Analyzing Somatic Genome Rearrangements in Human Cancers by Using Whole-Exome Sequencing

Lixing Yang,[1,11] Mi-Sook Lee,[2,11] Hengyu Lu,[3,11] Doo-Yi Oh,[2] Yeon Jeong Kim,[4,5] Donghyun Park,[4,5] Gahee Park,[4] Xiaojia Ren,[6] Christopher A. Bristow,[7] Psalm S. Haseley,[1,6] Soohyun Lee,[1] Angeliki Pantazi,[8] Raju Kucherlapati,[6,8] Woong-Yang Park,[2,4] Kenneth L. Scott,[3,12] Yoon-La Choi,[2,9,12,*] and Peter J. Park[1,6,10,12,*]

Although exome sequencing data are generated primarily to detect single-nucleotide variants and indels, they can also be used to identify a subset of genomic rearrangements whose breakpoints are located in or near exons. Using >4,600 tumor and normal pairs across 15 cancer types, we identified over 9,000 high confidence somatic rearrangements, including a large number of gene fusions. We find that the 5′ fusion partners of functional fusions are often housekeeping genes, whereas the 3′ fusion partners are enriched in tyrosine kinases. We establish the oncogenic potential of *ROR1-DNAJC6* and *CEP85L-ROS1* fusions by showing that they can promote cell proliferation in vitro and tumor formation in vivo. Furthermore, we found that ~4% of the samples have massively rearranged chromosomes, many of which are associated with upregulation of oncogenes such as *ERBB2* and *TERT*. Although the sensitivity of detecting structural alterations from exomes is considerably lower than that from whole genomes, this approach will be fruitful for the multitude of exomes that have been and will be generated, both in cancer and in other diseases.

Introduction

Genomic profiling of tumors with high-throughput sequencing technologies has provided an unprecedented opportunity for in-depth studies of genome rearrangements. Whole-genome sequencing (WGS) data are now routinely used for detection of a wide range of rearrangements with base-pair resolution of breakpoints, including those breakpoints in non-coding regions. These events are typically identified on the basis of read depth,[1] discordant paired-end reads,[2] split-read (reads spanning the breakpoint) alignment,[3] genome assembly,[4] local assembly,[5] or by a combination of these methods.[6] RNA-seq data can be used to interrogate gene fusions when the fusion is expressed at a sufficiently high amount.

Whole-exome sequencing (WES) data are generated to detect single-nucleotide variants (SNVs) and small indels. An enormous number of exomes have been generated by researchers around the world: the latest release from the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project[7] includes 6,500 samples; the Exome Aggregation Consortium (ExAC), an international collaboration to collect exome data, has more than 60,000 exomes in its current release. Despite the decreasing cost of WGS, WES data will continue to be generated because many somatic variants occur at low variant allelic frequency, and the necessary high-depth (e.g., >100–500×) sequencing is affordable only with a capture-based approach given current technologies. An important question, therefore, is whether genomic rearrangements can also be detected in exomes. If that were possible, we would be able to identify a large number of rearrangements with datasets that were generated for other purposes.

Here, we describe an approach to identify structural variations (SVs) from WES data. In a typical WES protocol, genomic DNA is sheared into fragments (~150–250 bp), and those containing exons are enriched by hybridization with shorter biotinylated probes (~50–100 nucleotides long). These probes are usually densely tiled across exons, extending just past the exon-intron boundaries. Thus, when the breakpoint of an SV occurs in or near the targeted region, the DNA fragment that contains the breakpoint can be captured if there is sufficient overlap between a probe and the DNA on either side of the breakpoint (Figure 1A). The sensitivity of detection from WES is clearly much lower than that from WGS, given that just a subset of rearrangements with breakpoints in or near exons can be detected and the fragment capture process introduces inefficiencies. However, with the large number of available exomes and the higher coverage than WGS, we demonstrate that re-analyzing existing large-scale WES data for genomic rearrangements can yield valuable findings.

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA; [2]Department of Health Sciences and Technology, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul 06351, Korea; [3]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; [4]Samsung Genome Institute, Samsung Medical Center, Seoul 06351, Korea; [5]Samsung Biomedical Research Institute, Samsung Advanced Institute of Technology (SAIT), Samsung Electronics Co., Seoul 06351, Korea; [6]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115, USA; [7]Department of Genomic Medicine and Institute for Applied Cancer Science, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; [8]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; [9]Department of Pathology and Translational Genomics, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul 06351, Korea; [10]Ludwig Center, Harvard Medical School, Boston, MA 02115, USA
[11]These authors contributed equally to this work
[12]These authors contributed equally to this work
*Correspondence: ylachoi@skku.edu (Y.-L.C.), peter_park@harvard.edu (P.J.P.)
http://dx.doi.org/10.1016/j.ajhg.2016.03.017.

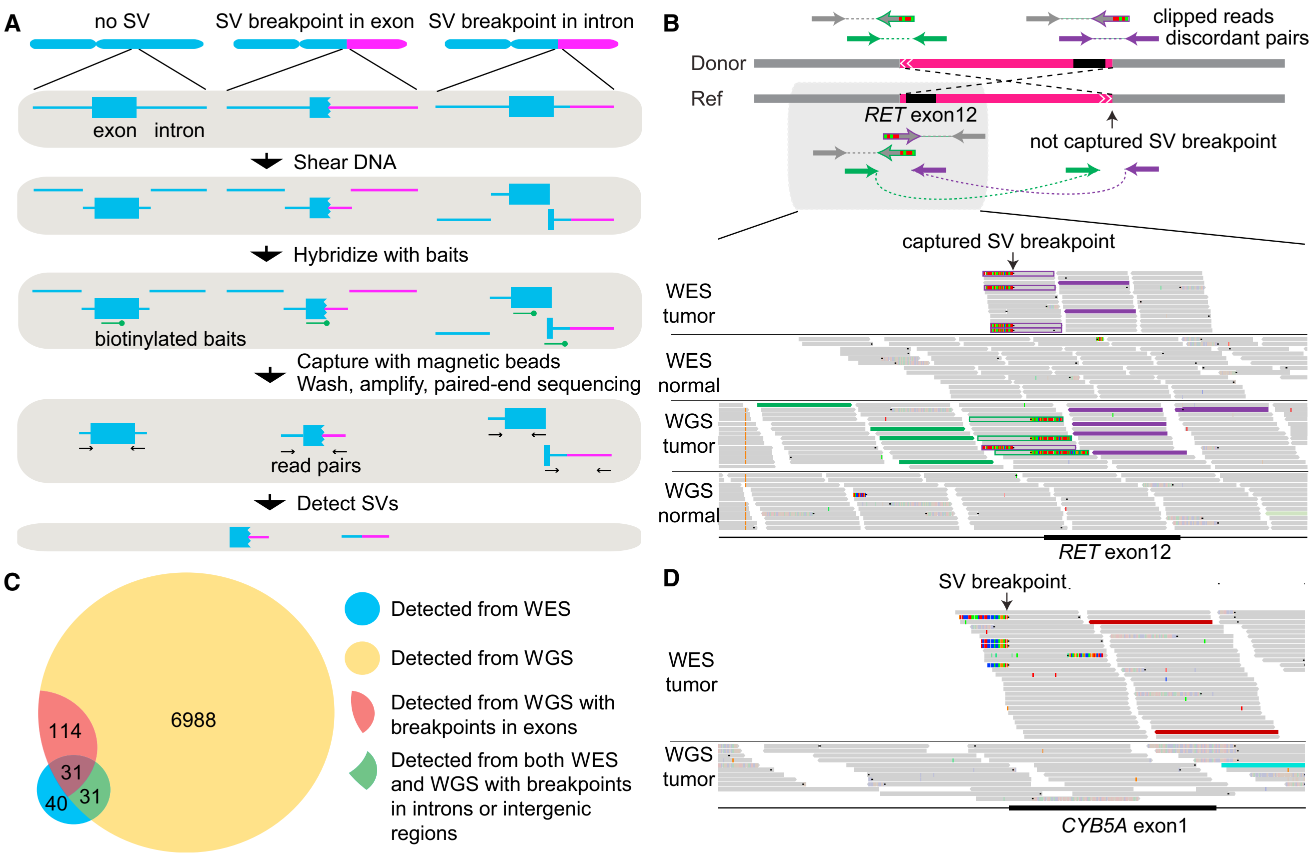

Figure 1. Detecting Somatic SVs from WES Data

(A) Workflow showing how DNA fragments are captured and sequenced when SV breakpoints occur in exons and near exon-intron boundaries.

(B) A true somatic *CCDC6-RET* fusion resulting from a balanced inversion (chr10:61,655,977–43,611,997) in a thyroid cancer (TCGA-FK-A3SE) is detected by both WES and WGS. The scheme of the inversion is shown on the top (not to scale). The Integrated Genome Viewer screen shot for the captured breakpoint is shown on the bottom. Green and purple read pairs represent discordant pairs from two different breakpoints; one breakpoint is captured by WES and the other is not. The gray reads are concordant read pairs. The half-gray and half-striped reads with green or purple outlines are partially aligned (clipped) reads spanning the breakpoints.

(C) A Venn diagram showing the overlap between somatic SVs called from WES and WGS data.

(D) A true somatic deletion (chr18:71,930,713–71,958,983) in a lung adenocarcinoma (TCGA-91-6840) is detected by WES but not by WGS and is validated by PCR. The coverage in WES is >100× and there are six discordant read pairs (two displayed), whereas the coverage of WGS in the same region is 30× and no discordant read pair is present. The red reads are discordant read pairs supporting the somatic deletion.

We applied our proposed method to survey somatic SVs in 4,609 samples across 15 tumor types from The Cancer Genome Atlas (TCGA). We focus on somatic variants here, but the approach we describe applies to detection of both germline and somatic rearrangements. We chose the TCGA data because they are high-quality, multi-dimensional data from a large number of samples, including cases that have undergone both WES and WGS. The availability of these two data types for the same samples allows us to characterize the sensitivity and specificity of exome-based SV detection. Although exome-based fusion detection has been recently used to identify recurrent *NAB2-STAT6* (MIM: 602381 and 601512) fusion in solitary fibrous tumors,[8] our study expands this approach to a much larger scale to discover additional cancer-driving gene fusions and characterize their features. Our results demonstrate the association of oncogene upregulation with massive rearrangements. We also report experimental validation that two of the candidate fusions we identified are cancer drivers, including the report of an activating genetic event related to *ROR1* (MIM: 602336).

Material and Methods

TCGA Sample Acquisition and WES

The details of data production were described in a previous publication.[9] The procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national). Tumor samples were obtained from the TCGA network with appropriate consent from the relevant institutional review board. Tumors were resected, flash-frozen, and shipped to a centralized processing center (Biospecimen Core Resource) for additional pathologic review and extraction of nucleic acids. The three genome sequencing centers (Baylor Human Genome Sequencing Center, Broad Institute, and The Genome Institute at Washington University) collectively sequenced the exomes from tumor tissues and matched normal tissues (mostly blood samples). Exome capturing procedures differ among sequencing centers and evolve over time. The details can be found in individual TCGA marker papers. Sequencing reads were aligned to the reference genome with the Burrows-Wheeler Aligner,[10] and quality control was performed. A single BAM file that includes reads, calibrated quantities, and alignments to the genome was generated for each sample.

Data Access

All primary sequence files can be downloaded by registered users from CGHub. Clinical data are available through the TCGA Data Portal. All coordinates are based on the hg19 human reference genome, downloaded from the UCSC Genome Browser.

Detecting Somatic Genome Rearrangements in WES Data

Somatic genome rearrangements were called by Meerkat, a software package we developed.[6] In brief, all discordant read pairs (reads that do not form a proper pair with expected orientations and distance between the reads) are first identified from the BAM files. Then, discordant read pairs supporting the same breakpoint are merged into clusters, which are used to call SV candidates. Reads spanning SV breakpoints (clipped reads and unmapped reads) are mapped back to the SV candidates (split-read mapping). Breakpoints are refined to the basepair resolution once split-read supports are identified. Variants are filtered by a large database of germline variants obtained by merging all matched normal BAM files from different tumor types together. The final somatic variants must have discordant read-pair support and split-read support totaling at least six reads and/or read pairs, with at least three discordant read-pair support. We have previously used these criteria to identify somatic SVs from WGS samples and have demonstrated that such a workflow offers great sensitivity and specificity. Samples with >100 somatic SVs were discarded from further analysis. Additional filters were applied to obtain high-confidence somatic rearrangements: at least four supporting discordant read pairs were required for each somatic event, and the size of an intra-chromosomal event could not be less than 20 kb. For comparison with WGS results, if the somatic rearrangement detected from WES data and the one detected from WGS data were the same type of event on the same chromosome(s) and the breakpoints differed by less than 50 bp, they were considered to be the same event. In most cases, the breakpoints predicted from WES and WGS were exactly the same. PCR primers were designed by Primer3.[11]

Detecting Activating Gene Fusions

RNA was extracted, prepared into Illumina TruSeq mRNA libraries, and sequenced by an Illumina sequencing platform with a target of 60 million read pairs per tumor (48 bp paired-end reads) and subjected to quality control. RNA reads were aligned to the reference genome with Mapsplice.[12] Gene expression was quantified for the transcript models (TCGA GAF2.1) with RSEM[13] and normalized within sample to a fixed upper quartile of total reads. RNA-seq results (normalized gene-level expression and exon-level expression) were downloaded from the Genome Data Analysis Center at the Broad Institute. RNA data were available only for tumor tissues because TCGA collected blood (rather than adjacent normal tissues, which are generally unavailable) as the matched normal control for the majority of the cases. Therefore, to normalize exonic expression, we computed a Z score for each exon on the basis of its expression across all samples in that tumor type. Gene Ontology (GO) term enrichment analyses were performed with DAVID.[14] All 5′ and 3′ fusion partners were entered into DAVID as a gene list to identify over-represented GO categories, and the functional annotation clustering of GO terms was performed. The p value was calculated by one-tail Fisher's exact test.

Analysis of Massive Rearrangements

A binomial model was used to identify the samples in which the number of somatic rearrangement breakpoints observed on one chromosome significantly exceeded the expected number, given the total number of somatic rearrangement breakpoints in one sample (the likelihood of observing at least n breakpoints on one chromosome given the total N breakpoints in that sample, with the probability p being the mappable coding-sequence (CDS) size for the chromosome divided by the mappable CDS size for the whole genome). Bonferroni correction was used to adjust for multiple testing. The mappability of the reference genome was downloaded from UCSC Genome browser and was used to normalize the chromosome size.

Statistical Analysis

All statistical analyses were conducted in R package (v.2.14.1). A p value of 0.01 was used for statistical significance.

Fusion Gene Cloning

Constructs of *CEP85L-ROS1(C9;R36)* (MIM: 165020), *GOPC* (MIM: 606845)-*ROS1(G7;R35)*, and *EML4* (MIM: 607442)-*ALK* (MIM: 105590) gene fusions were synthesized by CosmogeneTech and then transferred into pLenti6.3/V5-DEST (Life Technologies) and pLenti6.3-EF1α lentiviral vectors. *ROR1-DNAJC6* (MIM: 608375) fusion fragments were cloned from cDNA prepared from U87MG cells with overlapping ends, fused *ROR1-DNAJC6* was then generated by overlap-extension PCR, and the resulting fusion gene was then transferred into the pLenti6.3/V5-DEST vector. Expression of the *ROR1-DNAJC6* fusion gene was confirmed via RT-PCR and western blots with the following primer sets: forward, *ROR1,* 5′-GTGATGAAGATGGGACTGTGAA-3′; reverse, *DNAJC6*, 5′-CTA GAAGATGTGTCTTTGAGGGTGT-3′.

Ba/F3 Cell Viability and Inhibitor Assays

The Ba/F3 cell line was maintained in RPMI 1640 medium with 5% fetal bovine serum and 2.5 ng/ml recombinant mouse IL-3. *CEP85L-ROS1*, *BCR* (MIM: 151410) -*ABL* ([MIM: 189980] positive control), and *GFP* (negative control) were transduced into Ba/F3 cells. At 72 hr post-transduction, cells were re-suspended in medium without IL-3. Cell viability was determined with Cell Titer-Glo (Promega) at 7 days after IL-3 depletion. Ba/F3 cells stably expressing *CEP85L-ROS1* (no IL-3 medium) and parental Ba/F3 cells (IL-3 medium) were seeded in 96-well plates in quadruplicates at 1,000 cells per well. For the dose-dependent inhibitor assay, cells were treated with dimethyl sulfoxide (DMSO) or crizotinib (5 nM to 0.5 μM) and cell viability was determined with Cell Titer-Glo (Promega). Cell survival was normalized to non-treated (DMSO control treated) cells. IC_{50}, which is the concentration of an inhibitor causing 50% inhibition of cell survival normalized to non-treated cells, was calculated from a sigmoidal curve. The response of *CEP85L-ROS1*-expressing cells (without IL3) to crizotinib was compared to parental cells without treatment of crizotinib as control. Two independent experiments were performed.

Western Blot

Whole-cell and mouse tumor tissue lysates were prepared with radioimmunoprecipitation assay (50 mM Tris-HCl, 150 mM NaCl, 1% NP-40, and 0.25% sodium deoxycholate) plus protease inhibitors cocktail (GenDepot). Cell and tissue lysates were separated by SDS-PAGE and transferred to polyvinylidene difluoride membranes. The blots were probed with antibodies for ROS1, phosphorylated, and total STAT3 (MIM: 102582), AKT and ERK (Cell Signaling Technology), and ROR1 (Abcam) were then detected with chemiluminescent substrate (EMD Millipore). All western blot images are representative of at least three independent experiments.

In Vitro Cell Proliferation and Transforming Assays

NIH 3T3 cells were obtained from the Korean Cell Line Bank, and BEAS-2B cells (ATCC CRL-9609) were obtained from the American Type Culture Collection (Manassas, VA). They were expanded in DMEM supplemented with 10% FBS, 100 units/ml penicillin, and 100 mg/ml streptomycin. NIH 3T3 cells and BEAS-2B cells were transduced with *LacZ* (negative control), *CEP85L-ROS1*, *GOPC-ROS1* (positive control), *ROR1-DNAJC6,* and *EML4-ALK* (positive control). Then, stable cell lines were selected with blasticidin. Cell proliferation was determined by a EZ-Cytox cell viability assay kit (Daeil Lab Service). The transforming activity was assessed by transformed foci formation in Matrigel. NIH 3T3 stable cells expressing *CEP85L-ROS1*, *GOPC-ROS1*, and *EML4-ALK,* and BEAS-2B stable cells expressing *ROR1-DNAJC6* and *EML4-ALK* were seeded in Matrigel (BD Sciences; 10,000 cells per well), on which medium with 10% FBS was overlaid. The images of transformed foci were taken after culturing for 7 or 14 days.

Anchorage Independent Growth Assay

MCF-10A cells were cultured as described previously[15] and transduced with *CEP85L-ROS1*, *PIK3CA*^{H1047R} (positive control), and *GFP* (negative control). Soft agar assays were performed in six-well plates in triplicate. First, bottom layers were prepared at 0.8% Noble agar (Affymetrix) with complete MCF-10A growth medium. After solidification, 10,000 cells were mixed with 0.45% agar in complete growth medium and laid on top of the bottom layer. 2 mL of medium was added in each well after 3 days, and the medium was refreshed every 3 days. For NIH 3T3 and BEAS-2B cells expressing *LacZ*, *CEP85L-ROS1*, *GOPC-ROS1*, *ROR1-DNAJC6*, and *EML4-ALK* in 0.35% agar (BD Sciences), 20,000 cells were seeded on top of 0.5% agar in each well. Cells were cultured for 14 or 21 days, colonies were stained with 0.05% crystal violet, and images were taken by phase-contrast microscope (Olympus CKX41) and analyzed by i-Solution Lite image analysis software, and cells were counted in ten randomly selected fields.

Xenograft Tumor Formation Assay

All animal experiments were approved by the institutional review board of Samsung Medical Center. 5×10^6 cells were re-suspended in 1:1 PBS and Matrigel (BD Biosciences) and then subcutaneously injected into the right dorsal flank of six-week-old male nude mice (Orient Bio). Mice were monitored three times weekly until reaching maximal tumor size (approximately 2 cm × 2 cm). Mice were then sacrificed and photographed on day 23 after injection, and tumors were collected for further analysis.

Results

Detecting Somatic SVs in WES

In a standard WES protocol (Figure 1A), probes are designed to capture coding exons. The enriched exonic regions are subsequently amplified and subjected to paired-end sequencing. Due to the capturing and amplification steps, the coverage of resulting sequencing data is uneven across the genome. SV detection tools using read-depth information will suffer from this uneven sequencing coverage, whereas tools that depend on discordant read pairs and split reads to detect genomic rearrangements can be used in WES data as long as the breakpoints are captured and sequenced. We first tested the efficacy of detecting somatic SVs using discordant read pairs and split reads but not read depth. We selected 120 TCGA samples that had both WES and WGS data (Table S1) for initial analysis, with the assumption that somatic SVs called on both platforms are true positives (example in Figure 1B). We did not define the truth set purely on the basis of WGS data because some SVs are missed and some SV calls are artifacts even in WGS.

A major challenge in reliably identifying somatic SVs in WES data is to remove a large number of artifacts arising from chimeric molecules in the library preparation. This requires designing data processing steps to remove WES-specific artifacts. When we applied the Meerkat algorithm we originally developed[6] for WGS to WES data, we found a small subset of the samples containing a large number (>100) of somatic SVs, with the majority of SVs not found in the matched WGS (Figure S1A; examples shown in Figures S1B and S1C). WES-specific artifacts were distinguishable by their even distribution across all chromosomes, enrichment of small tandem duplications, and no homology at the breakpoints (Figures S1D–S1F). These samples therefore failed our quality control steps and were discarded from further analysis. For the remaining comparisons, we also removed two WGS cases whose normal data had poor quality (Figure S2).

We designed additional computational filters (see Material and Methods) to remove such artifacts in the remaining samples by testing different combinations of thresholds and comparing the resulting set against WGS calls. This filtration resulted in high-confidence somatic calls from WES data with a substantial reduction in the number of WES-specific calls (Figure S3A). Overall, 61% of the WES calls were shared by WGS (Figure 1C). Many calls found in WGS are missed by WES; out of 145 SVs detected from WGS data with breakpoints in exons (excluding UTRs), 21% (31/145) were recovered from WES data. This low rate is mainly due to the insufficient number of supporting read pairs (Figure S3B) in addition to the uneven read coverage in the targeted regions in WES (Figure S3C). The

Table 1. Summary of Somatic SVs in 15 Tumor Types

Tumor Type	Abbreviation	Sample Size	Bad Samples	Good Samples	Total SVs	Average SVs per Sample	Massively Rearranged
Urothelial bladder cancer	BLCA	185	3	182	370	2.03	6
Breast cancer	BRCA	781	93	688	3123	4.54	65
Glioblastoma multiforme	GBM	318	63	255	626	2.45	24
Head and neck squamous cell carcinoma	HNSC	377	0	377	413	1.10	4
Clear cell kidney carcinoma	KIRC	322	13	309	191	0.62	4
Papillary kidney carcinoma	KIRP	147	0	147	80	0.54	4
Lower grade glioma	LGG	272	0	272	218	0.80	6
Liver hepatocellular carcinoma	LIHC	98	0	98	350	3.57	2
Lung adenocarcinoma	LUAD	485	27	458	791	1.73	12
Lung squamous cell carcinoma	LUSC	460	23	437	837	1.92	9
Prostate adenocarcinoma	PRAD	235	1	234	331	1.41	6
Cutaneous melanoma	SKCM	311	1	310	577	1.86	24
Stomach adenocarcinoma	STAD	234	0	234	570	2.44	11
Papillary thyroid carcinoma	THCA	485	2	483	342	0.71	0
Uterine corpus endometrial carcinoma	UCEC	149	24	125	352	2.82	1
Total	–	4,859	250	4,609	9,171	1.99	178

allele fractions of somatic SVs detected in WES are generally smaller than those in WGS data (Figure S3D). We suspect that the exon capture efficiency is lower for the chimeric DNA molecules that contain the breakpoints, resulting in lower coverage and hence not enough supporting reads for detecting SVs. Conversely, it is important to note that ~39% of the WES calls were not found in WGS. At least a few of these are true positives that were detected by the higher sequencing coverage in WES data than in WGS (Figure 1D and Figure S4). The concordance between WES and WGS calls depends on the quality of the libraries and may vary among datasets.

To test the accuracy of our calls, we performed PCR on all high-confidence somatic SVs called from WES data for which we could obtain the DNA. We found that 78% (21/27) were validated (Table S2). Overall, these results suggest that, despite its modest sensitivity, WES-based SV analysis is likely to yield additional SV candidates that are biologically meaningful.

A Catalog of Gene Fusions and the Properties of Driver Fusions

We analyzed WES data for 4,859 cancer samples across 15 tumor types from TCGA (Table 1). A total of 9,171 high-confidence somatic SVs were detected from 4,609 samples (Table S3) after excluding 250 samples because of low quality. The breast cancers (MIM: 114480) have the highest number of somatic SVs, whereas the kidney cancers (both clear cell [MIM: 144700] and papillary cell [MIM: 605074] carcinomas) have the fewest, consistent with our previous findings[6] (Table 1). The genes with somatic rearrangements are expressed significantly higher (~2-fold increase) than the ones without any rearrangements (Figure S5). Although a previous study[16] associated somatic SV breakpoints with expression, the SV and expression data came from different sets of samples. Here, we used a large number of samples that have each undergone both WES and RNA-seq for a more direct comparison.

Our exome-based SV calling identified many biologically important variants. Some SVs disrupted tumor suppressors, such as *TP53* (MIM: 191170), *CDKN2A* (MIM: 600160), and *PTEN* (MIM: 601728) (Table S4). Many SVs were known driver fusions (examples in Figure 2A). For example, we detected four *RET* (MIM: 164761) fusions (three *CCDC6* [MIM: 601985]-*RET* fusions and one *FKBP15-RET* fusion) in thyroid carcinomas, an *EML4-ALK* fusion in lung adenocarcinoma, and five *FGFR3* (MIM: 134934)-*TACC3* (MIM: 605303) fusions in three cancer types (glioblastoma [GBM], bladder cancer [MIM: 109800], and renal papillary cancer). *FGFR3-TACC3* was originally described in GBM, with 3 out of 97 tumors examined carrying this fusion.[17] This was an important

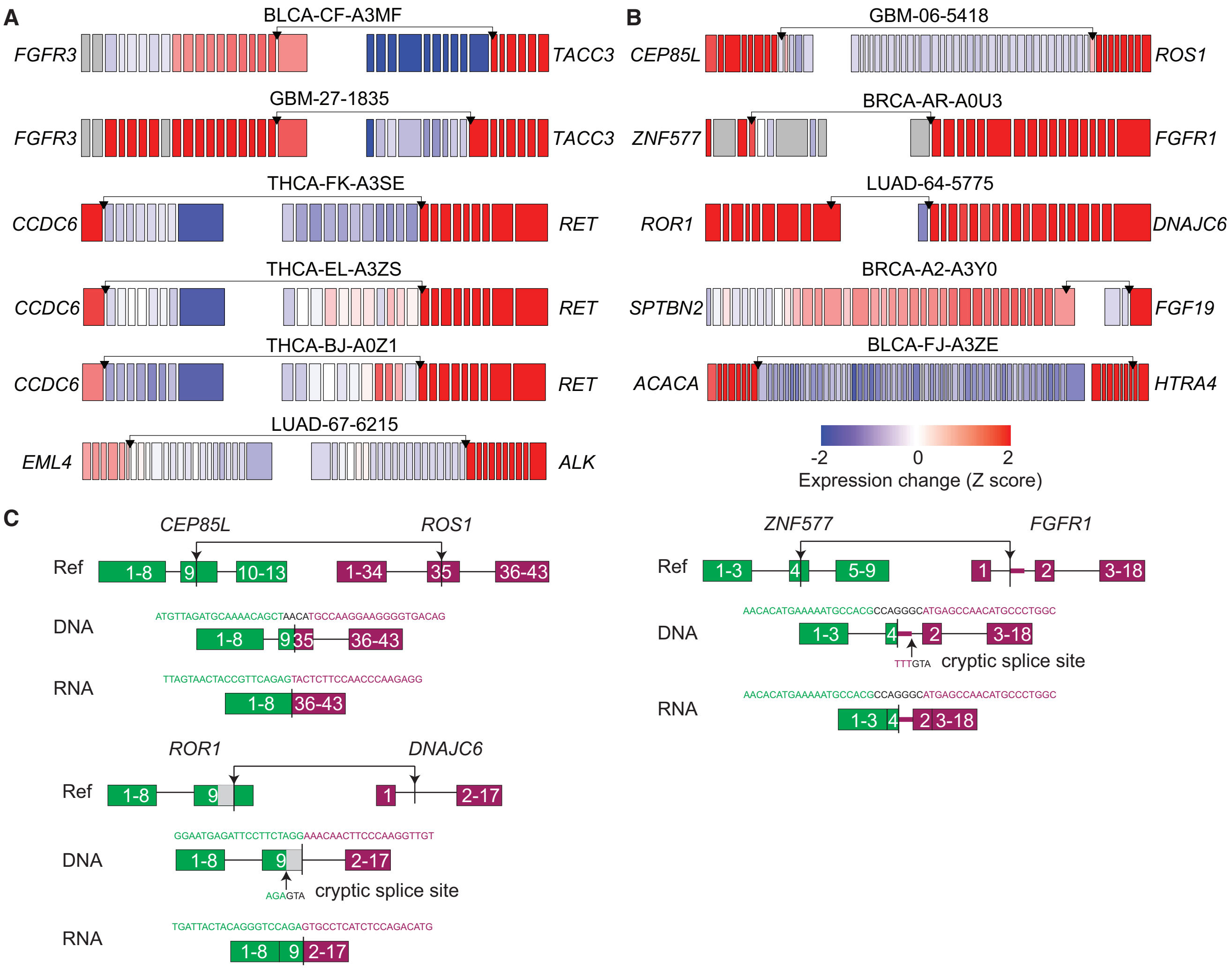

Figure 2. Activating Gene Fusions Detected
(A) Exon-specific expression profiles for known cancer-driving fusions.
(B) Exon-specific expression profiles for additional activating fusions. Black arrows in (A) and (B) denote fusion breakpoints. Each box represents an exon. The expression of each exon was normalized to its average expression across all individuals of the same tumor type. A gray box indicates that the exon is not expressed in more than 70% of the samples.
(C) Examples of fusion breakpoints at the DNA and RNA level for *CEP85L-ROS1*, *ZNF577-FGFR1*, and *ROR1-DNAJC6*. Green and purple boxes denote exons of 5′ and 3′ fusion partners, respectively. Breakpoint junction sequences are shown above the fusions, with letters in black denoting non-reference sequences. The thick purple line in *FGFR1* denotes exonized intronic sequence. The gray box in *ROR1* denotes the part of the exon being spliced out.

discovery because this subset of individuals could potentially benefit from targeted *FGFR* kinase inhibition. We had also found the same fusion in about 3% of the bladder cancer samples, based on analysis of WGS data, as we reported recently in the TCGA consortium paper.[18] Our analysis of the exome data reveals that *FGFR3-TACC3* also occurs in papillary kidney carcinoma. We also detected two prostate adenocarcinoma (MIM: 176807) cases with *TMPRSS2* (MIM: 602060)-*ERG* (MIM: 165080) fusions. As expected, the frequencies of these known drivers are much lower than the previously reported numbers due to limited sensitivity. However, we were able to discover a wide range of variants as a result of the large sample size.

Distinguishing drivers (alterations that increase the fitness of cells) from passengers (neutral alterations) is challenging for any type of genetic alteration. For SNVs and copy-number variants, computational methods (e.g., MutSigCV[19] and GISTIC,[20] respectively) aim to assess the statistical significance of the observed mutation frequencies by using a background model. Recurrence is the most obvious factor in estimating the likelihood of fusion being a driver; however, understanding the molecular characteristics of driver fusion is critical, given that some driver fusions have very low frequency, many studies have small sample sizes, or, as in the case here, detection sensitivity might be low. Furthermore, recurrent events can also result from frequent breaks of certain genomic regions such as fragile sites and might not drive cancer. We previously observed that most of the known driver fusions are activating fusions and that the 3′ fusion partners are

Table 2. Activating Fusions with 3′ Tyrosine Protein Kinases

ID	Chr A	Breakpoint A	Gene A	Chr B	Breakpoint B	Gene B	Discord Pair	Split Read	Homology
THCA-FK-A3SE	10	61655977	*CCDC6*	10	43611997	*RET*	13	17	3
THCA-EL-A3ZS	10	61659539	*CCDC6*	10	43611930	*RET*	4	4	0
THCA-BJ-A0ZJ	10	61626050	*CCDC6*	10	43611953	*RET*	13	5	1
THCA-ET-A3DQ	9	115932783	*FKBP15*	10	43610457	*RET*	5	2	−3
LUAD-67-6215	2	42491894	*EML4*	2	29447037	*ALK*	6	5	2
THCA-EM-A4FR	5	41038833	*MROH2B*	2	29481156	*ALK*	7	7	3
GBM-06-5418	6	118801608	*CEP85L*	6	117642526	*ROS1*	55	46	−4
BRCA-AR-A0U3	19	52383621	*ZNF577*	8	38317439	*FGFR1*	104	29	−7
BRCA-AR-A0TT	19	16243092	*RAB8A*	19	4115139	*MAP2K2*	45	32	0
GBM-06-5411	1	204951828	*NFASC*	1	156844167	*NTRK1*	534	398	2
LGG-E1-5319	1	155784108	*GON4L*	1	156813488	*INSRR*	29	24	1

Genes on the left denoted by "Gene A" are 5′ fusion partners, and genes on the right denoted by "Gene B" are 3′ fusion partners.

almost always upregulated, typically with expression change at the fusion breakpoints[21,22] (Figure 2A). To identify activating gene fusions, we thus propose three criteria: (1) the gene fusion must maintain the same transcription orientation; (2) the fused 3′ partner must be upregulated; and (3) a significant expression change must be observed at or near the fusion breakpoints in at least one of the two source regions (e.g., red versus blue exons on the two sides of the *TACC3* breakpoint in the *FGFR3-TACC3* fusion in Figure 2A). There are driver fusions that do not have an upregulated 3′ partner, but these are hard to identify unless they recur across many samples. Expression change at the breakpoint was also used to identify fusion candidates from expression array data, followed by 5′ rapid amplification of cDNA ends to search for the fusion partners.[23–25] Using the three criteria above, we uncovered a total of 150 activating fusions (Table S5). Five activating fusions (*CEP85L-ROS1*, *ZNF577-FGFR1* [MIM: 136350], *ROR1-DNAJC6*, *SPTBN2* [MIM: 604985]-*FGF19* [MIM: 603891], *ACACA* [MIM: 200350]-*HTRA4* [MIM: 610700]) are shown in Figure 2B as examples. We note that these activating fusions are candidate driver fusions, but the criteria we used are not sufficient to define them as cancer drivers. In vitro and in vivo experiments are needed to definitively address their role in tumorigenesis (see Functional Validation of Fusion Genes In Vitro and In Vivo).

Not surprisingly, GO analysis of the activating fusions revealed that the 3′ fusion partners are enriched for protein tyrosine kinases ($p = 1.7E-4$) (Table 2) as previously observed.[26–28] The protein tyrosine kinases *RET*, *ALK*, and *ROS1* are known oncogenes and often form fusions with various partners in lung (MIM: 211980), thyroid, and colorectal cancers (MIM: 114500)[22,29–33] (e.g., for *RET*: *CCDC6*, *FKBP15*, *TBL1XR1* [MIM: 608628], *AKAP13* [MIM: 604686], *KIF5B* [MIM: 602809]; for *ALK*: *EML4*, *STRN* [MIM: 614765], *GTF2IRD1* [MIM: 604318], *MROH2B*, *C2orf44* [MIM: 616234]; for *ROS1*: *SLC34A2* [MIM: 604217], *CD74* [MIM: 142790], *SDC4* [MIM: 600017], *EZR* [MIM: 123900], *LRIG3* [MIM: 608870]). Some of the kinase fusions detected from WES were known previously. For instance, *NFASC* (MIM: 609145)-*NTRK1* ([MIM: 191315] neurotrophic tyrosine receptor kinase type 1) was found in two TCGA GBM samples via RNA-seq data and validated as a cancer driver.[34] Other fusions identified here were not reported previously: for example, *INSRR* (MIM: 147671), an insulin receptor-related receptor, is paralogous to many oncogenes such as *ROS1*, *NTRK1*, and *ALK*, but has never been described as a fusion partner in cancer even though it is involved in the AKT and MAPK signaling pathways and its expression has been correlated with a favorable prognosis in neuroblastoma.[35] The fusion *GON4L* (MIM: 610393)-*INSRR* found in low-grade glioma activates the protein kinase domain of *INSRR*, suggesting that it is likely to be a driver fusion.

We also found that the 5′ fusion partners of activating fusions are often housekeeping genes, such as those related to the cytoskeleton ($p = 7.4E-5$) and biosynthesis ($p = 2.8E-3$) (Table S6). For example, CCDC6, FKBP15, and EML4 are cytoskeleton proteins that fuse to RET and ALK. Furthermore, both the 5′ fusion partners ($p = 8.5E-3$) and the 3′ fusion partners ($p = 4.9E-3$) of the activating fusions are enriched in chromatin regulators (Tables S7 and S8). Many of the chromatin regulator fusions occur in the breast cancer samples. *USP21* (ubiquitin specific protease 21 [MIM: 604729]), which deubiquitinates histone H2A and removes the transcriptional repression tag, is upregulated in 33% of the breast cancer samples.[36] *KDM2A* (MIM: 605657), a histone demethylase that maintains heterochromatin and genome stability, and *C11orf30* (MIM: 608574), a protein-coding gene that can repress transcription and might play a central role in the DNA-repair function of *BRCA2* (MIM: 600185), are upregulated in 17% and 11% of the breast cancer samples, respectively.[36] The chromatin regulators are upregulated upon fusions and might alter expressions of many other genes and play important roles in tumor progression.

Given the functional categories enriched in the fusion partners, we propose a general model of driver fusions in cancer. The 3′ partners are often oncogenes, which can promote cell growth and proliferation but are typically not expressed in differentiated cells. The 5′ partners are enriched in housekeeping genes, which are expressed in normal cells but whose production is controlled by various mechanisms, including negative feedback loops. Upon fusion, the active housekeeping gene in cancer cells turns on its oncogenic partner. However, because no housekeeping protein is produced, the housekeeping genes remain on. As a result, both the 5′ and 3′ fusion partners are upregulated. In the case of *TMPRSS2-ERG* in prostate cancers (the predominant recurrent aberration in that tumor type), *TMPRSS2* is activated by the androgen receptor and serves as a housekeeping gene in the prostate tissue. The 3′ fusion partners are different ETS family oncogenes (e.g., *ERG*, *ETV1* [MIM: 600541], *ETV4* [MIM: 600711], and *ETV5* [MIM: 601600])[23] that are activated by *TMPRSS2*.

With sequencing data available from both DNA and RNA, it is also possible to interrogate how the fusion genes are spliced. Three cases are shown in Figure 2C: (1) The *CEP85L-ROS1* fusion occurs between exon 9 of *CEP85L* and exon 35 of *ROS1*. The breakpoints at the DNA level are out of frame; however, upon alternative splicing (the fusion exon 9-35 being spliced out), the fusion is in frame at the RNA level. (2) The *ZNF577-FGFR1* fusion is between exon 4 of *ZNF577* and intron 1 of *FGFR1*. A small portion of the *FGFR1* intron becomes part of an exon through a cryptic splice site, and the resulting transcript is in frame. (3) The *ROR1-DNAJC6* fusion is between exon 9 of *ROR1* and intron 1 of *DNAJC6*. After fusion, part of the *ROR1* exon 9 is spliced out through a cryptic splice site along with the intron 1 of *DNAJC6*, resulting in an in-frame transcript. These examples illustrate how alternative splicing and/or cryptic splice sites can be used after gene-fusion events to produce in-frame transcripts even if the fusions are out of frame at the DNA level. Therefore, prediction of functional consequences for gene fusions on the basis of the DNA sequence must account for these mechanisms.

Functional Validation of Fusion Genes In Vitro and In Vivo

We performed extensive in vitro and in vivo validation for two fusions. Various fusions involving the *ROS1* receptor tyrosine kinase have been identified previously, primarily in non-small cell lung cancer (NSCLC),[33] and they are known to induce cell foci formation and anchorage-independent growth.[37,38] The *CEP85L-ROS1* fusion in particular was reported in angiosarcoma and epithelioid hemangioendothelioma,[25] and we found it in GBM in our analysis. However, its function in tumorigenesis has not yet been established. To test the oncogenic potential of this fusion, we utilized Ba/F3, a murine pro-B cell line that depends on interleukin-3 (IL-3) for survival and proliferation. Ba/F3's dependence on IL-3 is readily transferred to expressed oncogenes, thus representing a sensitive assay to quantitate oncogenic activity of fusion genes after Ba/F3 transduction and IL-3 removal from growth medium.[39–41] Introduction of the *CEP85L-ROS1* fusion gene into Ba/F3 cells revealed a robust, >100-fold increase ($p < 0.0001$) in survival after IL3 removal in comparison to GFP-expressing control cells (Figure 3A). Notably, the growth-promoting activity exhibited by *CEP85L-ROS1* was similar to that of *BCR–ABL1*, whose oncogenic activity has been well characterized.[42] Next, we delivered *CEP85L-ROS1* fusion into MCF-10A human breast epithelial cells[43] which are widely used in anchorage-independent growth assays to assess the transforming activity of oncogenes.[44] As shown in Figure 3B, expression of *CEP85L-ROS1* in MCF-10A cells significantly increased colony formation (11-fold, $p < 0.0001$), as did the oncogenic PIK3CA^{H1047R} control.[45] We also found that *CEP85L-ROS1* expression in NIH 3T3 murine fibroblasts induced their anchorage independent growth and cellular proliferation in vitro (Figures S6A and S6B) and tumor-forming activity in vivo (Figures 3C and 3D). Immunoblot analysis showed elevated phosphorylation of ERK1/2 (T202/Y204) in all three cell lines (Ba/F3, MCF-10A, and NIH 3T3; Figures S6C–S6E), which suggested that the MAPK signaling pathway was activated. We tested the effectiveness of this fusion as a drug target. Crizotinib is a small molecular protein kinase inhibitor for *ALK* and *ROS1*. It is approved for use in NSCLC cases with *ALK* fusion, and it has shown great anti-tumor activity in clinical trials targeting advanced NSCLC with a *ROS1* rearrangement.[46] We observed a marked inhibitory activity of crizotinib on *CEP85L-ROS1*-transformed Ba/F3 cells in comparison to parental cells (*CEP85L-ROS1* $IC_{50} = 0.012\ \mu M$; parental $IC_{50} = 0.489\ \mu M$) as shown in Figure 3E. Our results show that individuals harboring a *ROS1* fusion in tumor types other than NSCLC might also benefit from the *ROS1* inhibitor.

Our second candidate fusion for experimental validation was *ROR1-DNAJC6* in lung adenocarcinoma. *ROR1* is a receptor tyrosine kinase that modulates neurite growth in the CNS and might interact with the Wnt signaling pathway.[47] It has not yet been reported as a cancer-driving fusion partner. Our experiments showed that the *ROR1-DNAJC6* fusion can promote in vitro cell proliferation in BEAS-2B cells (non-cancerous human bronchial epithelium; Figures 4A and S7). It can also induce anchorage-independent cell growth (Figures 4B–4D) in both BEAS-2B and NIH 3T3 cells, and promote in vivo tumor formation in mice (Figure 4E) as well. Interestingly, the receptor tyrosine kinase *ROR1* is the 5′ partner in this fusion, in contrast to most other fusions in which protein tyrosine kinases are activated as 3′ fusion partners. Another example with a protein tyrosine kinase on the 5′ side is the *FGFR3-TACC3* fusion,[17] in which *FGFR3* loses its 3′ UTR and escapes from silencing to promote cellular growth.

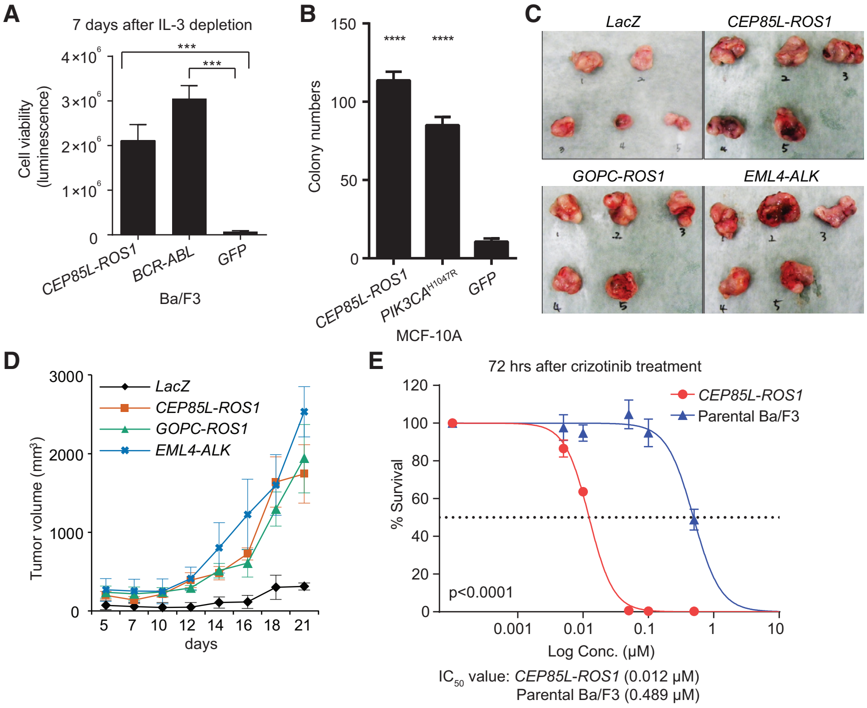

Figure 3. Functional Validation of *CEP85L-ROS1*

(A) *CEP85L-ROS1* expression relieves Ba/F3 cells from dependency on IL-3.

(B) Anchorage-independent colony formation assays for *CEP85L-ROS1* in MCF-10A cells (mean colony count from ten random areas).

(C and D) The transforming potential of the *CEP85L-ROS1* fusion in vivo. The tumor volume was calculated with the modified ellipsoidal formula (volume = 1/2 [length × width^2]) and the greatest longitudinal diameter (length) and the greatest transverse diameter (width) were used. Mice were sacrificed and photographed on day 23.

(E) Compared to parental cells (IC_{50} = 0.489 μM), *CEP85L-ROS1*-transformed Ba/F3 cells are significantly more sensitive (log-rank test) to crizotinib (IC_{50} = 0.012 μM). Error bars indicate SD.

Our results showing the oncogenic potential of these two fusions demonstrate that previously unknown cancer-driving fusions can be detected from WES data, including some that are potential drug targets.

Massive Rearrangements

A small percentage of cancers might have one or more chromosomes massively rearranged, often with copy numbers oscillating between two or three states (chromothripsis),[48,49] segments amplified to many copies (chromoanasynthesis),[6,50] or chains of rearrangements (chromoplexy).[51] These rearrangements have been proposed to form through shattering and rejoining of DNA fragments by non-homologous end joining,[48] pulverization of chromosomes in the micronuclei,[52] and template switching during DNA replication.[6,50] When we searched for chromosomes with statistically significant enrichment of SV breakpoints compared to the rest of the genome by using WES data (taking into account the gene densities on different chromosomes), we found a total of 196 chromosomes in 178 samples (3.8% of 4,609 samples; Table 1, Figure 5A, and Table S9). Our statistical threshold was based on the binomial test with a cutoff of p = 0.01 after the Bonferroni correction (see Material and Methods); given this stringent threshold, the number of samples we report with massively rearranged chromosomes is likely to be an underestimation.

The frequency of massive rearrangements was highly variable across chromosomes (Figure 5A), with up to an ~100-fold difference in the normalized frequencies (e.g., chr17 versus chrX). The highest frequencies were found in chromosomes 17 and 22, consistent with a previous study[53] that found amplification breakpoints to be most frequent on chromosome 17. Different chromosomes were enriched for the SV clusters from different tumor types (Figure 5B). Chromosomes 7 and 12 are enriched for rearrangements in GBMs, and chromosome 22 is enriched for melanomas (MIM: 155600). On chromosome 17, 23 out of 35 occurrences are in breast cancers (examples in Figure 5C), and their breakpoints are highly abundant at the *ERBB2* (MIM: 164870) locus (Figure 5D). Significantly higher copy numbers and expression at the *ERBB2* locus suggest that the massively rearranged chromosome 17 is associated with upregulation of oncogene *ERBB2* (Figure 5E). Those breast cancers with any massively rearranged chromosome, as well as those with massively rearranged chromosome 17 among the HER2+ subtype, have poorer prognosis with marginal statistical significance (p = 0.06 and 0.08, respectively; Figure S8).

There are co-occurrence patterns among the chromosomes that have massive rearrangements. For example, of the nine melanomas with chromosome 22 rearrangements, seven involve other chromosomes, including five involving chromosome 5 (Figures 5F and S9). Conversely, there are three melanomas with massively rearranged chromosome 5, and all of them co-occur with massively rearranged chromosome 22 (Table S9). In melanoma cases, it is known that ~70% have *TERT* ([MIM: 187270] on chromosome 5) upregulated by promoter mutations.[54,55] We found that the individuals with massively rearranged chromosome 22 have significantly higher expression of *TERT* when chromosome 5 is also involved (Figure 5G). In GBM, *CDK4* (MIM: 123829) is often amplified and expressed at a significantly higher amount in individuals with massively rearranged chromosome 12 (Figure S10A). On the other hand, the expression of *EGFR* (MIM: 131550) is not significantly

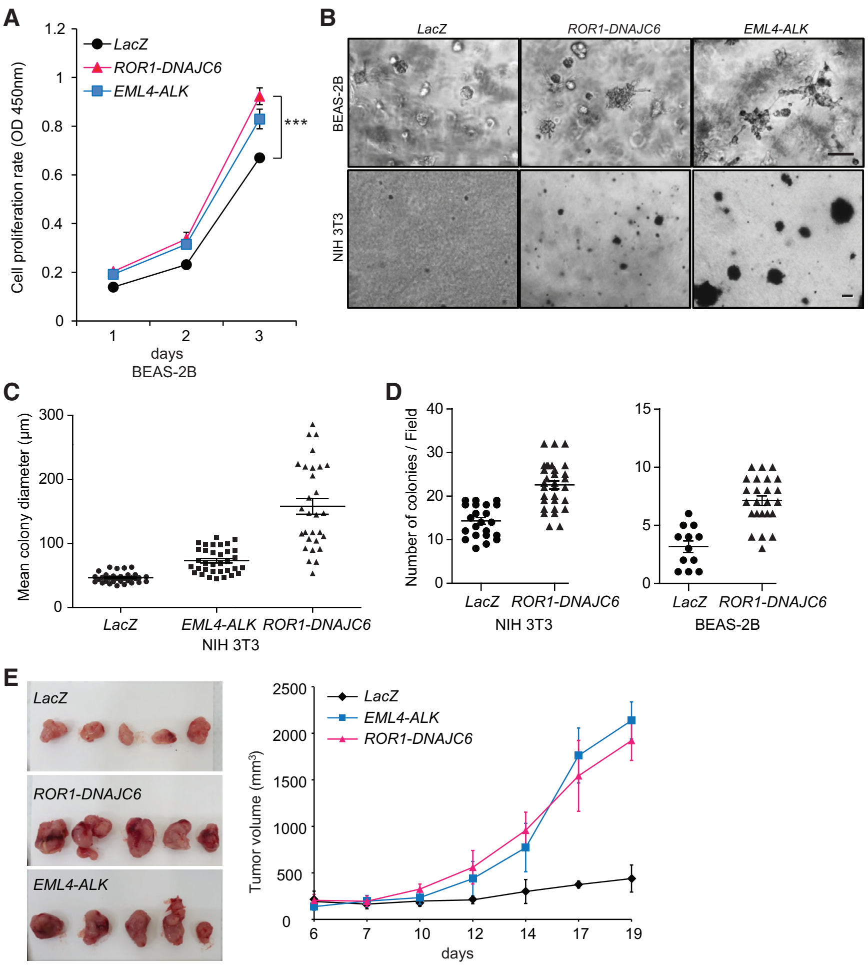

Figure 4. Functional Validation of *ROR1-DNAJC6*

(A) Growth rate of cells expressing ROR1-DNAJC6 fusion protein in BEAS-2B cells.
(B) BEAS-2B cells cultured in Matrigel after 7 days and NIH 3T3 cells cultured in soft agar after 14 days expressing ROR1-DNAJC6 fusion protein. Scale bar, 50 μm.
(C and D) Anchorage independent growth in soft agar. BEAS-2B or NIH 3T3 cells transformed with *ROR-DNAJC6* were cultured in soft agar for 21 days.
(E) The transforming potential of *ROR1-DNAJC6* fusion in vivo.

different in individuals with massively rearranged chromosome 7 because the ones without massively rearranged chromosome 7 also have *EGFR* amplifications (Figure S10B). This is consistent with our previous study[6] showing that most (14 out of 16) GBM samples have *EGFR* amplified and that some of the amplifications are achieved through very complex rearrangements. These results suggest that massive rearrangements are often associated with upregulation of oncogenes, which provides selective advantage to the cells, and these rearrangements are thus maintained in the genome.

Discussion

Here, we report the somatic genome rearrangements detected in the WES data for nearly 5,000 human cancer samples. WES data present challenges for SV identification, with ligation artifacts formed during exome capture and/or DNA amplification steps often manifesting as small tandem duplications. Many of the samples we excluded on the basis of quality were whole-genome amplified (WGA) samples, but other WGA samples did not suffer from the same problem. Although it is not possible to determine whether a specific tandem duplication is a true or artifactual one, their genome-wide distribution is strongly indicative of the sample data quality. The large number of samples with both WES and WGS data allowed us to set proper filtering thresholds.

SV identification based on WES data has much lower sensitivity than that based on WGS data. Therefore, it is not sensible to generate WES data to profile SVs or to replace WGS with WES. Our goal here was to reanalyze existing WES data, given that the number of samples with WES data is larger than that with WGS by an order of magnitude. In TCGA, for example, almost all of the samples were profiled by WES, whereas about 10% of the cases were profiled by WGS.

The number of exomes sequenced will continue to grow, especially as we search for somatic mutations with low variant allelic frequency. For instance, some somatic driver SNVs in cancer have been shown to occur in <5% of the cells. In neuroscience, there is now a great deal of interest in identifying somatic mutations in the brain to potentially explain neurological diseases such as epilepsy and developmental brain malformations.[56] For such variants, high-coverage WES will be the preferred platform for most investigators until WGS at very high coverage becomes more affordable. Identification of even a fraction of the SVs in these datasets will be valuable. As we showed in one example (Figure 1D), somatic SV with low variant allele frequency cannot be detected by WGS as a result of its much lower coverage than WES. Importantly, the framework we described here is also applicable to germline rearrangements, and the number of germline exomes from individuals with a variety of disease phenotypes as well as from healthy individuals is already enormous.

As another application of exome-based SV analysis, we investigated massive rearrangements in our cohort and found that WES data can capture the presence of these events and their association with other factors. Because

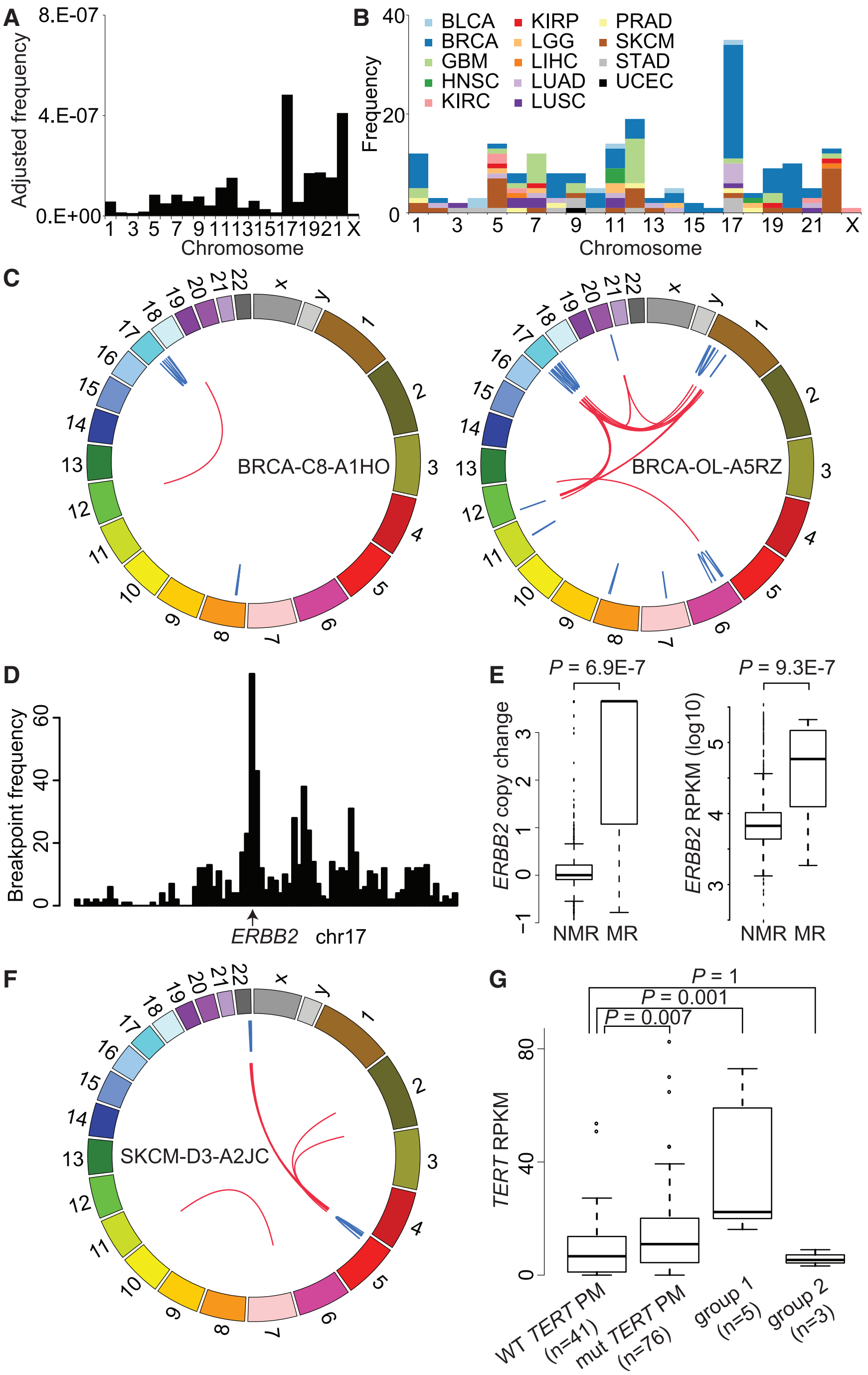

Figure 5. Massive Rearrangements Are Often Associated with Upregulation of Oncogenes

(A) The frequencies of massively rearranged chromosomes normalized by the uniquely mappable size of CDS in each chromosome.

(B) The frequencies of massively rearranged chromosomes colored by tumor type.

(C) Examples of two breast cancers with massively rearranged chr17. Blue and red lines denote intra-chromosomal and inter-chromosomal rearrangements, respectively.

(D) The breakpoint distribution of massively rearranged chr17 of breast cancers with the peak at *ERBB2*.

(E) Association (Wilcoxon one-side rank test) of massive rearrangements with copy change and expression of *ERBB2* in breast cancers. NMR, not massively rearranged; MR, massively rearranged. Error bars indicate SD.

(F) An example of massively rearranged chr22 that involves chr5 in melanoma.

(G) Association (Wilcoxon one-side rank test) of massively rearranged chr22 with *TERT* expression. Group 1 includes melanomas with massively rearranged chr22 that involves chr5. Group 2 includes melanomas with massively rearranged chr22 that does not involve chr5 with wild-type *TERT* promoter. Error bars indicate SD.

these events are rare (~4% of the cases), their enrichment in specific chromosomes or tumor types, as well as their correlations with copy number and gene expression, became apparent with a large sample size (hundreds of samples per tumor type). Our finding that massive rearrangements are often associated with oncogene upregulation would not have been possible from WGS data. Copy-number profiles from microarray have been used to detect chromothripsis events on the basis of oscillating copy numbers on one or more chromosomes,[48] including in our own work.[57] However, interchromosomal events cannot be detected from array profiles, and the association between chromosome 22 massive rearrangements and upregulation of *TERT* could only be detected with WES data. Overall, our study of somatic genome rearrangements utilizing WES data provides insights into how gene fusions drive cancer and demonstrates the utility of re-analyzing existing data.

Supplemental Data

Supplemental Data include ten figures and nine tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.03.017.

Acknowledgments

The results published here are in whole or in part based upon data generated by The Cancer Genome Atlas (TCGA) project established by the National Cancer Institute and the National Human Genome Research Institute. Information about the TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov. This work was supported by the Harvard Ludwig Center (P.J.P.), NIH grant U24CA144025 (R.K. and P.J.P.), National

Research Foundation grants NRF-2013M3C8A1078501, NRF-2013R1A2A2A01068922 (Y.L.C.) and 2015R1C1A1A02037066 (M.L.) funded by the Korean Ministry of Science, Information and Communications Technology & Future Planning, Cancer Prevention and Research Institute of Texas grants RP140102 (H.L.) and RP120046 (K.L.S), and NIH grant U01CA168394 (K.L.S). This work made use of the Bionimbus Protected Data Cloud.

Received: September 30, 2015
Accepted: March 14, 2016
Published: May 5, 2016

Web Resources

Bionimbus, http://bionimbus.opensciencedatacloud.org
CGhub, https://cghub.ucsc.edu/
DAVID, http://david.abcc.ncifcrf.gov/
Genome Data Analysis Center (GDAC), http://gdac.broadinstitute.org
Meerkat software, http://compbio.med.harvard.edu/Meerkat
OMIM, http://www.omim.org/
The Cancer Genome Atlas Data Portal, https://tcga-data.nci.nih.gov/tcga/
UCSC Genome Browser, http://genome.ucsc.edu

References

1. Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. *21*, 974–984.
2. Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods *6*, 677–681.
3. Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871.
4. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat. Genet. *44*, 226–232.
5. Wang, J., Mullighan, C.G., Easton, J., Roberts, S., Heatley, S.L., Ma, J., Rusch, M.C., Chen, K., Harris, C.C., Ding, L., et al. (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat. Methods *8*, 652–654.
6. Yang, L., Luquette, L.J., Gehlenborg, N., Xi, R., Haseley, P.S., Hsieh, C.H., Zhang, C., Ren, X., Protopopov, A., Chin, L., et al. (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. Cell *153*, 919–929.
7. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al.; NHLBI Exome Sequencing Project (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216–220.
8. Chmielecki, J., Crago, A.M., Rosenberg, M., O'Connor, R., Walker, S.R., Ambrogio, L., Auclair, D., McKenna, A., Heinrich, M.C., Frank, D.A., and Meyerson, M. (2013). Whole-exome sequencing identifies a recurrent NAB2-STAT6 fusion in solitary fibrous tumors. Nat. Genet. *45*, 131–132.
9. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.
10. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
11. Rozen, S., and Skaletsky, H. (1999). Primer3 on the WWW for general users and for biologist programmers. In Bioinformatics Methods And Protocols, S. Misener and S.A. Krawetz, eds. (Springer), pp. 365–386.
12. Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. *38*.
13. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.
14. Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.
15. Debnath, J., Muthuswamy, S.K., and Brugge, J.S. (2003). Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. Methods *30*, 256–268.
16. Drier, Y., Lawrence, M.S., Carter, S.L., Stewart, C., Gabriel, S.B., Lander, E.S., Meyerson, M., Beroukhim, R., and Getz, G. (2013). Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res. *23*, 228–235.
17. Singh, D., Chan, J.M., Zoppoli, P., Niola, F., Sullivan, R., Castano, A., Liu, E.M., Reichel, J., Porrati, P., Pellegatta, S., et al. (2012). Transforming fusions of FGFR and TACC genes in human glioblastoma. Science *337*, 1231–1235.
18. Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature *507*, 315–322.
19. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.
20. Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S., et al. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc. Natl. Acad. Sci. USA *104*, 20007–20012.
21. Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550.
22. Cancer Genome Atlas Research Network (2014). Integrated genomic characterization of papillary thyroid carcinoma. Cell *159*, 676–690.
23. Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., et al. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science *310*, 644–648.
24. Wang, L., Motoi, T., Khanin, R., Olshen, A., Mertens, F., Bridge, J., Dal Cin, P., Antonescu, C.R., Singer, S., Hameed, M., et al. (2012). Identification of a novel, recurrent

HEY1-NCOA2 fusion in mesenchymal chondrosarcoma based on a genome-wide screen of exon-level expression data. Genes Chromosomes Cancer *51*, 127–139.

25. Giacomini, C.P., Sun, S., Varma, S., Shain, A.H., Giacomini, M.M., Balagtas, J., Sweeney, R.T., Lai, E., Del Vecchio, C.A., Forster, A.D., et al. (2013). Breakpoint analysis of transcriptional and genomic profiles uncovers novel gene fusions spanning multiple human cancer types. PLoS Genet. *9*, e1003464.
26. Mitelman, F., Johansson, B., and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. Nat. Rev. Cancer *7*, 233–245.
27. Yoshihara, K., Wang, Q., Torres-Garcia, W., Zheng, S., Vegesna, R., Kim, H., and Verhaak, R.G. (2015). The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene *34*, 4845–4854.
28. Stransky, N., Cerami, E., Schalm, S., Kim, J.L., and Lengauer, C. (2014). The landscape of kinase fusions in cancer. Nat. Commun. *5*, 4846.
29. Lipson, D., Capelletti, M., Yelensky, R., Otto, G., Parker, A., Jarosz, M., Curran, J.A., Balasubramanian, S., Bloom, T., Brennan, K.W., et al. (2012). Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. Nat. Med. *18*, 382–384.
30. Soda, M., Choi, Y.L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., et al. (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature *448*, 561–566.
31. Kohno, T., Ichikawa, H., Totoki, Y., Yasuda, K., Hiramoto, M., Nammo, T., Sakamoto, H., Tsuta, K., Furuta, K., Shimada, Y., et al. (2012). KIF5B-RET fusions in lung adenocarcinoma. Nat. Med. *18*, 375–377.
32. Ju, Y.S., Lee, W.-C., Shin, J.-Y., Lee, S., Bleazard, T., Won, J.-K., Kim, Y.T., Kim, J.-I., Kang, J.-H., and Seo, J.-S. (2012). A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. Genome Res. *22*, 436–445.
33. Takeuchi, K., Soda, M., Togashi, Y., Suzuki, R., Sakata, S., Hatano, S., Asaka, R., Hamanaka, W., Ninomiya, H., Uehara, H., et al. (2012). RET, ROS1 and ALK fusions in lung cancer. Nat. Med. *18*, 378–381.
34. Kim, J., Lee, Y., Cho, H.-J., Lee, Y.-E., An, J., Cho, G.-H., Ko, Y.-H., Joo, K.M., and Nam, D.-H. (2014). NTRK1 fusion in glioblastoma multiforme. PLoS ONE *9*, e91940.
35. Weber, A., Huesken, C., Bergmann, E., Kiess, W., Christiansen, N.M., and Christiansen, H. (2003). Coexpression of insulin receptor-related receptor and insulin-like growth factor 1 receptor correlates with enhanced apoptosis and dedifferentiation in human neuroblastomas. Clin. Cancer Res. *9*, 5683–5692.
36. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1.
37. Davies, K.D., and Doebele, R.C. (2013). Molecular pathways: ROS1 fusion proteins in cancer. Clin. Cancer Res. *19*, 4040–4045.
38. Shaw, A.T., Hsu, P.P., Awad, M.M., and Engelman, J.A. (2013). Tyrosine kinase gene rearrangements in epithelial malignancies. Nat. Rev. Cancer *13*, 772–787.
39. Warmuth, M., Kim, S., Gu, X.J., Xia, G., and Adrián, F. (2007). Ba/F3 cells and their use in kinase drug discovery. Curr. Opin. Oncol. *19*, 55–60.
40. Liang, H., Cheung, L.W., Li, J., Ju, Z., Yu, S., Stemke-Hale, K., Dogruluk, T., Lu, Y., Liu, X., Gu, C., et al. (2012). Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer. Genome Res. *22*, 2120–2129.
41. Grubbs, E.G., Ng, P.K., Bui, J., Busaidy, N.L., Chen, K., Lee, J.E., Lu, X., Lu, H., Meric-Bernstam, F., Mills, G.B., et al. (2015). RET fusion as a novel driver of medullary thyroid carcinoma. J. Clin. Endocrinol. Metab. *100*, 788–793.
42. Daley, G.Q., Van Etten, R.A., and Baltimore, D. (1990). Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. Science *247*, 824–830.
43. Soule, H.D., Maloney, T.M., Wolman, S.R., Peterson, W.D., Jr., Brenz, R., McGrath, C.M., Russo, J., Pauley, R.J., Jones, R.F., and Brooks, S.C. (1990). Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. Cancer Res. *50*, 6075–6086.
44. Shin, S.-I., Freedman, V.H., Risser, R., and Pollack, R. (1975). Tumorigenicity of virus-transformed cells in nude mice is correlated specifically with anchorage independent growth in vitro. Proc. Natl. Acad. Sci. USA *72*, 4435–4439.
45. Isakoff, S.J., Engelman, J.A., Irie, H.Y., Luo, J., Brachmann, S.M., Pearline, R.V., Cantley, L.C., and Brugge, J.S. (2005). Breast cancer-associated PIK3CA mutations are oncogenic in mammary epithelial cells. Cancer Res. *65*, 10992–11000.
46. Shaw, A.T., Ou, S.-H.I., Bang, Y.-J., Camidge, D.R., Solomon, B.J., Salgia, R., Riely, G.J., Varella-Garcia, M., Shapiro, G.I., Costa, D.B., et al. (2014). Crizotinib in ROS1-rearranged non-small-cell lung cancer. N. Engl. J. Med. *371*, 1963–1971.
47. Green, J.L., Kuntz, S.G., and Sternberg, P.W. (2008). Ror receptor tyrosine kinases: orphans no more. Trends Cell Biol. *18*, 536–544.
48. Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27–40.
49. Maher, C.A., and Wilson, R.K. (2012). Chromothripsis and human disease: piecing together the shattering process. Cell *148*, 29–32.
50. Liu, P., Erez, A., Nagamani, S.C., Dhar, S.U., Kołodziejska, K.E., Dharmadhikari, A.V., Cooper, M.L., Wiszniewska, J., Zhang, F., Withers, M.A., et al. (2011). Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell *146*, 889–903.
51. Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., et al. (2013). Punctuated evolution of prostate cancer genomes. Cell *153*, 666–677.
52. Crasta, K., Ganem, N.J., Dagher, R., Lantermann, A.B., Ivanova, E.V., Pan, Y., Nezi, L., Protopopov, A., Chowdhury, D., and Pellman, D. (2012). DNA breaks and chromosome pulverization from errors in mitosis. Nature *482*, 53–58.
53. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhsng, C.Z., Wala, J., Mermel, C.H., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet. *45*, 1134–1140.

54. Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science *339*, 957–959.

55. Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. Science *339*, 959–961.

56. Poduri, A., Evrony, G.D., Cai, X., and Walsh, C.A. (2013). Somatic mutation, genomic variation, and neurological disease. Science *341*, 1237758.

57. Kim, T.-M., Xi, R., Luquette, L.J., Park, R.W., Johnson, M.D., and Park, P.J. (2013). Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. Genome Res. *23*, 217–227.

The American Journal of Human Genetics, Volume *98*

Supplemental Data

Analyzing Somatic Genome Rearrangements in Human Cancers by Using Whole-Exome Sequencing

Lixing Yang, Mi-Sook Lee, Hengyu Lu, Doo-Yi Oh, Yeon Jeong Kim, Donghyun Park, Gahee Park, Xiaojia Ren, Christopher A. Bristow, Psalm S. Haseley, Soohyun Lee, Angeliki Pantazi, Raju Kucherlapati, Woong-Yang Park, Kenneth L. Scott, Yoon-La Choi, and Peter J. Park

Table of contents

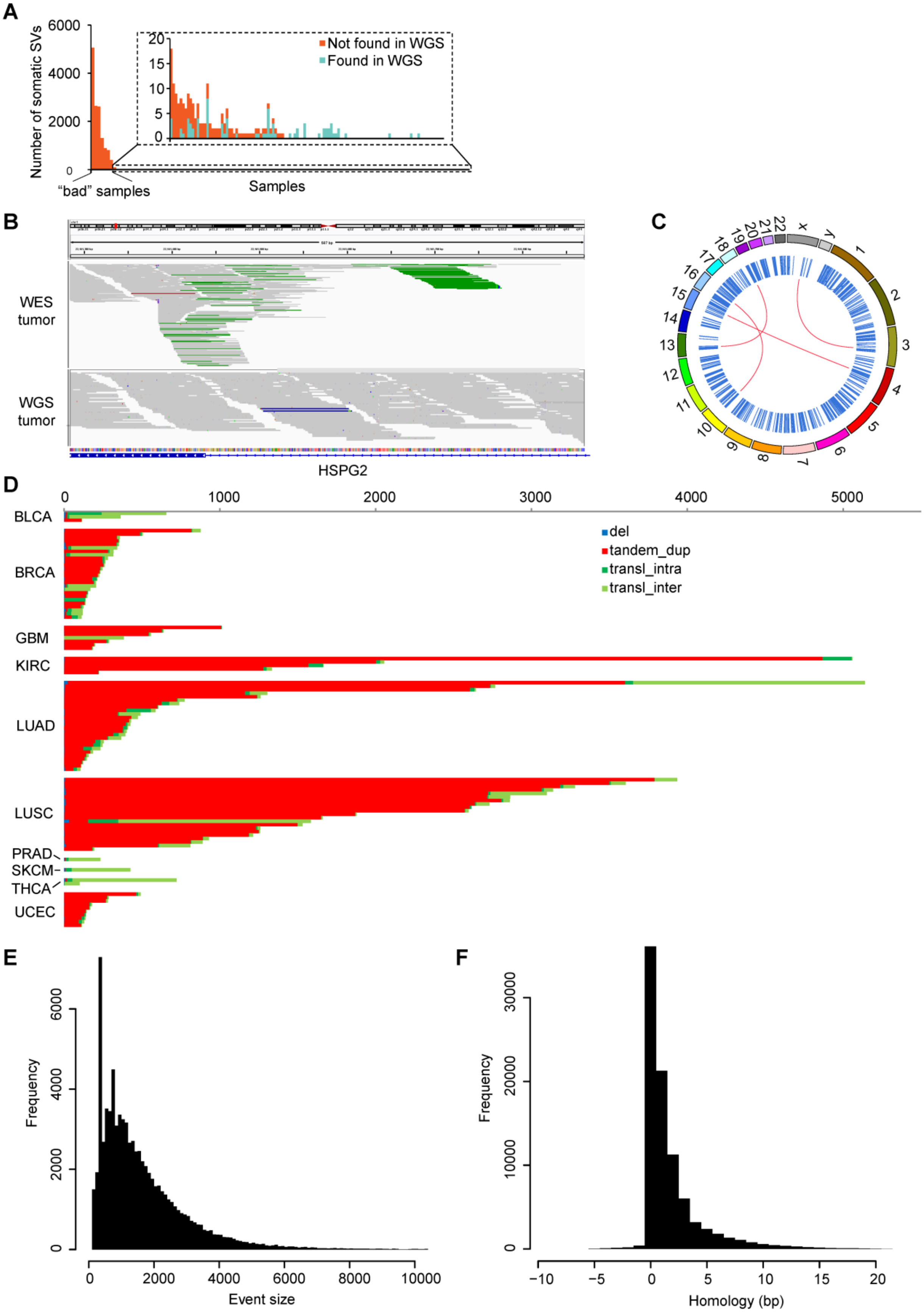
A
Number of somatic SVs
6000
4000
2000
0
"bad" samples
Samples
20
15
10
5
0
Not found in WGS
Found in WGS
B
WES tumor
WGS tumor
HSPG2
C
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
D
0
1000
2000
3000
4000
5000
BLCA
BRCA
GBM
KIRC
LUAD
LUSC
PRAD
SKCM
THCA
UCEC
del
tandem_dup
transl_intra
transl_inter
E
Frequency
6000
4000
2000
0
0
2000
4000
6000
8000
10000
Event size
F
Frequency
30000
20000
10000
0
−10
−5
0
5
10
15
20
Homology (bp)

Figure S1. WES-specific artifacts. A, Comparisons of WES and WGS SV calls. A small number of low-quality samples have an unusually large number of WES-specific somatic SVs. **B**, An IGV screen shot for one artifact. The green lines in the top panel (WES) denote discordant read pairs supporting a tandem duplication; such discordant read pairs are not observed in the bottom panel (WGS data for the same patient). **C**, A Circos plot showing artifacts evenly distributed across all chromosomes. The red lines denote inter-chromosomal events and the blue lines denote intra-chromosomal events. **D**, The number and the type of SVs across a large number of patients, with each horizontal line corresponding to a sample. Artifacts are enriched for tandem duplications. **E** and **F**, Histograms of event size and homology distribution for artifacts, respectively. A negative number for sequence homology corresponds to the size of an insertion.

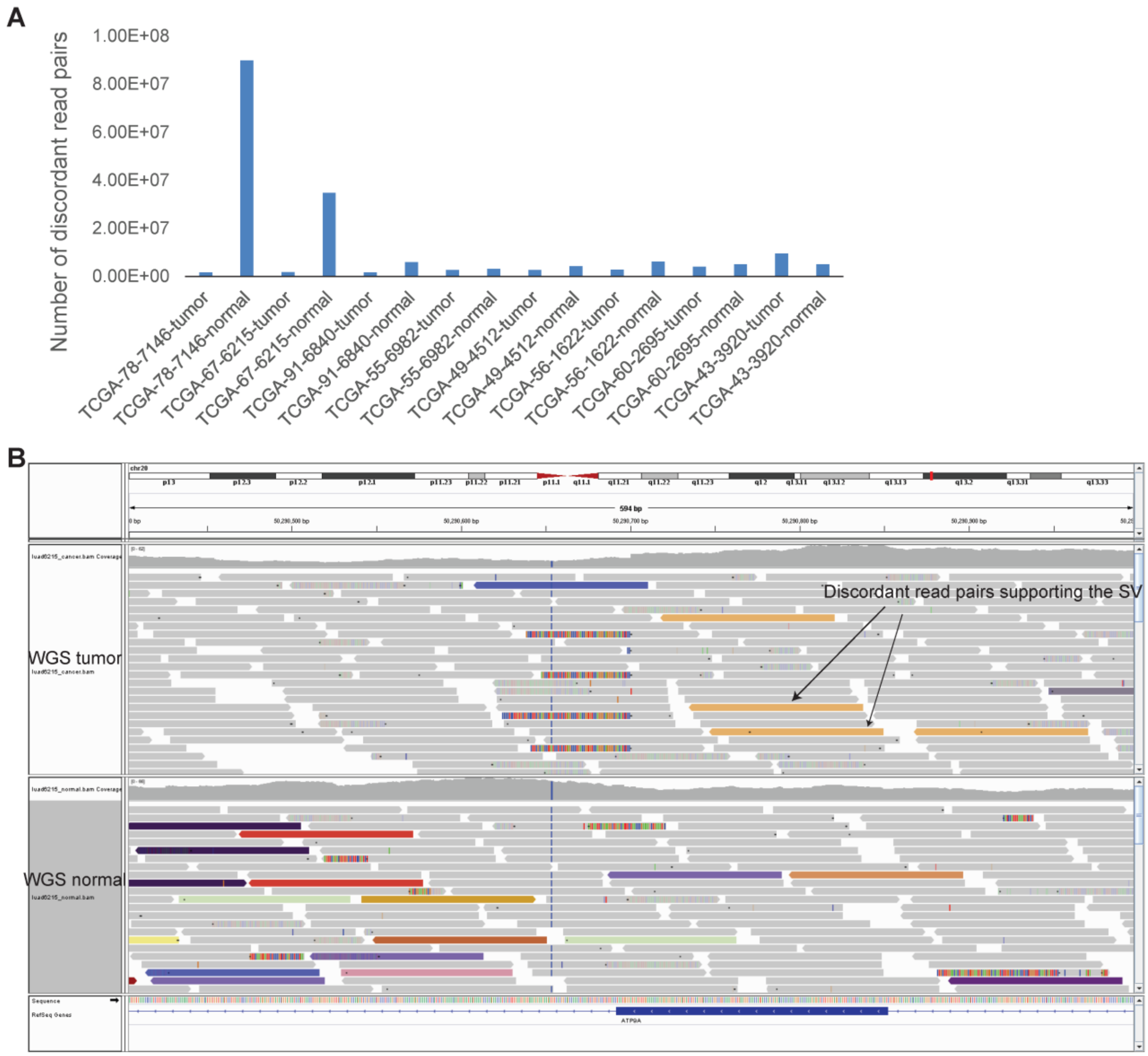

Figure S2. Noises in WGS samples. A, Number of discordant read pairs in WGS samples. Eight tumor-normal pairs are shown. Some normal samples have excessive discordant read pairs (TCGA-78-7146 and TCGA-67-6215). These discordant pairs are generated from library construction and sequencing, rather than from real SVs in the genome. **B**, The top and bottom panels show the read-level view of WGS tumor and matched normal samples. The orange bars are discordant reads supporting the somatic SV and the bars with different colors in bottom panel are discordant read pairs that the mate are mapped to different chromosomes (chromosomes indicated by color). The event was filtered in WGS because of the many discordant read pairs present in the match normal sample.

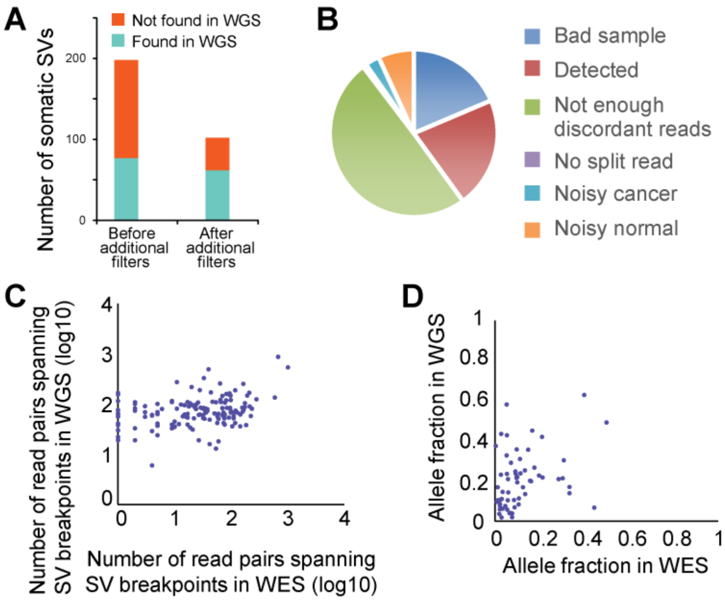

Figure S3. Comparison of somatic rearrangements detected from WES and WGS. A, Comparison of somatic rearrangements detected from WES before and after additional filters (with poor-quality samples excluded) shows that the fraction of WES-specific calls is substantially reduced. **B**, The "catchable" somatic SVs detected from WGS data (with breakpoints in exons excluding UTRs). About one-fifth of the SVs are detected by WES but the rest are missed for the reasons listed. "Bad sample" refers to the events being in the sample with >100 somatic SVs detected, and therefore, such sample was subsequently discarded from further analysis. "Noisy cancer" and "Noisy normal" refer to the SVs in which the algorithm did not make a call because of increased noise in the data at the SV location, as reflected in, e.g., aberrant discordant read pairs (see Fig. S2). **C**, Coverage comparison of WES and WGS. The number of read pairs spanning an SV breakpoint in WES and WGS is equivalent to its physical coverage. Each dot is a somatic SV detected from WGS that is also detectable in WES (in exons excluding UTRs). A portion of breakpoints detected in WGS have 10-100x physical coverage in WGS but with <10x coverage in WES. There are also 12 loci with no read pair spanning breakpoint in WES data and are not represented in this plot. **D**, Allele fractions of somatic SVs that are shared by WES and WGS.

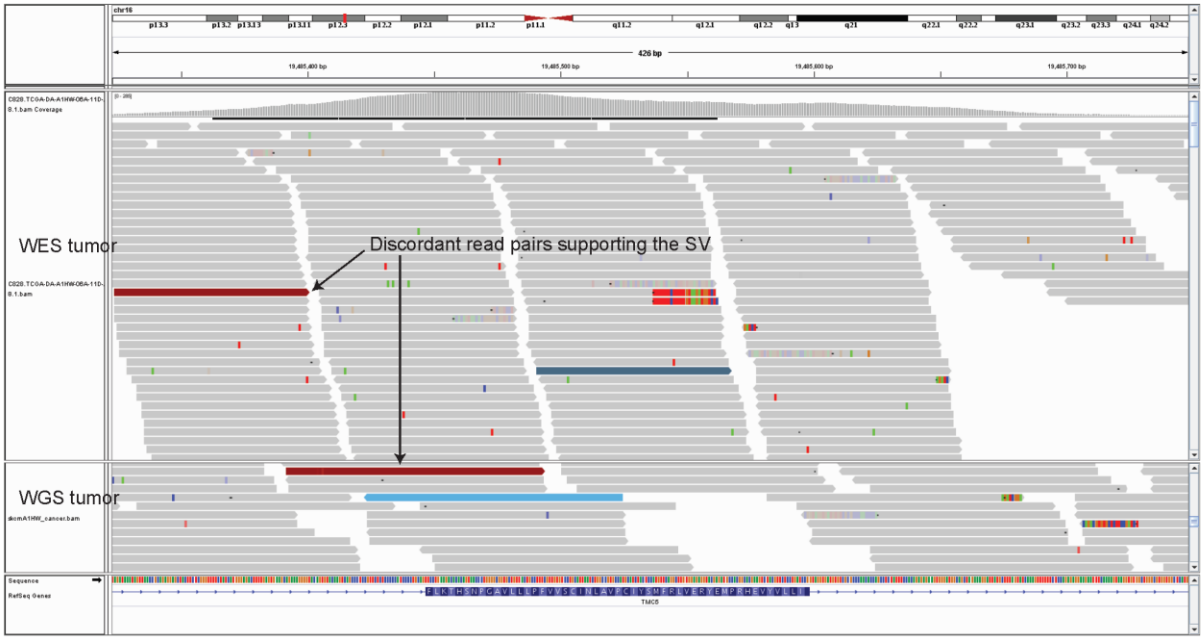

Figure S4. Example of an SV detected in WES but not in WGS due to higher coverage of WES. A somatic deletion in melanoma TCGA-DA-A1HW (chr16:19485535-19690623) is detected from WES but not in WGS. The coverage in WES is 300x and there are 4 discordant read pairs (only 1 is displayed). In contrast, the coverage of WGS in the same region is 90x with only 1 discordant read pair present. The event was validated as somatic by PCR.

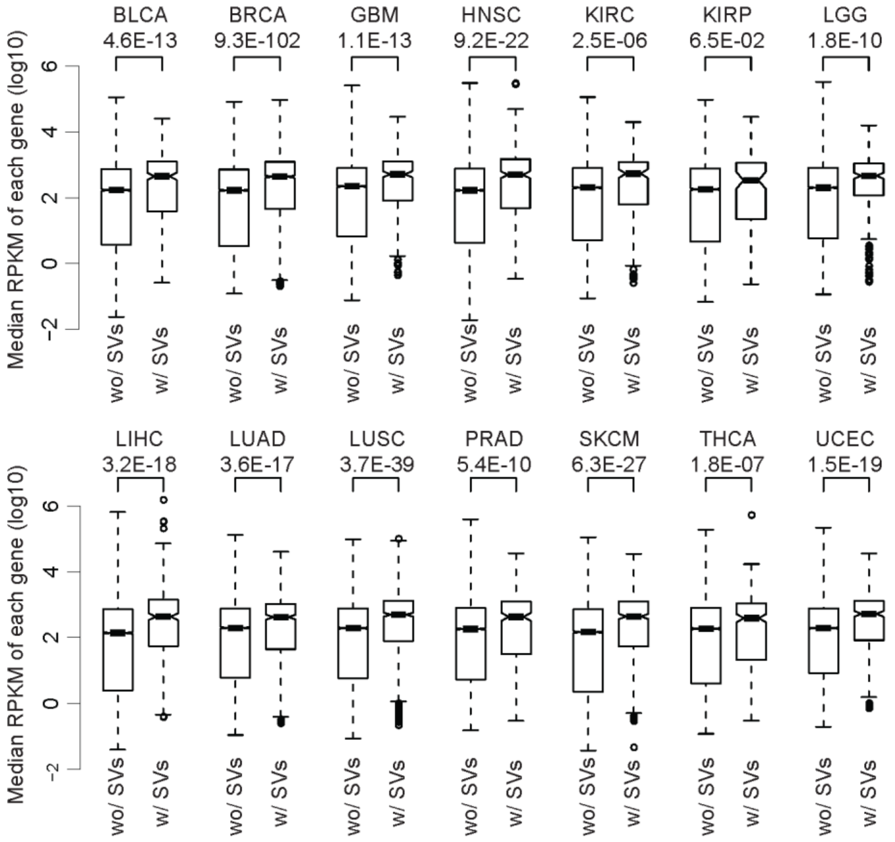

Figure S5. The expression of genes with and without somatic SVs for 14 tumor types. The *P* values by Wilcoxon one-side rank test are shown below the tumor type names. The median expression level of each gene across all individuals for one tumor type was plotted. This shows somatic SVs occur in relatively highly expressed genes.

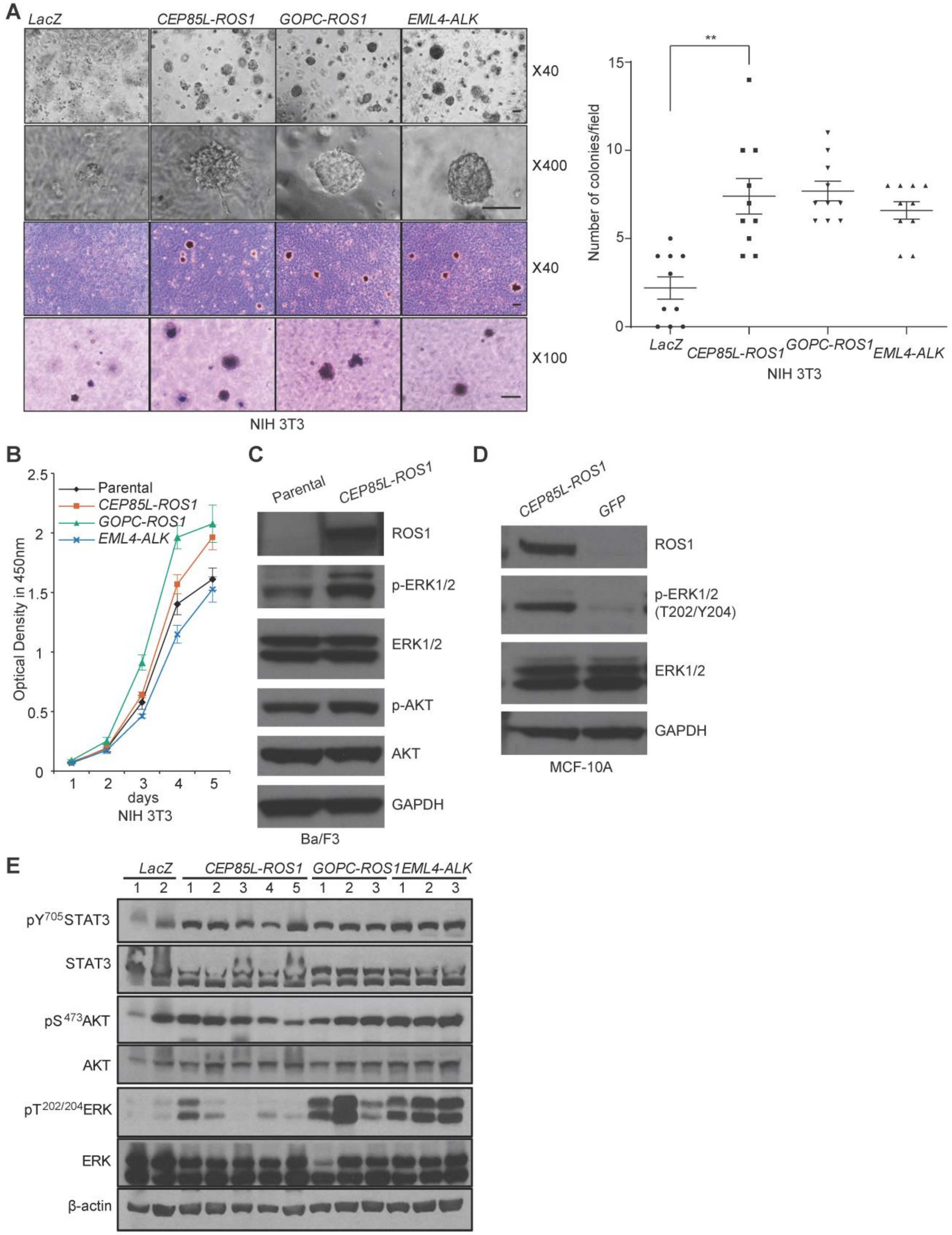

Figure S6. Functional validation for *CEP85L-ROS1* fusion. A, *In vitro* transforming assay of NIH 3T3 cells. Cells expressing the indicated fusion genes were cultured in Matrigel (upper left panels) for 14 days or soft agar (lower left panels) for 21 days. For visualization, some colonies

were stained with 0.05% crystal violet. Images were taken by a phase-contrast microscope and the colonies were counted at 40X magnification. Scale bar, 50μm. Each dot in the right panel represents total colony number in a unit microscopic field. ** denote $P<0.01$ with Wilcoxon one-side rank test. **B**, NIH 3T3 cell growth rate. The values represent the average of three determinations, and the error bars indicate standard deviations. **C**, *CEP85L-ROS1* strongly activates ERK1/2 but not AKT in Ba/F3 cell line by western blots. **D**, Immunoblots of *CEP85L-ROS1* expression and MAPK pathway activation in MCF-10A cells. **E**, Activation of Erk and STAT3 but not AKT in solid tumors from nude mice by western blots.

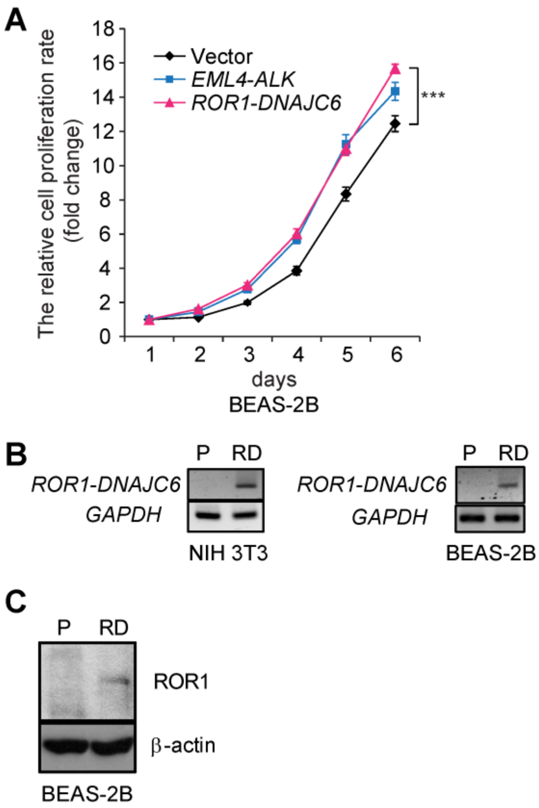

Figure S7. Functional validation for *ROR1-DNAJC6* fusion. A, Growth rate of BEAS-2B cells expressing *ROR1-DNAJC6* fusion measured by optical density at 450 nm (OD450). **B**, Identification of *ROR1-DNAJC6* mRNA expression using RT-PCR in NIH 3T3 and BEAS-2B cells infected with *ROR1-DNAJC6* lentivirus. **C**, Immunoblot using an anti-ROR1 antibody on lysates from the BEAS-2B stable cells expressing ROR1-DNAJC6 fusion protein.

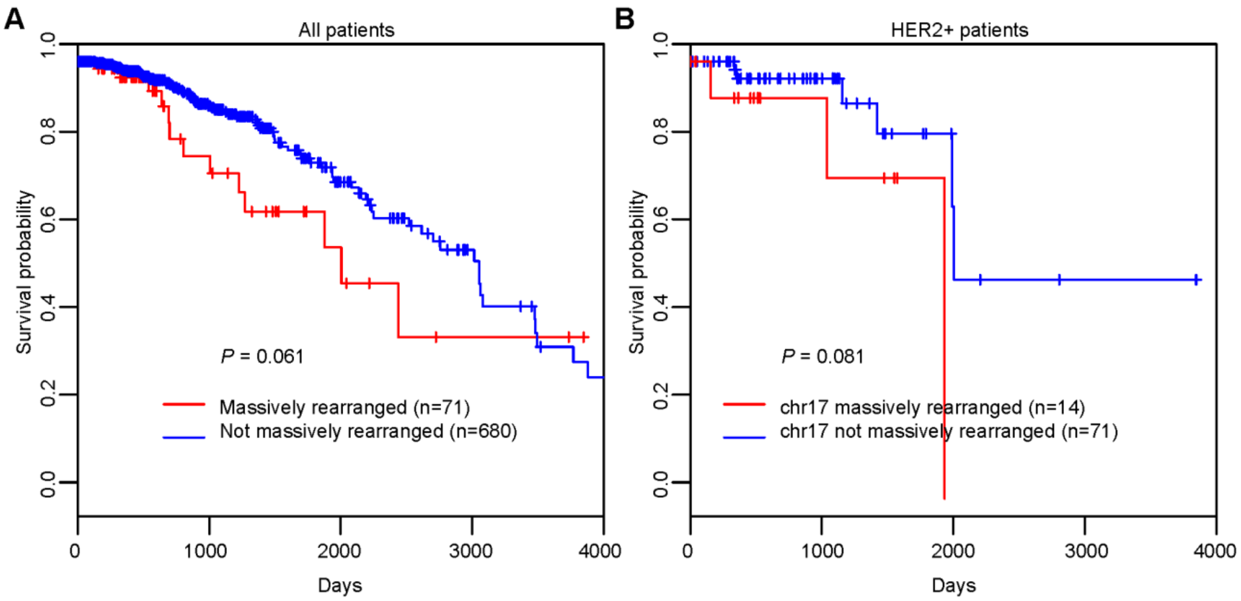

Figure S8. Survival plots for breast cancer patients with and without massive rearrangements. A, All individuals were divided between those with and without rearrangements. **B**, The HER2+ subgroup was divided between those with and without massively-rearranged chromosome 17. The p-values, computed using the log-rank test, were marginally significant in both cases (*P* = 0.061 and 0.081, respectively).

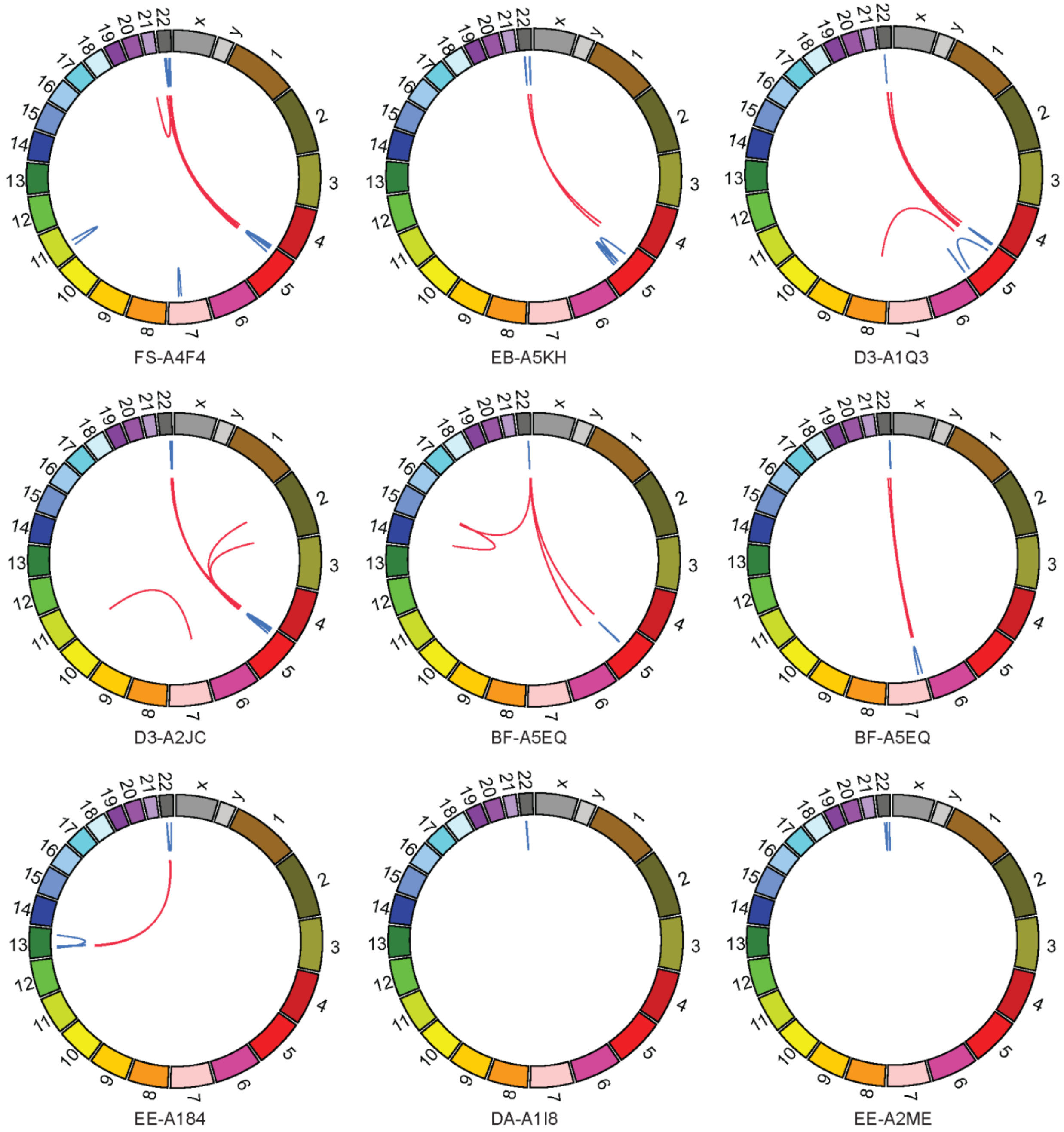

Figure S9. Circos plots for nine melanoma cases with massive rearranged chromosome 22. Five of them involve chromosome 5, two involve other chromosomes but not chromosome 5, and two do not involve any other chromosomes.

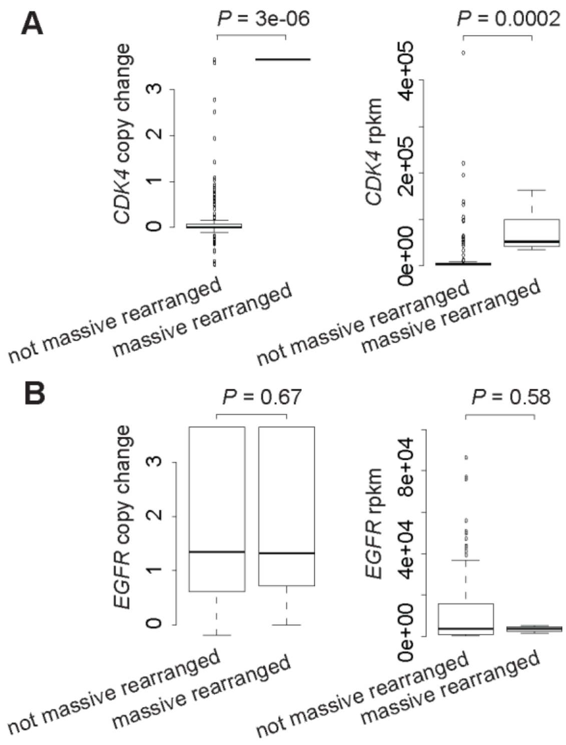

Figure S10. Comparisons of copy numbers and expressions for *CDK4* and *EGFR* in GBM with and without massive rearrangements. A, *CDK4* (chromosome 12). **B**, *EGFR* (chromosome 7). Zero in copy change denotes copy neutral and positive number in copy change denotes copy gain. Wilcoxon one-side rank test was used.

ID	chrA	posA	oriA	geneA	chrB	posB	oriB	geneB	event_type	disc_pair	split_read	homology	validation
SKCM-ER-A19E	14	23532191	-1	ACIN1	14	23683359	1		tandem_dup	9	9	2	0
SKCM-ER-A19E	17	74082405	-1	EXOC7	17	74154545	1	RNF157	tandem_dup	20	21	0	1
SKCM-DA-A1HW	16	18044955	1		16	19548665	-1	CP110	del	6	2	11	1
SKCM-DA-A1HW	16	19485535	1	TMC5	16	19690623	-1	C16orf62	del	5	5	1	1
SKCM-DA-A1HW	11	33054030	1	DEPDC7	11	41760952	-1		del_ins	5	5	-2	1
SKCM-DA-A1HW	15	85328052	-1	ZNF592	15	102151414	-1		invers_r	10	8	1	1
SKCM-DA-A1HW	11	36103407	1	LDLRAD3	4	1311103	1	MAEA	transl_inter	6	5	0	1
LUAD-55-6982	1	142955754	-1		12	57920445	-1	MBD6	transl_inter	5	22	-7	1
LUAD-55-6982	1	169347544	-1	BLZF1	12	68432633	-1		transl_inter	15	11	1	1
LUAD-44-2659	14	24084435	1		14	39855262	-1		del	14	4	2	1
LUAD-44-2659	14	33185034	1	AKAP6	14	39855193	1		invers_f	4	4	3	1
LUAD-44-2659	20	39540716	1		20	39794027	1	PLCG1	invers_f	6	1	-29	1
LUAD-44-2659	22	39078409	-1	TOMM22	22	39125367	1	GTPBP1	tandem_dup	8	7	1	0
LUAD-44-2659	10	61068056	-1	FAM13C	7	44294066	1	CAMK2B	transl_inter	4	4	-17	0
LUAD-05-4396	16	30392278	-1	1-Sep	5	154738915	-1		transl_inter	4	7	1	1
LUAD-49-4512	11	65300191	-1	SCYL1	8	56378712	-1	XKR4	transl_inter	7	5	0	0
LUAD-05-5429	17	65794643	-1		17	73230745	1	NUP85	tandem_dup	12	5	0	0
LUAD-05-5429	12	56493903	-1	ERBB3	16	11917359	-1	BCAR4	transl_inter	8	1	0	1
LUAD-49-6742	19	14951975	-1	OR7A10	19	46543671	-1	IGFL4	invers_r	12	10	0	1
LUAD-91-6840	18	71930713	1	CYB5A	18	71958983	-1	CYB5A	del	9	1	2	1
SKCM-ER-A19L	6	80725550	-1	TTK	6	80746118	-1	TTK	invers_r	10	10	2	1
SKCM-ER-A19L	17	29284811	-1	ADAP2	17	31226566	1		tandem_dup	7	1	0	1
SKCM-DA-A1I8	22	18640545	1	USP18	22	21445272	-1		del	9	9	0	0
SKCM-DA-A1I8	22	19740422	-1		22	21355490	-1	THAP7	invers_r	7	2	-20	1
SKCM-EB-A24D	12	7048176	1	ATN1	12	7077799	-1	PHB2	del_ins	5	4	-2	1
SKCM-EB-A24D	9	128274797	1	MAPKAP1	9	131356713	1	SPTAN1	invers_f	7	10	0	1
SKCM-EE-A2GT	6	32010309	1	TNXB	6	35512873	-1		del	7	24	1	1

Table S2. PCR validation.

ID	chrA	posA	oriA	geneA	chrB	posB	oriB	geneB	event_type	disc_pair	split_read	homology
BLCA-DK-A3IS	9	20173504	1		9	21970869	-1	CDKN2A	del	4	10	2
BLCA-K4-A5RJ	9	21852271	1	MTAP	9	21994341	-1	CDKN2A	del	4	3	3
HNSC-CV-7427	9	21971883	1	CDKN2A	9	32431704	-1	ACO1	del	17	25	0
HNSC-DQ-5625	9	21969406	1	CDKN2A	9	22012184	-1	CDKN2B-AS1	del	8	3	9
LUAD-78-7542	9	21971881	1	CDKN2A	9	22012127	-1	CDKN2B-AS1	del	6	1	0
SKCM-EE-A29H	9	21968345	1	CDKN2A	9	22008169	-1	CDKN2B	del	17	14	0
BRCA-AO-A1KS	17	7468851	1	SENP3	17	7587689	1	TP53	invers_f	12	11	3
GBM-06-0237	16	6261180	1	RBFOX1	17	7576799	-1	TP53	transl_inter	16	11	4
HNSC-CV-7095	17	7578543	1	TP53	17	7689076	-1	DNAH2	del	14	8	1
LIHC-BC-A10Y	17	7577216	1	TP53	9	74887791	1		transl_inter	8	9	0
PRAD-G9-6329	17	7481771	1	EIF4A1	17	7577502	-1	TP53	del	9	9	3
PRAD-HC-A48F	17	5347154	1	DHX33	17	7578439	-1	TP53	del	5	4	0
PRAD-HC-A48F	17	7578250	1	TP53	18	66524823	-1	CCDC102B	transl_inter	8	7	1
LGG-HT-7873	10	89717526	-1	PTEN	6	93021957	-1		transl_inter	11	3	0
LUAD-17-Z017	10	89672707	-1	PTEN	10	129870486	1	PTPRE	tandem_dup	17	9	0
LUSC-66-2770	10	89690926	1	PTEN	10	89717215	-1	PTEN	del	19	25	1
PRAD-EJ-5521	10	69075557	1	CTNNA3	10	89690880	1	PTEN	invers_f	5	7	0
SKCM-BF-A3DN	10	86864700	-1		10	89624234	-1	PTEN	invers_r	9	20	2
SKCM-ER-A42K	1	242236694	1		10	89653956	1	PTEN	transl_inter	6	4	3
STAD-B7-5818	10	89653739	-1	PTEN	10	89744296	-1		invers_r	20	19	2

Table S4. Somatic SVs distupting tumor suppressors.

ID	chrA	posA	oriA	geneA	chrB	posB	oriB	geneB	event_type	disc_pair	split_read	homology
Cytoskeleton genes												
THCA-FK-A3SE	10	61655977	-1	CCDC6	10	43611997	-1	RET	invers_r	13	17	3
THCA-EL-A3ZS	10	61659539	-1	CCDC6	10	43611930	-1	RET	invers_r	4	4	0
THCA-BJ-A0ZJ	10	61626050	-1	CCDC6	10	43611953	-1	RET	invers_r	13	5	1
THCA-ET-A3DQ	9	115932783	-1	FKBP15	10	43610457	-1	RET	transl_inter	5	2	-3
LUAD-67-6215	2	42491894	1	EML4	2	29447037	1	ALK	invers_f	6	5	2
BRCA-AR-A0TX	20	55012426	1	CASS4	12	75712009	1	CAPS2	transl_inter	12	11	0
HNSC-CV-7243	5	75866511	1	IQGAP2	10	117242409	-1	ATRNL1	transl_inter	8	7	2
BRCA-C8-A12X	2	204319150	-1	RAPH1	2	234864004	-1	TRPM8	invers_r	37	20	2
LIHC-DD-A3A7	22	38137110	1	TRIOBP	22	46643348	1	C22orf40	invers_f	7	2	8
PRAD-HC-8264	12	32299559	1	BICD1	12	66221789	-1	HMGA2	del	8	7	1
KIRC-CJ-4882	10	102045759	-1	BLOC1S2	10	102232322	-1	WNT8B	invers_r	4	2	3
LIHC-DD-A116	19	1026749	1	CNN2	19	992884	-1	WDR18	tandem_dup	17	36	1
LUAD-97-A4M1	5	629160	1	CEP72	5	6748341	-1	PAPD7	del_ins	9	8	-1
PRAD-HC-8262	1	156302064	-1	CCT3	1	155823591	1	GON4L	del	4	2	0
BRCA-PE-A5DC	11	70279907	1	CTTN	11	71943775	-1	INPPL1	del	50	42	0
BRCA-AR-A1AH	3	196989153	-1	DLG1	3	197792772	1	LOC348840	tandem_dup	19	15	4
LUAD-49-4490	18	5479153	-1	EPB41L3	22	41282396	-1	XPNPEP3	transl_inter	7	4	2
BRCA-B6-A0I1	19	12963954	1	MAST1	11	65033937	-1	POLA2	transl_inter	5	5	5
BLCA-DK-A1A7	17	30963585	-1	MYO1D	17	32116519	1	ACCN1	tandem_dup	5	1	0
KIRP-P4-A5EB	5	58682616	-1	PDE4D	5	65029155	-1	NLN	invers_r	6	6	0
LUAD-69-7980	X	50446818	-1	SHROOM4	X	45013382	1	CXorf36	del	5	9	2
SKCM-EB-A3XF	22	31485928	1	SMTN	3	49700916	-1	BSN	transl_inter	4	2	6
BRCA-A2-A3Y0	11	66453566	-1	SPTBN2	11	69517170	1	FGF19	tandem_dup	4	3	1
BRCA-E9-A1NF	15	99670591	1	SYNM	15	99535129	1	PGPEP1L	invers_f	8	6	0
KIRP-GL-7966	16	2120616	1	TSC2	16	2041973	-1	SYNGR3	tandem_dup	4	5	2
BRCA-BH-A18K	19	22825352	1	ZNF492	19	56549328	-1	NLRP5	del_ins	22	38	-5
Biosynthesis genes												
BRCA-C8-A12V	1	27060158	1	ARID1A	1	155172955	1	THBS3	invers_f	6	4	-4
LGG-DH-5140	22	38373695	-1	SOX10	11	73130191	1	FAM168A	transl_inter	7	4	0

LUSC-43-A475	6	42023974	1	TAF8	6	110764199	1	SLC22A16	invers_f	19	12	-1
KIRC-CJ-4882	10	102045759	-1	BLOC1S2	10	102232322	-1	WNT8B	invers_r	4	2	3
BRCA-AQ-A54N	12	121693336	-1	CAMKK2	12	121882684	1	KDM2B	tandem_dup	19	34	4
BRCA-AR-A24W	X	40541858	-1	MED14	X	133102817	1	GPC3	tandem_dup	4	6	1
LUSC-NC-A5HL	5	176700705	1	NSD1	5	176452827	-1	ZNF346	tandem_dup	8	8	-1
BRCA-C8-A278	1	164786924	1	PBX1	1	156354277	-1	RHBG	tandem_dup	8	1	-11
BRCA-A1-A0SK	6	43141951	1	SRF	1	169484431	1	F5	transl_inter	23	13	1
LGG-IK-7675	9	32544233	-1	TOPORS	9	18824908	-1	ADAMTSL1	invers_r	4	5	2
LUAD-55-8615	2	85535039	1	TCF7L1	14	36340716	-1	BRMS1L	transl_inter	6	7	4
KIRP-J7-8537	X	48897474	-1	TFE3	17	7132983	1	DVL2	transl_inter	23	14	1
LUSC-L3-A524	6	43752524	1	VEGFA	6	43571575	-1	POLH	tandem_dup	4	2	-19

Table S6. 5' fusion partners of activating fusions enriched in house-keeping genes.

ID	chrA	posA	oriA	geneA	chrB	posB	oriB	geneB	event_type	disc_pair	split_read	homology
BRCA-AR-A1AY	11	66919403	1	KDM2A	11	67200591	-1	RPS6KB2	del	4	2	1
BRCA-A7-A13D	17	7755424	1	KDM6B	17	7259504	-1	TMEM95	tandem_dup	18	6	-2
BRCA-C8-A12V	1	27060158	1	ARID1A	1	155172955	1	THBS3	invers_f	6	4	-4
BRCA-BH-A18H	7	151962195	-1	MLL3	X	44098470	1	EFHC2	transl_inter	17	22	2
GBM-06-5856	12	121916361	-1	KDM2B	7	54825303	1	SEC61G	transl_inter	11	11	-15
PRAD-EJ-8469	X	44918730	1	KDM6A	X	11418883	1	ARHGAP6	invers_f	9	8	-3
LUSC-NC-A5HL	5	176700705	1	NSD1	5	176452827	-1	ZNF346	tandem_dup	8	8	-1
LUSC-34-5928	17	30293540	1	SUZ12	17	30349632	-1	LRRC37B	del	4	19	1

Table S7. 5’ fusion partners of activating fusions enriched in chromatin regulators.

ID	chrA	posA	oriA	geneA	chrB	posB	oriB	geneB	event_type	disc_pair	split_read	homology
BRCA-AR-A0U3	11	68777066	-1	MRGPRF	11	66949518	-1	KDM2A	invers_r	5	3	1
BRCA-AQ-A54N	12	121693336	-1	CAMKK2	12	121882684	1	KDM2B	tandem_dup	19	34	4
BRCA-OL-A5D7	19	1219348	1	STK11	19	5024719	-1	KDM4B	del_ins	6	7	-2
BRCA-AR-A0TT	1	155058733	1	EFNA3	1	161129991	-1	USP21	del_ins	15	1	-1
BRCA-AR-A250	17	59370199	1	BCAS3	17	47889001	-1	MYST2	tandem_dup	12	15	3
BRCA-BH-A1EN	17	37343318	-1	CACNB1	11	76201326	-1	C11orf30	transl_inter	8	11	0
BLCA-DK-A6AV	12	56641865	-1	ANKRD52	12	56562982	1	SMARCC2	del	15	9	5

Table S8. 3’ fusion partners of activating fusions enriched in chromatin regulators.