of note, CIN samples have been associated with worse prognosis (42) and stemness (43,44). Future work is needed to disentangle the relationship between NFPT and CIN.

As indicated in the subtype analysis, fusion events may be associated with DNA methylation, therefore, we examined the differences of fusion abundance across methylation subtypes for each cancer. Among the nine tested cancer types with enough samples for both fusions and methylation (BRCA, ESCA, KIRP, PAAD, PRAD, SARC, STAD, THCA, UCS), we found five (>50%) with P values <0.05 (Kruskal–Wallis test), with extremely high significance in BRCA (P = 1.1E-12) and SARC (P = 7.5E-10) (Supplementary Figure S4). Furthermore, we observed a general trend of better survival rate for patients with lower fusion events (although some did not reach significance, Supplementary Figure S5), with the most significant difference in LGG (Figure 3A). This is also consistent with the subtype analysis.

Fusions were positively associated with DNA damage and tumor stemness

Fusions are mostly products of DNA damage. We correlated our fusions with previously reported DNA damage

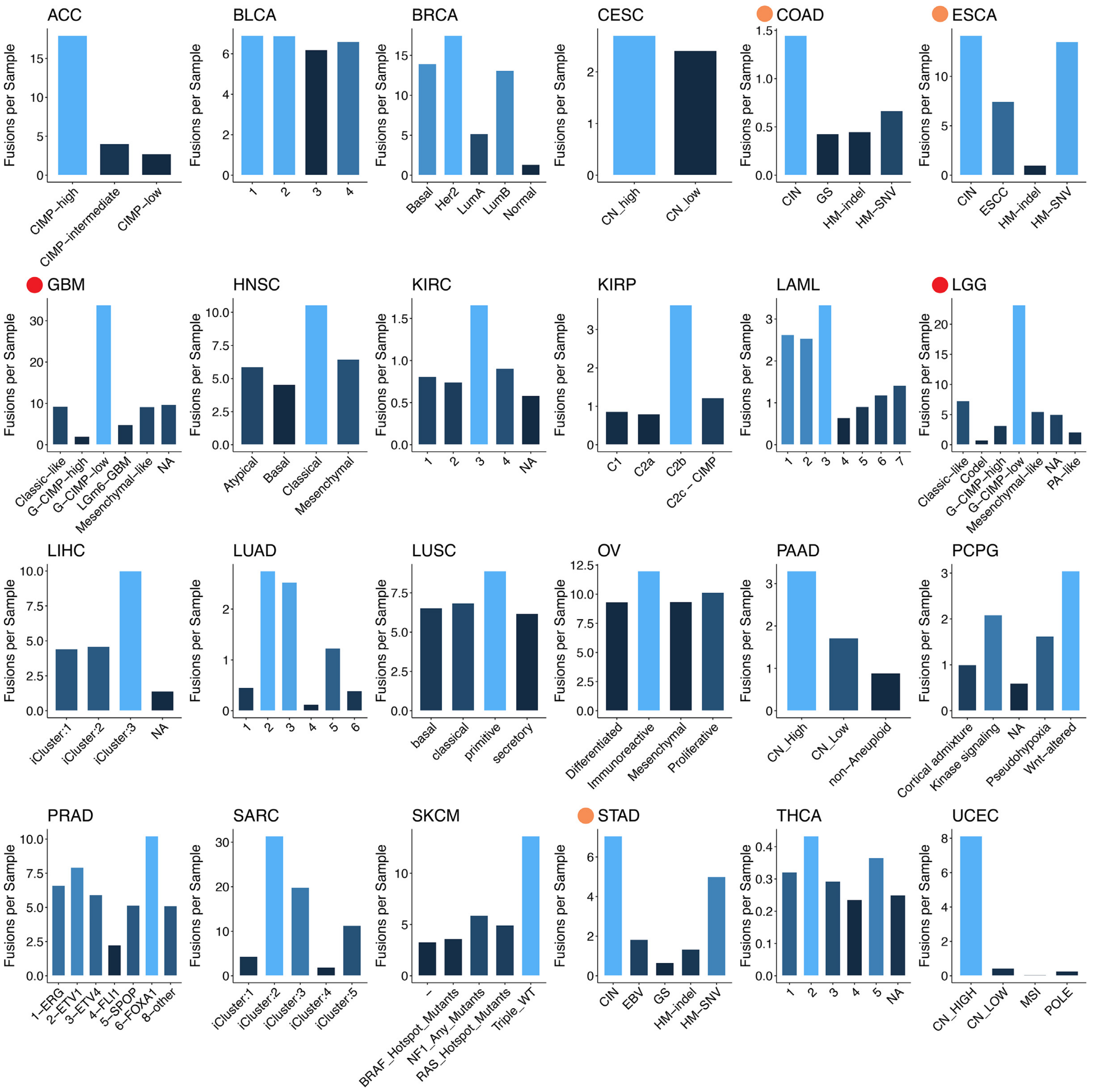

Figure 2. Fusion distribution across subtypes for each cancer. *X*-axis shows cancer subtypes and *Y*-axis is number of fusion per tumor (NFPT). Most cancers show subtype-specific fusion enrichment. Chromosome instable (CIN) and CN (somatic copy number)-High tumors show high NFPTs. Brain tumors (GBM and LGG) are marked by filled circles (top left). Digestive system cancers (COAD, ESCA, STAD) are also marked by filled circles. NA, not available. Color scales represent NFPT values (*Y*-axis) to create better visual effect. A detailed summary is presented in Table 1.

scores in TCGA (45). Expectedly, most cancers showed significant positive correlation between fusion events and DNA damage (Figure 3B), with the strongest signal from BRCA (Pearson $R = 0.44$, $P < 2.2E-16$, Figure 3C). We further examined the relationship between fusions and cancer stemness (46), which was associated with damage mutations of TP53 in some cancers. Although some signals were weak, most cancer types showed positive correlation between fusion abundance and stemness (Figure 3C, only significant signals were shown). The strongest signal was also in BRCA for mRNA-based stemness (Pearson $R = 0.33$, $P < 2.2E-16$, Figure 3D). Furthermore, we found the fewest fusions per sample in tumors classified as Inflammatory (C3) (Figure 3E), which shows the best survival rate among all six immune groups (C5 and C6 groups were excluded in Figure 3E due to small sample sizes) of patients (47).

Table 1. A summary of the associations between fusion frequencies and cancer subtypes

Cancer	Subtype	MSS	SCNA-high	CIN	TP53	PTEN	CDKN2A	Survival
ACC	CIMP-high							Poor
BRCA	Her2							Poor
COAD	CIN	Y		Y				Poor
ESCA	CIN	Y		Y				Poor
STAD	CIN	Y		Y				Poor
GBM	G-CIMP-low						Y	
LGG	G-CIMP-low						Y	
HNSC	Classical		Y		Y			Poor
KIRC	3					Y	Y	
KIRP	C2b						Y	Poor
LIHC	iCluster:3				Y	Y	Y	
LUAD	2,3		Y		Y			
LUSC	primitive			Y		Y		
PAAD	CN-high		Y					
PCPG	Wnt-altered							Poor
SARC	iCluster:2		Y					
SKCM	Triple-WT		Y	Y				
UCEC	CN-high	Y	Y		Y			Poor

MSS, microsatellite stable; SCNA, somatic copy number alteration; CIN, chromosome instable; Y means Yes.

Fusions were negatively associated with virus infection

A considerable number of cancers were induced by virus infection (17), which may cause gene fusions directly by genome integration or indirectly by disrupting genome stability. We investigated the fusion difference of samples with or without virus infection. Surprisingly, four viruses were found associated with decreased fusion events and one with increased fusions (Figure 4A–E). The association was significant for HBV (hepatitis virus B, $P = 0.012$, one-sided t-test) in LIHC and HHV5 (human herpesvirus 5, $P = 0.033$, one-sided t-test) in COAD, and the negative trend was clear for HPV16 (human papillomavirus 16) in HNSC, HHV4 (also known as Epstein–Barr virus, or EBV) in STAD, and HPV45 in CESC. Therefore, cancers with virus infection, except for HHV5, seems to have less fusions, which is similar to MSI-High cancers—both types of cancers probably require much fewer fusion events to induce tumor (48).

HBV was an important risk factor of liver cancer (49). We found all six DNAJB1–PRKACA fusions exclusively in HBV-negative liver cancer samples (Figure 4F). The fusion product showed that both the DnaJ domain and the kinase domain were intact (Figure 4G). Although DNAJB1 was able to target HBX to inhibit HBV replication (50), the DNAJB1–PRKACA fusion would cause a rare but lethal cancer called fibrolamellar hepatocellular carcinoma (FL-HCC) in adolescents (51). In addition, all six CPS1–CPS1 fusions found only in HBV-negative samples (Figure 4F) possibly disrupted the urea cycle of metabolism in liver and caused liver cancer (49).

HPV was known to cause cancer (52). We found all 14 (five with lncRNAs) and 6 (three with lncRNAs) FAT1 fusions only in HPV-negative samples (Figure 4F) of HNSC and CESC, respectively. FAT1, together with TP53, CDKN2A, and EGFR, was frequently mutated in HPV-negative tumors but rarely in HPV-positive tumors (48). Similarly, six out of seven (five with lncRNAs) EGFR fusions were found in HPV-negative HNSC samples (Figure 4F), and, interestingly, the remaining one (TCGA-CR-7368–01A) was previously classified as 'Atypical' (35). Of note, this is consistent with our cancer subtype analysis.

A novel fusion mechanism mediated by eRNAs created complex fusion networks

Many lncRNAs, including enhancer RNAs (eRNA), function as scaffolding molecules, guiding various factors to their target positions to control gene expression (14,19). We propose that FPL may wrongly connect its protein-coding partner (FG1) with the eRNA partner targeted genes (FG2), forming oncogenic fusion RNA-gene interactions (FGI, FG1–eRNA–FG2, Supplementary Figure S6). We collected expression-based eRNA-gene interactions (19) and Hi-C interaction-based (20) and CRISPR (clustered regularly interspaced short palindromic repeats)-based (21) experimentally verified physical enhancer-promoter interactions, and imposed our FPL fusions on them to construct FPL-induced FGI networks (3,122 interactions, Figure 5, Supplementary Table S4, see 'Materials and Methods' section). Surprisingly, we found about 13% (P-value < 2E-16, compared to random guess, chi-squared test) of these predicted FGIs (FG1–FG2) were directly detected by fusion calling algorithms using RNA-seq data. It is possible that many FGIs were not detected by fusion tools because of their lower expression level than their parental fusions.

These secondary fusion events were possibly generated by a new mechanism involving eRNA-promoter long-range interaction, well-known for trans-activation (20) and RNA–DNA interaction, which can create fusion RNAs without DNA structural changes or rearrangements, demonstrated by the iMARGI assay (29). To support our finding of FGI, we re-analyzed the iMARGI assay generated interaction data. We found ∼12.9% (P-value < 2E-16, compared to random guess, chi-squared test) of our FGI fusions had physical RNA–DNA interaction determined by the iMARGI assay. Therefore, both RNA-seq and iMARGI assay provided support for our secondary fusions (in total, about 22.3% FGIs were supported).

Since eRNA (lncRNA) can target a large number of genes, this may be a potentially highly efficient way to create fusion transcripts. This was demonstrated by many fusion hubs in the FGI network, including DCAF12 and AXIN1 in LGG, PTK2 in OV, MYH14 and LRIG2 in

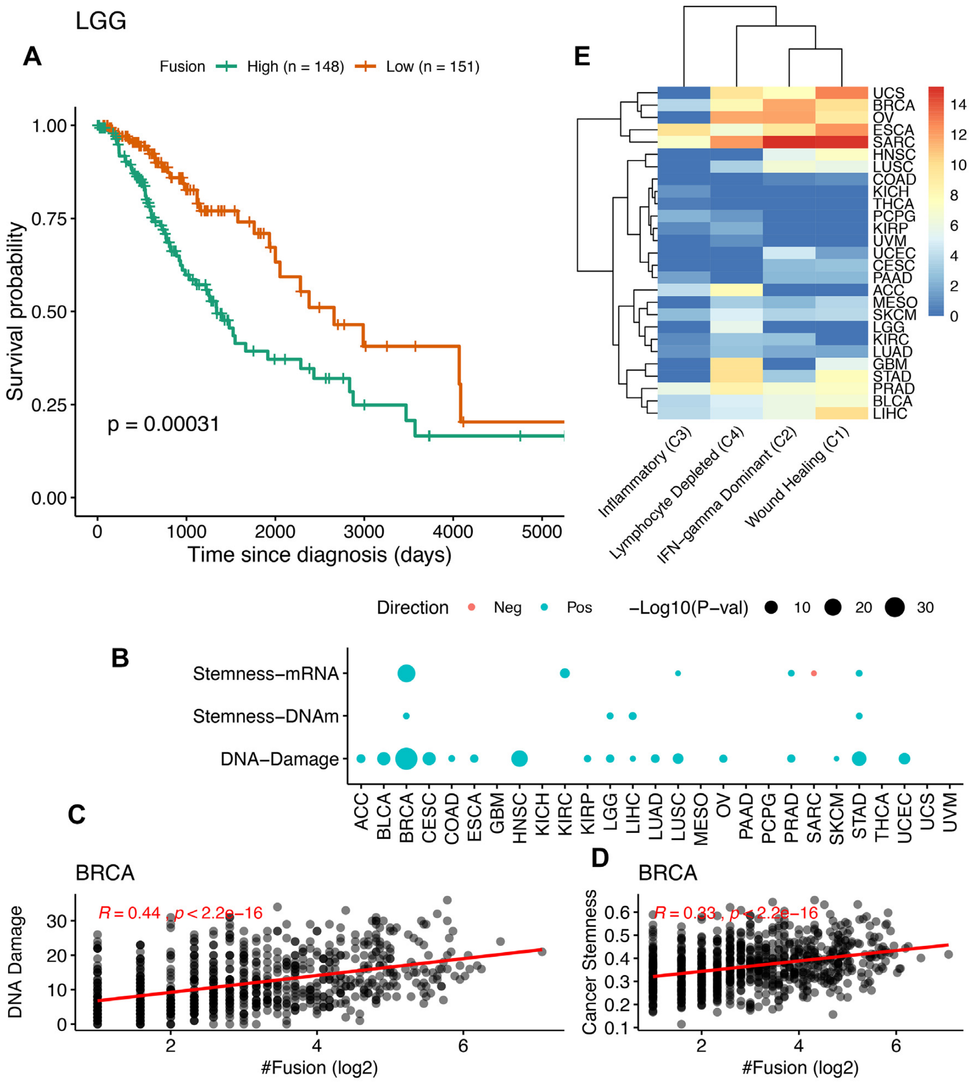

Figure 3. Associations between fusion frequencies and cancer features or patient survival. (**A**) Higher fusions in LGG is very significantly associated with poorer survival of patients (log-rank test, $P = 3.1\text{E-}4$). (**B**) Correlation between fusions and DNA damage and cancer stemness (mRNA expression-based and DNA methylation-based). The strongest signals are from BRCA for DNA damage (Pearson $R = 0.44$, $P < 2.2\text{E-}16$) (**C**) and cancer stemness (Pearson $R = 0.33$, $P < 2.2\text{E-}16$) (**D**). (**E**) Fusion distribution across four immune groups for each cancer. Inflammatory group C3 has the fewest fusions.

LUSC, CSNK1E in LIHC, TAB1 in GBM, MRPL49 in KIRP, BPTF in BRCA and TP53 and ELP5 in PRAD. Of note, BPTF was essential for triple-negative breast cancer (TNBC, the most malignant BRCA) metastasis (53). PTK2 was critical for tumor invasiveness and drug resistance, including in ovarian cancer (54). Interestingly, TP53 and ELP5 were functionally related (55) and shared many common FGI interactions. This eRNA-mediated FGI network may help us explain some high-frequency fusions for a single protein-coding gene and pinpoint the underlying tumorigenic mechanisms for some cancers.

Cancer druggability can potentially be augmented and drug responses were possibly altered by lncRNA fusions

Drugs targeting FGI hubs probably will be highly effective, which may provide new angles to tackle relevant cancers. In the FGI network, we chose 49 hub genes (with ≥10 interactions) and searched for potential drugs targeting them in DGIdb (30). Nine genes (BPTF, CREBBP, CSNK1E, EGFR, ERBB2, PTK2, SERPINA6, TMPRSS2 and TP53) had available drugs (Figure 5) and five of them had Food and Drug Administration (FDA)-approved drugs, including CREBBP (one drug) fusions in LGG and BLCA and SERPINA6 (eight drugs) in LIHC (Supplementary Figure S7A).

Drug responses potentially altered by eRNA fusions were also examined, by using a previous report of eRNA–drug interactions (19). A total of 52 412 eRNA–drug interactions, formed by 419 eRNAs and 650 drugs, were potentially affected by our eRNA fusions. (Supplementary Figure S7B). However, we note that the expression level of lncRNA fusions needs to be considered when accurate estimation of fusion expression are available in the future. Although higher fusion expression possibly mediates larger impact on druggability and drug responses, due to within-tumor het-

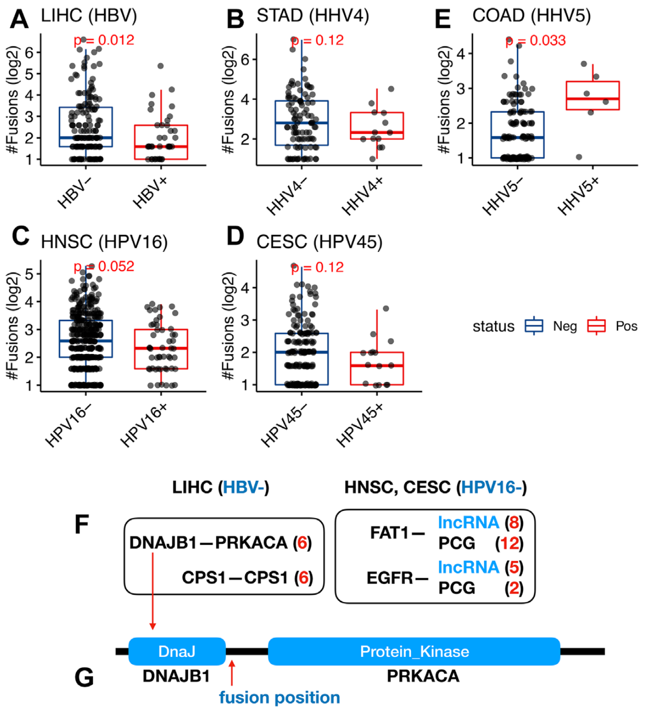

Figure 4. Fusion frequencies are negatively correlated with virus infection. Cancers showing more fusions include HBV-negative LIHC (**A**), HHV4-negative STAD (**B**), HPV16-negative HNSC (**C**) and HPV45-negative CESC (**D**). (**E**) HHV5-negative COAD tumors show less fusion than HHV5-positive ones. (**F**) Example of four fusions only found in HBV-negative LIHC (DNAJB1–PRKACA, CPS1–CPS1) or HPV16-negative HNSC and CESC samples (FAT1- and EGFR- fusions with other protein-coding genes (PCG) and some lncRNAs). Red numbers represent corresponding fusion frequencies. (**G**) Schematic diagram of the DNAJB1–PRKACA fusion protein with domains indicated. This chimeric protein was reported to cause rare fibrolamellar liver cancer. DnaJ was reported to inhibit HBV replication. One-sided *t*-test was used (A–E).

erogeneity, low fusion expression may represent driver fusion activity in cancer stem cells, a small population of tumor cells (56,57).

FPL-derived novel fusion proteins were potentially functional in breast cancer

Despite the huge volume of identified fusion events, experimental validations of fusions functioning in cancer are very limited, and as far as we know, are only reported for mRNA–mRNA fusions (2,11). Here, we manually inspected mRNA–lncRNA fusions and focused on two genes of interest, KDM4B (Lysine Demethylase 4B) and EPS15L1 (Epidermal Growth Factor Receptor Substrate 15-Like 1) with lncRNA fusions. Their fusions were prognostic for BRCA patient survival (log-rank test, $P = 1.5\text{E-}4$ for KDM4B and 1.8E-4 for EPS15L1, Cox regression hazard ratio = 3.24 for KDM4B and 4.32 for EPS15L1 after adjustment for confounding factors; Figure 6A and B, Supplementary Figure S8). We further selected two FPL fusion events for functional analysis, KDM4B–G039927 (the fusion protein was termed as KDM4Bf, 'f' for fusion) and EPS15L1–lncOR7C2–1 (termed EPS15L1f, 'f' for fusion) (Figure 6C), whose protein products were both detected by mass spectrometry (MS, both $P = 0.0020$, Figure 6D–E, chimeric open reading frame sequences were provided in Supplementary Table S5), for functional validation.

KDM4B is a hypoxia-inducible histone lysine demethylase, promoting DNA damage and genome instability through demethylating retrotransposons, especially in breast cancer (58). For KDM4Bf (Figure 6C), loss of the C-terminal domains of KDM4B probably abolished its pro-inflammatory function (59) that is important in anti-cancer therapies and made it structurally similar to KDM4D (60), which is enriched in testis (61) and is associated with cancer metastasis (62).

EPS15L1 regulates epidermal growth factor receptor (EGFR) signaling (63). For EPS15L1f, two ubiquitin-interacting domains located in the C-terminal of EPS15L1 were lost (64), which may render it constitutively active without being targeted by the ubiquitin system for degradation by the proteasome. Proteomics analysis showed that proteins from various oncogenic pathways were dysregulated (Figure 6F and Supplementary Table S6), including down-regulation of DEPDC5 (inhibitor of mechanistic target of rapamycin complex 1 or mTORC1 pathway), COPRS (differentiation stimulator), ERCC8 (DNA repair) and FZD6 (negative regulation of Wnt signaling), and up-regulation of G6PD (energy production).

Interestingly, phosphoproteomics analysis showed that EPS15L1f was also associated with decreased phosphorylation of DFNA5 (also known as GSDME, Gasdermin E) at Ser-252 (*P*-value = 7.7E-5, ranked first out of 11 643 tested phosphorylation sites), and increased phosphorylation of GRB7 (Growth Factor Receptor Bound Protein 7) at Ser-86 (*P*-value = 1.5E-3, ranked sixth among 11 643 sites) (Figure 6G and Supplementary Table S7). The tumor suppressor GSDME has recently been reported by two groups that it can induce pyroptosis, which is pro-inflammatory and enhances anti-tumor immunity through inducing tumor cell death and recruiting cytotoxic immune cells to the tumor sites (65,66), including in breast cancer. Activity of GSDME was potentially affected by phosphorylation (67). In addition, TNBC cells require GRB7 (68), whose phosphorylation by focal adhesion kinase (FAK) could stimulate tumor cell migration (69).

KDM4Bf and EPS15L1f promote SKBR3 cell proliferation

The above data promoted further examination of fusion events contributing to the tumorigenesis. KDM4B and EPS15L1 fusion events, which generate shorter ORFs, were selected to test whether these short chimeric proteins could promote breast cancer cell proliferation. By real-time PCR assay performed in a panel of breast cancer cells, SKBR3 cells were found to have low expression levels of both wild-type KDM4B and EPS15L1(Figure 7A). Upon overexpression of full-length (wild-type) and fusion (short) ORFs in SKBR3 cells (Figure 7B), the fusion KDM4Bf and EPS15L1f greatly increased SKBR3 cell proliferation relative to empty control and full length ORFs (Figure 7C and D), suggesting that lncRNA fusion events could play vital functions in tumorigenesis.

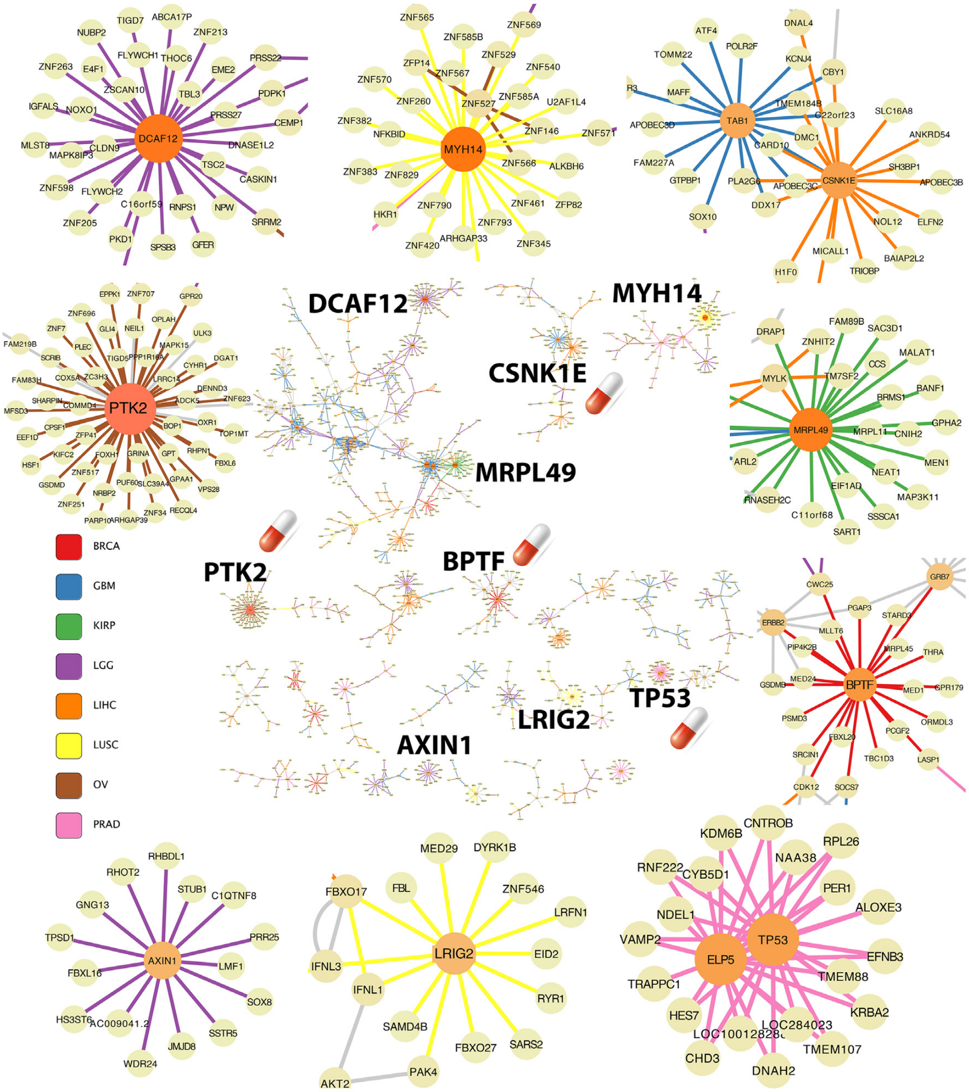

Figure 5. The network of secondary fusion events (FGI) mediated by enhancer (RNA)-promoter long-range interactions. Shown in the center is the main structure of this network (for better visualization, only connected subnetworks with at least 15 genes were included). Nine subnetworks (each with one or two hub genes, DCAF12, AXIN1, PTK2, MYH14, LRIG2, CSNK1E, MRPL49, BPTF and TP53) are shown as examples, displayed closely to their positions (marked genes) in the full network in the center. Edge colors stand for eight different cancers as indicated in the legend and the edges for remaining cancers are colored as gray. Node size represents the number of direct interactions. Four (CSNK1E, PTK2, BPTF and TP53) of the nine genes are marked by a drug icon, indicating available drugs in the DGIdb database.

DISCUSSION

Long noncoding RNAs have been increasingly recognized as important players in various diseases, including cancers. We developed an atlas of lncRNA-involved tumor-specific fusion events across cancer types, by integrating three large lncRNA annotation databases and using two high-accuracy fusion calling algorithms. We explored this fusion atlas and revealed interesting characteristics of fusions in cancer, which have not been reported before and provided novel angles to understand connections between fusion and cancer and uncovered potential mechanisms of fusions generated in cancer. This work also enriched our understanding of lncRNA functions. A summary of our findings were summarized in Supplementary Figure S9.

Most fusions were deemed to originate from genomic structural variations (1,2), which may be the products of complex events such as chromothripsis (70). We found most of our fusions were from duplications and translocations, which were also supported by our cancer subtype analysis showing positive correlation between frequent fusions and high SCNAs that possibly resulted from abundant extra-chromosomal DNAs (ecDNA) in cancer (41). However, we identified a group of secondary fusion events, which formed the FGI network, were possibly generated by a novel mechanism that involved eRNA-mediated long-range target interaction and RNA–DNA interaction. A large number fusion hub genes, some with FDA-approved drugs available, were found in this FGI network, possibly due to the large number targets for individual eRNAs and high efficiency of fusion-generating by RNA–DNA interactions (29). We

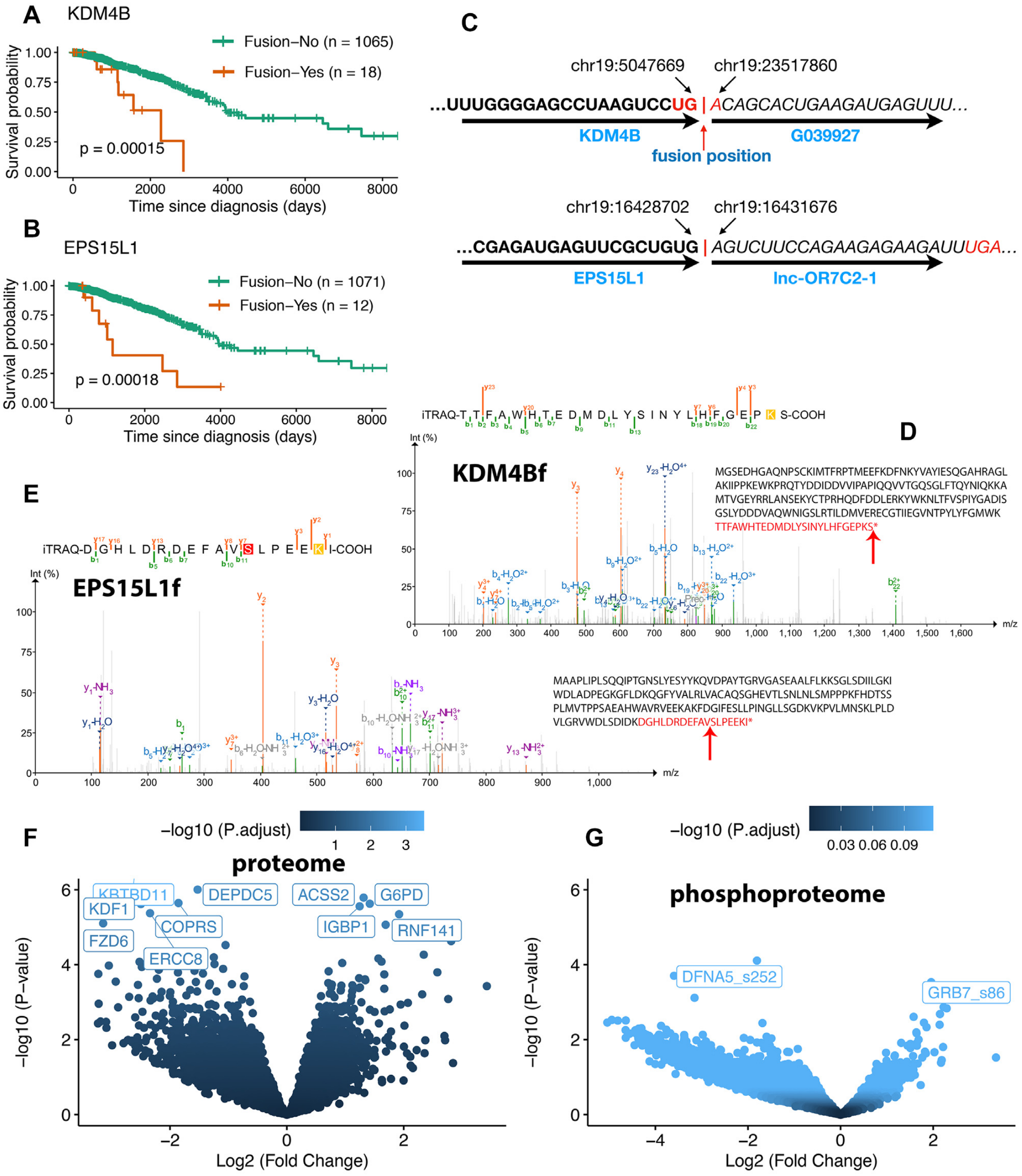

Figure 6. Characterizing two protein genes with lncRNA fusions in BRCA. (**A**) BRCA patients with KDM4B fusions have very poor outcome. (**B**) BRCA patients with EPS15L1 fusions have very poor outcome. (**C**) Chimeric transcript sequences near the fusion positions (the red pipe symbol) for the selected event KDM4B–G039927 and EPS15L1–lncOR7C2–1. UGA (red color) is the stop codon. (**D**) One of the mass spectrum (*x*-axis is *m*/*z* and *y*-axis is intensity) that matches the KDM4Bf fusion peptide (shown on the top). The full chimeric fusion protein sequence was shown, with mass spectrum identified peptide marked in red and the fusion breakpoint marked by an arrow. (**E**) Similar to (**D**), one of the mass spectrum that matches the EPS15L1f fusion peptide. The full chimeric fusion protein sequence was shown, with mass spectrum identified peptide marked in red and the fusion breakpoint marked by an arrow. (**F**) Differential protein levels for tumor samples with EPS15L1 fusions, with the top ten proteins shown. (**G**) Differential protein phosphorylation levels for tumor samples with EPS15L1 fusions. Only DFNA5 (also known as GSMDE) and GRB7 showing suggestive signals are labeled.

created the FGI network by using long-range enhancer-promoter interactions from different technologies, including Hi-C and CRISPR. With the accumulation of data from more studies and more advanced technologies, we expect this network will be largely improved and expanded. However, we noted that some fusions in this network may not be functional, which need much future work to unravel.

Two features of fusions seemed surprising initially—the negative correlation of fusion frequencies with microsatellite instability and virus infection, which were also supported by our cancer subtype analysis. However, they became reasonable after considering the possible mechanisms of tumorigenesis. Both MSI-High and virus infected cells possibly did not require higher fusion events to increase

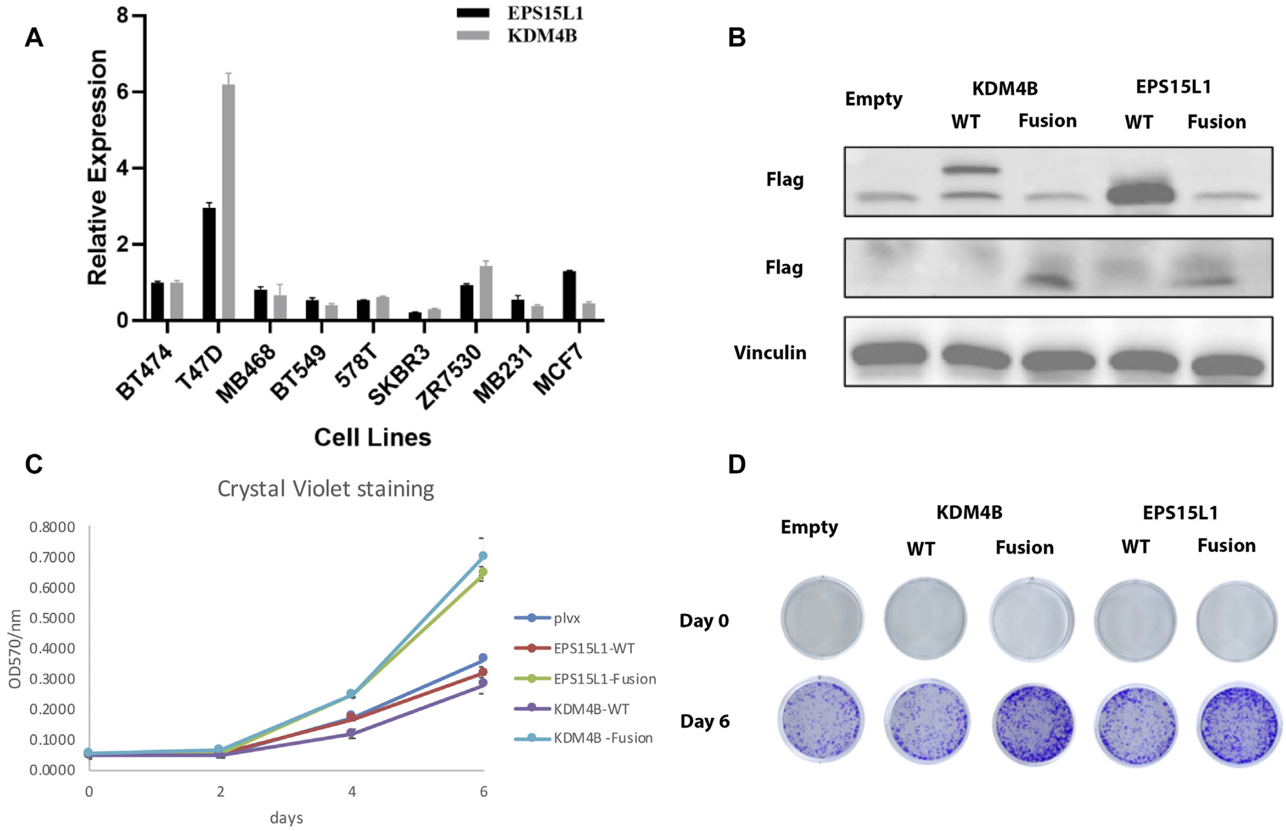

Figure 7. Fusion proteins from KDM4B–G039927 and EPS15L1–lncOR7C2–1 fusion events promote breast cancer cell proliferation. (**A**) Real-time PCR detection of EPS15L1 and KDM4B in a panel of breast cancer cells. Values represent mean ± s.d.. (**B**) Western blot analysis of SKBR3 stable cells expressing wild-type (WT) and fusion (Fusion) ORFs of KDM4B and EPS15L1. Vinculin is used as loading control. (**C**) Cells were cultured for different days as indicated, and then subjected to Crystal Violet staining; Error bar: standard deviation. (**D**) Cell growth assay for SKBR3 cells expressing various ORFs. Cells are seed the same numbers and grow for 6 days. Cells are stained with 0.1% Crystal Violet.

their oncogenic potential and induce cancer. A previous study reported that driver mutations in genes, such as TP53 and FAT1, were exclusively found in HPV-negative cancer cells. These results were also consistent with previous finding that fusion events were mutually exclusive with driver mutations (10). Altogether, most cancer possibly only require one major type of driving events. Although we performed correlations between all fusions and various phenotypes, we observed similar results when only using lncRNA fusions (Supplementary Figures S10–S13).

One of the limitations for lncRNA identification was the generally low expression of lncRNAs. Moreover, we imposed strict filters to our initial tumor fusion events, which would further underestimate lncRNA fusions. As stated by the author of the Arriba algorithm, the accuracy would be slightly lower after restricting Arriba to its self-declared high-confidence fusions. We demonstrated, in FGI network validation and survival analysis by fusions of a single gene, that our large number of raw tumor fusions should also be useful by enhancing the high-confidence fusions results. Furthermore, we expect many tumor-specific neoantigens would be contributed by FPL fusion proteins, exemplified by the KDM4B and EPS15L1 fusions, possibly comparable to those contributed by FPP fusions (10).

Collectively, our work completes the whole picture of fusions in cancer. Comprehensive analysis of the tumor-specific and lncRNA-dominated fusion landscape across various cancer types in new angles reveals insights into fusions in cancer and enriches our understanding of fusion functions in tumorigenesis and cancer progression. Our work also introduces numerous possibilities in cancer drug development and cancer treatment.

DATA AVAILABILITY

Raw sequencing reads for SKBR3 are available through the NCBI via GEO Accession GSE157986.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank National Supercomputer Center in Guangzhou for their computing service (Tianhe-2). The authors would also like to thank TCGA and GTEx for providing full data access.

Author Contributions: Y.Y.X. and M.B.G. conceived the project. M.B.G. and L.Z. collected the raw sequencing data. M.B.G. wrote the fusion pipeline and L.Z. assisted M.B.G. in the execution of the pipeline. M.B.G. performed the analysis, interpreted the results and wrote the manuscript. Y.Y.X. and Z.M.D. participated in the analysis and critically reviewed the manuscript. Z.D.X. designed the experiment, performed the validation and wrote the experimental validation part of the manuscript. L.T.D. and H.L. assisted Z.D.X. in the experimental validation. All authors approved the manuscript.

FUNDING

National Natural Science Foundation of China (NSFC) [31871323, U1611265, 61872395]; National Natural Science Funds of Guangdong Province for Distinguished Young Scholar [2014A030306044]; National Natural Science Foundation of China [81974436]; Program for Guangdong Introducing Innovative and Enterpreneurial Teams [2019ZT08Y485]; Fundamental Research Funds for the Central Universities [19ykzd06]. Funding for open access charge: NSFC.
Conflict of interest statement. None declared.

REFERENCES

1. Group,P.T.C., Calabrese,C., Davidson,N.R., Demircioglu,D., Fonseca,N.A., He,Y., Kahles,A., Lehmann,K.V., Liu,F., Shiraishi,Y. *et al.* (2020) Genomic basis for RNA alterations in cancer. *Nature*, **578**, 129–136.
2. Yang,L., Lee,M.S., Lu,H., Oh,D.Y., Kim,Y.J., Park,D., Park,G., Ren,X., Bristow,C.A., Haseley,P.S. *et al.* (2016) Analyzing Somatic Genome Rearrangements in Human Cancers by Using Whole-Exome Sequencing. *Am. J. Hum. Genet.*, **98**, 843–856.
3. Brien,G.L., Stegmaier,K. and Armstrong,S.A. (2019) Targeting chromatin complexes in fusion protein-driven malignancies. *Nat. Rev. Cancer*, **19**, 255–269.
4. Mertens,F., Johansson,B., Fioretos,T. and Mitelman,F. (2015) The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer*, **15**, 371–381.
5. Cancer Genome Atlas Research, N., Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
6. Consortium,G.T. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
7. Hu,X., Wang,Q., Tang,M., Barthel,F., Amin,S., Yoshihara,K., Lang,F.M., Martinez-Ledesma,E., Lee,S.H., Zheng,S. *et al.* (2018) TumorFusions: an integrative resource for cancer-associated transcript fusions. *Nucleic Acids Res.*, **46**, D1144–D1149.
8. Jang,Y.E., Jang,I., Kim,S., Cho,S., Kim,D., Kim,K., Kim,J., Hwang,J., Kim,S., Kim,J. *et al.* (2020) ChimerDB 4.0: an updated and expanded database of fusion genes. *Nucleic Acids Res.*, **48**, D817–D824.
9. Kim,P. and Zhou,X. (2019) FusionGDB: fusion gene annotation DataBase. *Nucleic Acids Res.*, **47**, D994–D1004.
10. Gao,Q., Liang,W.W., Foltz,S.M., Mutharasu,G., Jayasinghe,R.G., Cao,S., Liao,W.W., Reynolds,S.M., Wyczalkowski,M.A., Yao,L. *et al.* (2018) Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep.*, **23**, 227–238.
11. Picco,G., Chen,E.D., Alonso,L.G., Behan,F.M., Goncalves,E., Bignell,G., Matchan,A., Fu,B., Banerjee,R., Anderson,E. *et al.* (2019) Functional linkage of gene fusions to cancer cell fitness assessed by pharmacological and CRISPR-Cas9 screening. *Nat. Commun.*, **10**, 2198.
12. Singh,S., Qin,F., Kumar,S., Elfman,J., Lin,E., Pham,L.P., Yang,A. and Li,H. (2020) The landscape of chimeric RNAs in non-diseased tissues and cells. *Nucleic Acids Res.*, **48**, 1764–1778.
13. Kung,J.T., Colognori,D. and Lee,J.T. (2013) Long noncoding RNAs: past, present, and future. *Genetics*, **193**, 651–669.
14. Rinn,J.L. and Chang,H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.*, **81**, 145–166.
15. Ke,L., Yang,D.C., Wang,Y., Ding,Y. and Gao,G. (2020) AnnoLnc2: the one-stop portal to systematically annotate novel lncRNAs for human and mouse. *Nucleic Acids Res.*, **48**, W230–W238.
16. Kopp,F. and Mendell,J.T. (2018) Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell*, **172**, 393–407.
17. Cao,S., Wendl,M.C., Wyczalkowski,M.A., Wylie,K., Ye,K., Jayasinghe,R., Xie,M., Wu,S., Niu,B., Grubb,R. 3rd *et al.* (2016) Divergent viral presentation among human tumors and adjacent normal tissues. *Sci. Rep.*, **6**, 28294.
18. Cortes-Ciriano,I., Lee,S., Park,W.Y., Kim,T.M. and Park,P.J. (2017) A molecular portrait of microsatellite instability across multiple cancers. *Nat. Commun.*, **8**, 15180.
19. Zhang,Z., Lee,J.H., Ruan,H., Ye,Y., Krakowiak,J., Hu,Q., Xiang,Y., Gong,J., Zhou,B., Wang,L. *et al.* (2019) Transcriptional landscape and clinical utility of enhancer RNAs for eRNA-targeted therapy in cancer. *Nat. Commun.*, **10**, 4562.
20. Jung,I., Schmitt,A., Diao,Y., Lee,A.J., Liu,T., Yang,D., Tan,C., Eom,J., Chan,M., Chee,S. *et al.* (2019) A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.*, **51**, 1442–1449.
21. Fulco,C.P., Nasser,J., Jones,T.R., Munson,G., Bergman,D.T., Subramanian,V., Grossman,S.R., Anyoha,R., Doughty,B.R., Patwardhan,T.A. *et al.* (2019) Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.*, **51**, 1664–1669.
22. Iyer,M.K., Niknafs,Y.S., Malik,R., Singhal,U., Sahu,A., Hosono,Y., Barrette,T.R., Prensner,J.R., Evans,J.R., Zhao,S. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.*, **47**, 199–208.
23. Zhao,Y., Li,H., Fang,S., Kang,Y., Wu,W., Hao,Y., Li,Z., Bu,D., Sun,N., Zhang,M.Q. *et al.* (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.*, **44**, D203–D208.
24. Volders,P.J., Anckaert,J., Verheggen,K., Nuytens,J., Martens,L., Mestdagh,P. and Vandesompele,J. (2019) LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.*, **47**, D135–D139.
25. Ma,L., Cao,J., Liu,L., Du,Q., Li,Z., Zou,D., Bajic,V.B. and Zhang,Z. (2019) LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res.*, **47**, D128–D134.
26. Haas,B.J., Dobin,A., Li,B., Stransky,N., Pochet,N. and Regev,A. (2019) Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.*, **20**, 213.
27. Colaprico,A., Silva,T.C., Olsen,C., Garofano,L., Cava,C., Garolini,D., Sabedot,T.S., Malta,T.M., Pagnotta,S.M., Castiglioni,I. *et al.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71.
28. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
29. Yan,Z., Huang,N., Wu,W., Chen,W., Jiang,Y., Chen,J., Huang,X., Wen,X., Xu,J., Jin,Q. *et al.* (2019) Genome-wide colocalization of RNA-DNA interactions and fusion RNA pairs. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 3328–3337.
30. Griffith,M., Griffith,O.L., Coffman,A.C., Weible,J.V., McMichael,J.F., Spies,N.C., Koval,J., Das,I., Callaway,M.B., Eldred,J.M. *et al.* (2013) DGIdb: mining the druggable genome. *Nat. Methods*, **10**, 1209–1210.
31. Wen,B., Wang,X. and Zhang,B. (2019) PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. *Genome Res.*, **29**, 485–493.
32. Sato,K., Kawazu,M., Yamamoto,Y., Ueno,T., Kojima,S., Nagae,G., Abe,H., Soda,M., Oga,T., Kohsaka,S. *et al.* (2019) Fusion Kinases Identified by Genomic Analyses of Sporadic Microsatellite Instability-High Colorectal Cancers. *Clin. Cancer Res.*, **25**, 378–389.

33. Monument,M.J., Lessnick,S.L., Schiffman,J.D. and Randall,R.T. (2012) Microsatellite instability in sarcoma: fact or fiction? *ISRN Oncol.*, **2012**, 473146.
34. Delattre,O., Zucman,J., Melot,T., Garau,X.S., Zucker,J.M., Lenoir,G.M., Ambros,P.F., Sheer,D., Turc-Carel,C., Triche,T.J. *et al.* (1994) The Ewing family of tumors–a subgroup of small-round-cell tumors defined by specific chimeric transcripts. *N. Engl. J. Med.*, **331**, 294–299.
35. Cancer Genome Atlas, N. (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, **517**, 576–582.
36. Cancer Genome Atlas, N. (2015) Genomic Classification of Cutaneous Melanoma. *Cell*, **161**, 1681–1696.
37. Cancer Genome Atlas Research, N., Kandoth,C., Schultz,N., Cherniack,A.D., Akbani,R., Liu,Y., Shen,H., Robertson,A.G., Pashtan,I., Shen,R. *et al.* (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**, 67–73.
38. Cancer Genome Atlas Research Network. Electronic address, e.d.s.c. and Cancer Genome Atlas Research, N. (2017) Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell*, **171**, 950–965.
39. Ceccarelli,M., Barthel,F.P., Malta,T.M., Sabedot,T.S., Salama,S.R., Murray,B.A., Morozova,O., Newton,Y., Radenbaugh,A., Pagnotta,S.M. *et al.* (2016) Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*, **164**, 550–563.
40. Liu,Y., Sethi,N.S., Hinoue,T., Schneider,B.G., Cherniack,A.D., Sanchez-Vega,F., Seoane,J.A., Farshidfar,F., Bowlby,R., Islam,M. *et al.* (2018) Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell*, **33**, 721–735.
41. Wu,S., Turner,K.M., Nguyen,N., Raviram,R., Erb,M., Santini,J., Luebeck,J., Rajkumar,U., Diao,Y., Li,B. *et al.* (2019) Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature*, **575**, 699–703.
42. Kleppe,A., Albregtsen,F., Vlatkovic,L., Pradhan,M., Nielsen,B., Hveem,T.S., Askautrud,H.A., Kristensen,G.B., Nesbakken,A., Trovik,J. *et al.* (2018) Chromatin organisation and cancer prognosis: a pan-cancer study. *Lancet Oncol.*, **19**, 356–369.
43. Yoon,J.Y., Brezden-Masley,C. and Streutker,C.J. (2020) Lgr5 and stem/progenitor gene expression in gastric/gastroesophageal junction carcinoma - significance of potentially retained stemness. *BMC Cancer*, **20**, 860.
44. Morel,A.P., Ginestier,C., Pommier,R.M., Cabaud,O., Ruiz,E., Wicinski,J., Devouassoux-Shisheboran,M., Combaret,V., Finetti,P., Chassot,C. *et al.* (2017) A stemness-related ZEB1-MSRB3 axis governs cellular pliancy and breast cancer genome stability. *Nat. Med.*, **23**, 568–578.
45. Knijnenburg,T.A., Wang,L., Zimmermann,M.T., Chambwe,N., Gao,G.F., Cherniack,A.D., Fan,H., Shen,H., Way,G.P., Greene,C.S. *et al.* (2018) Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep.*, **23**, 239–254.
46. Malta,T.M., Sokolov,A., Gentles,A.J., Burzykowski,T., Poisson,L., Weinstein,J.N., Kaminska,B., Huelsken,J., Omberg,L., Gevaert,O. *et al.* (2018) Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell*, **173**, 338–354.
47. Thorsson,V., Gibbs,D.L., Brown,S.D., Wolf,D., Bortone,D.S., Ou Yang,T.H., Porta-Pardo,E., Gao,G.F., Plaisier,C.L., Eddy,J.A. *et al.* (2018) The Immune Landscape of Cancer. *Immunity*, **48**, 812–830.
48. Eckhardt,M., Zhang,W., Gross,A.M., Von Dollen,J., Johnson,J.R., Franks-Skiba,K.E., Swaney,D.L., Johnson,T.L., Jang,G.M., Shah,P.S. *et al.* (2018) Multiple Routes to Oncogenesis Are Promoted by the Human Papillomavirus-Host Protein Network. *Cancer Discov.*, **8**, 1474–1489.
49. Cancer Genome Atlas Research Network. Electronic address, w.b.e. and Cancer Genome Atlas Research, N. (2017) Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell*, **169**, 1327–1341.
50. Sohn,S.Y., Kim,J.H., Baek,K.W., Ryu,W.S. and Ahn,B.Y. (2006) Turnover of hepatitis B virus X protein is facilitated by Hdj1, a human Hsp40/DnaJ protein. *Biochem. Biophys. Res. Commun.*, **347**, 764–768.
51. Kastenhuber,E.R., Lalazar,G., Houlihan,S.L., Tschaharganeh,D.F., Baslan,T., Chen,C.C., Requena,D., Tian,S., Bosbach,B., Wilkinson,J.E. *et al.* (2017) DNAJB1-PRKACA fusion kinase interacts with beta-catenin and the liver regenerative response to drive fibrolamellar hepatocellular carcinoma. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 13076–13084.
52. Jin,J. (2018) HPV Infection and Cancer. *JAMA*, **319**, 1058.
53. Koedoot,E., Fokkelman,M., Rogkoti,V.M., Smid,M., van de Sandt,I., de Bont,H., Pont,C., Klip,J.E., Wink,S., Timmermans,M.A. *et al.* (2019) Uncovering the signaling landscape controlling breast cancer cell migration identifies novel metastasis driver genes. *Nat. Commun.*, **10**, 2983.
54. Sulzmaier,F.J., Jean,C. and Schlaepfer,D.D. (2014) FAK in cancer: mechanistic findings and clinical applications. *Nat. Rev. Cancer*, **14**, 598–610.
55. Xu,S., Zhan,M., Jiang,C., He,M., Yang,L., Shen,H., Huang,S., Huang,X., Lin,R., Shi,Y. *et al.* (2019) Genome-wide CRISPR screen identifies ELP5 as a determinant of gemcitabine sensitivity in gallbladder cancer. *Nat. Commun.*, **10**, 5492.
56. Eppert,K., Takenaka,K., Lechman,E.R., Waldron,L., Nilsson,B., van Galen,P., Metzeler,K.H., Poeppl,A., Ling,V., Beyene,J. *et al.* (2011) Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat. Med.*, **17**, 1086–1093.
57. Ishizawa,K., Rasheed,Z.A., Karisch,R., Wang,Q., Kowalski,J., Susky,E., Pereira,K., Karamboulas,C., Moghal,N., Rajeshkumar,N.V. *et al.* (2010) Tumor-initiating cells are rare in many human tumors. *Cell Stem Cell*, **7**, 279–282.
58. Xiang,Y., Yan,K., Zheng,Q., Ke,H., Cheng,J., Xiong,W., Shi,X., Wei,L., Zhao,M., Yang,F. *et al.* (2019) Histone Demethylase KDM4B Promotes DNA Damage by Activating Long Interspersed Nuclear Element-1. *Cancer Res.*, **79**, 86–98.
59. Kirkpatrick,J.E., Kirkwood,K.L. and Woster,P.M. (2018) Inhibition of the histone demethylase KDM4B leads to activation of KDM1A, attenuates bacterial-induced pro-inflammatory cytokine release, and reduces osteoclastogenesis. *Epigenetics*, **13**, 557–572.
60. Katoh,M. and Katoh,M. (2004) Identification and characterization of JMJD2 family genes in silico. *Int. J. Oncol.*, **24**, 1623–1628.
61. Iwamori,N., Zhao,M., Meistrich,M.L. and Matzuk,M.M. (2011) The testis-enriched histone demethylase, KDM4D, regulates methylation of histone H3 lysine 9 during spermatogenesis in the mouse but is dispensable for fertility. *Biol. Reprod.*, **84**, 1225–1234.
62. Soini,Y., Kosma,V.M. and Pirinen,R. (2015) KDM4A, KDM4B and KDM4C in non-small cell lung cancer. *Int. J. Clin. Exp. Pathol*, **8**, 12922–12928.
63. van Bergen En Henegouwen, P.M. (2009) Eps15: a multifunctional adaptor protein regulating intracellular trafficking. *Cell Commun. Signal.*, **7**, 24.
64. Polo,S., Sigismund,S., Faretta,M., Guidi,M., Capua,M.R., Bossi,G., Chen,H., De Camilli,P. and Di Fiore,P.P. (2002) A single motif responsible for ubiquitin recognition and monoubiquitination in endocytic proteins. *Nature*, **416**, 451–455.
65. Zhang,Z., Zhang,Y., Xia,S., Kong,Q., Li,S., Liu,X., Junqueira,C., Meza-Sosa,K.F., Mok,T.M.Y., Ansara,J. *et al.* (2020) Gasdermin E suppresses tumour growth by activating anti-tumour immunity. *Nature*, **579**, 415–420.
66. Wang,Q., Wang,Y., Ding,J., Wang,C., Zhou,X., Gao,W., Huang,H., Shao,F. and Liu,Z. (2020) A bioorthogonal system reveals antitumour immune function of pyroptosis. *Nature*, **579**, 421–426.
67. Rogers,C., Erkes,D.A., Nardone,A., Aplin,A.E., Fernandes-Alnemri,T. and Alnemri,E.S. (2019) Gasdermin pores permeabilize mitochondria to augment caspase-3 activation during apoptosis and inflammasome activation. *Nat. Commun.*, **10**, 1689.
68. Giricz,O., Calvo,V., Pero,S.C., Krag,D.N., Sparano,J.A. and Kenny,P.A. (2012) GRB7 is required for triple-negative breast cancer cell invasion and survival. *Breast Cancer Res. Treat.*, **133**, 607–615.
69. Han,D.C., Shen,T.L. and Guan,J.L. (2000) Role of Grb7 targeting to focal contacts and its phosphorylation by focal adhesion kinase in regulation of cell migration. *J. Biol. Chem.*, **275**, 28911–28917.
70. Cortes-Ciriano,I., Lee,J.J., Xi,R., Jain,D., Jung,Y.L., Yang,L., Gordenin,D., Klimczak,L.J., Zhang,C.Z., Pellman,D.S. *et al.* (2020) Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet.*, **52**, 331–341.