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

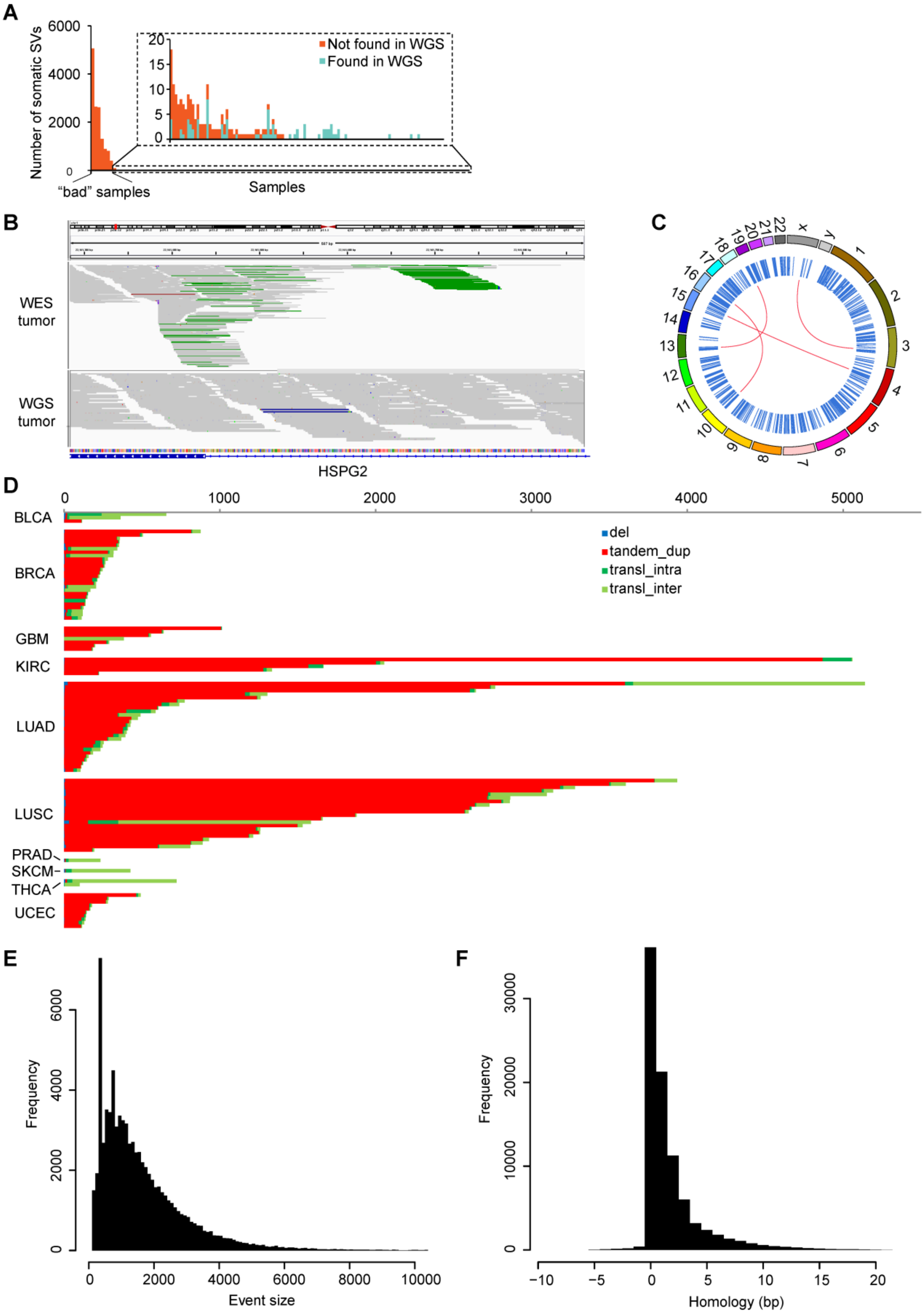

A
Number of somatic SVs
6000
4000
2000
0
"bad" samples
Samples
20
15
10
5
0
🟧Not found in WGS
🟦Found in WGS
B
WES tumor
WGS tumor
HSPG2
C
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
D
0
1000
2000
3000
4000
5000
BLCA
BRCA
GBM
KIRC
LUAD
LUSC
PRAD
SKCM
THCA
UCEC
■del
■tandem_dup
■transl_intra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

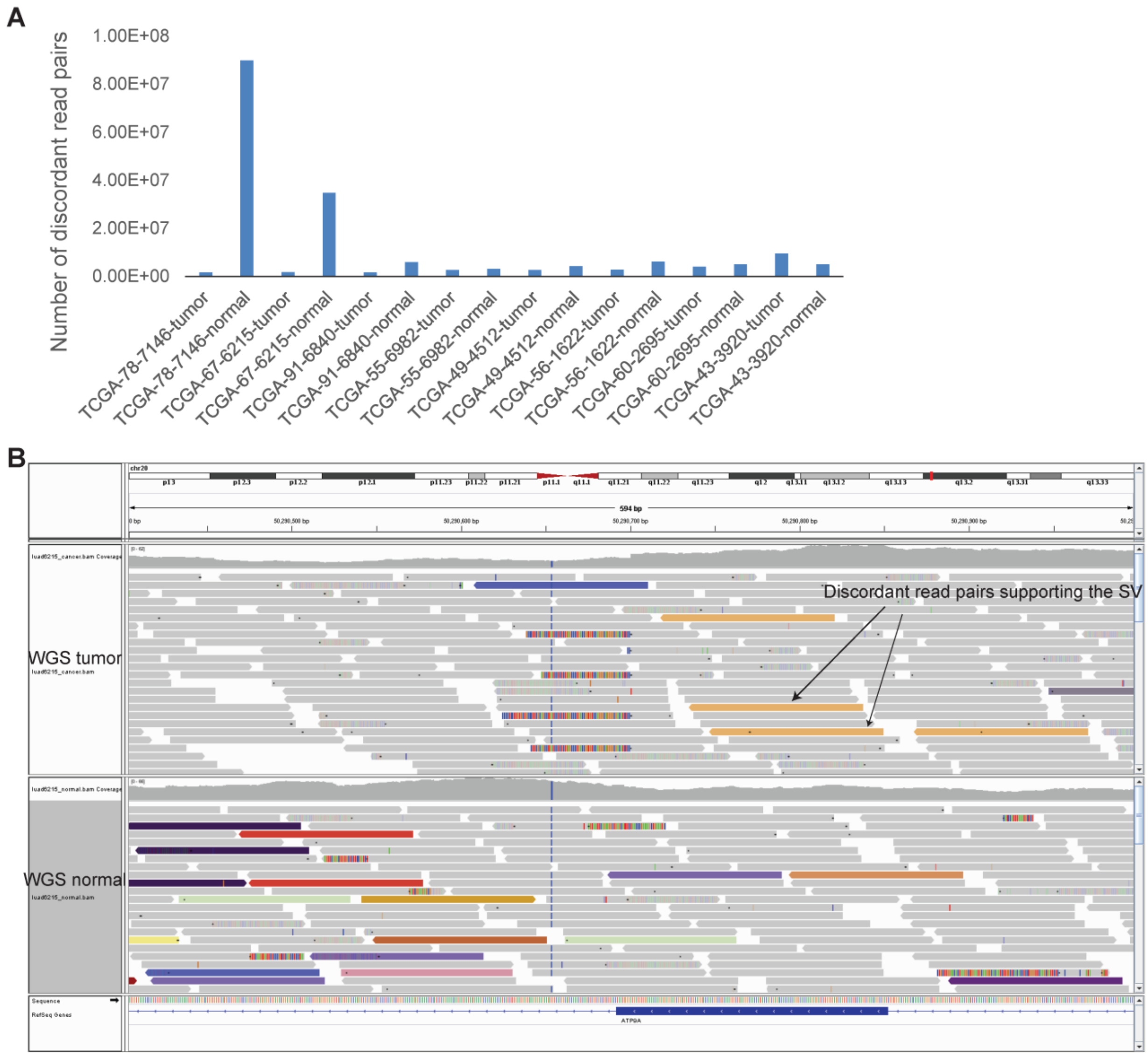


Figure S2. Noises in WGS samples. A, Number of discordant read pairs in WGS samples. Eight tumor-normal pairs are shown. Some normal samples have excessive discordant read pairs (TCGA-78-7146 and TCGA-67-6215). These discordant pairs are generated from library construction and sequencing, rather than 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.

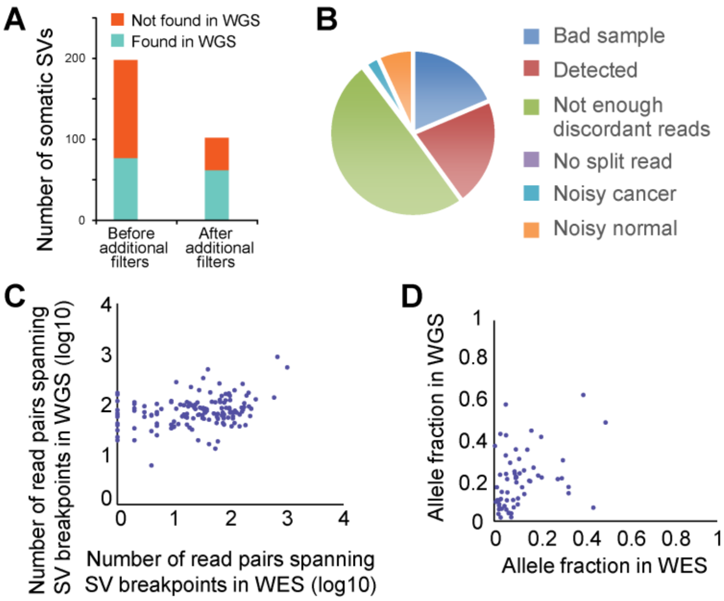


Figure S3. Comparison of somatic rearrangements detected from WES and WGS. A, Comparison of somatic rearrangements detected from WES before and after additional filters (with poor-quality samples excluded) shows that the fraction of WES-specific calls is substantially reduced. **B**, The "catchable" 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.

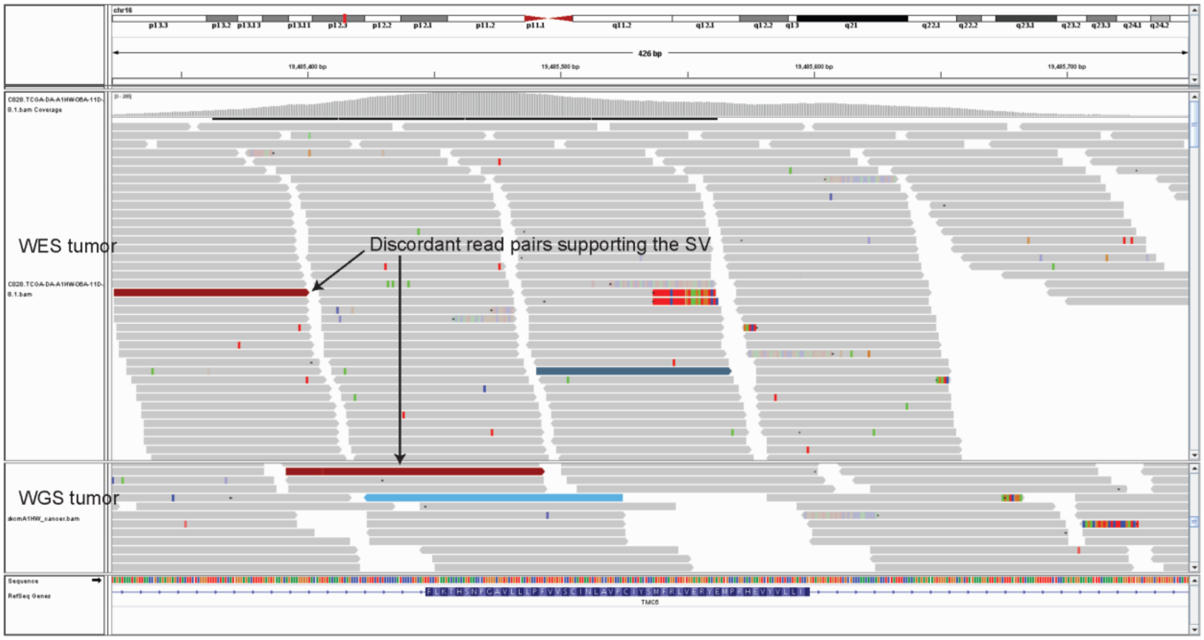


Figure S4. Example of an SV detected in WES but not in WGS due to higher coverage of WES. A somatic deletion in melanoma TCGA-DA-A1HW (chr16:19485535-19690623) is detected from WES but not in WGS. The coverage in WES is 300x and there are 4 discordant read pairs (only 1 is displayed). In contrast, the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

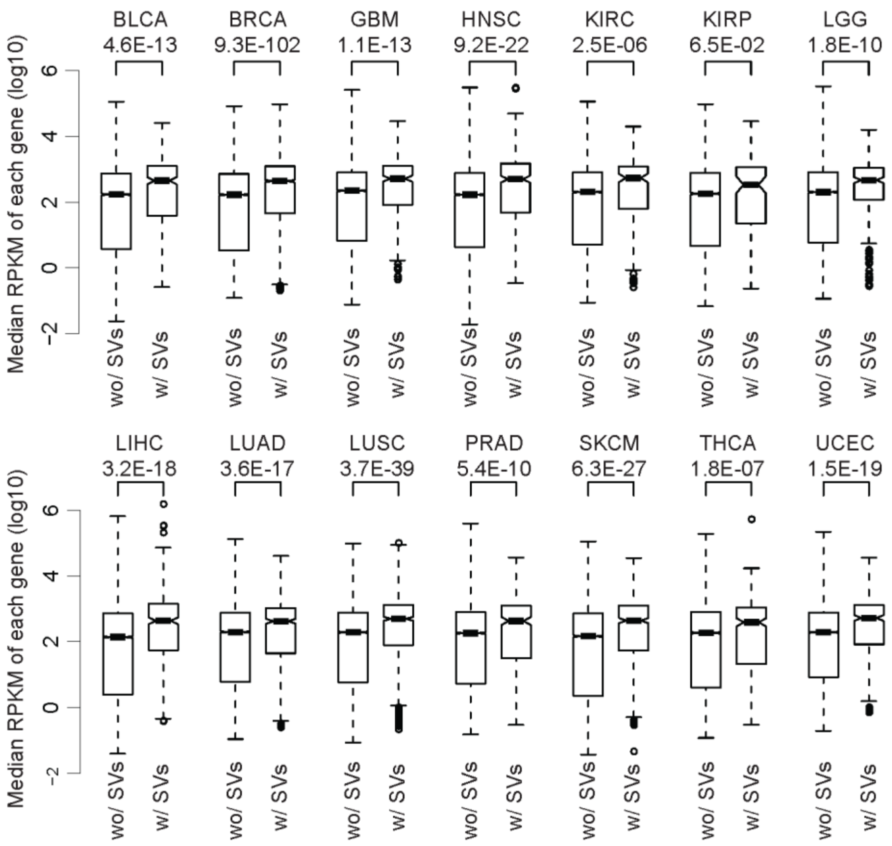


Figure S5. The expression of genes with and without somatic SVs for 14 tumor types. The *P* values by Wilcoxon one-side rank test are shown below the tumor type names. The median expression level of each gene across all individuals for one tumor type was plotted. This shows somatic SVs occur in relatively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

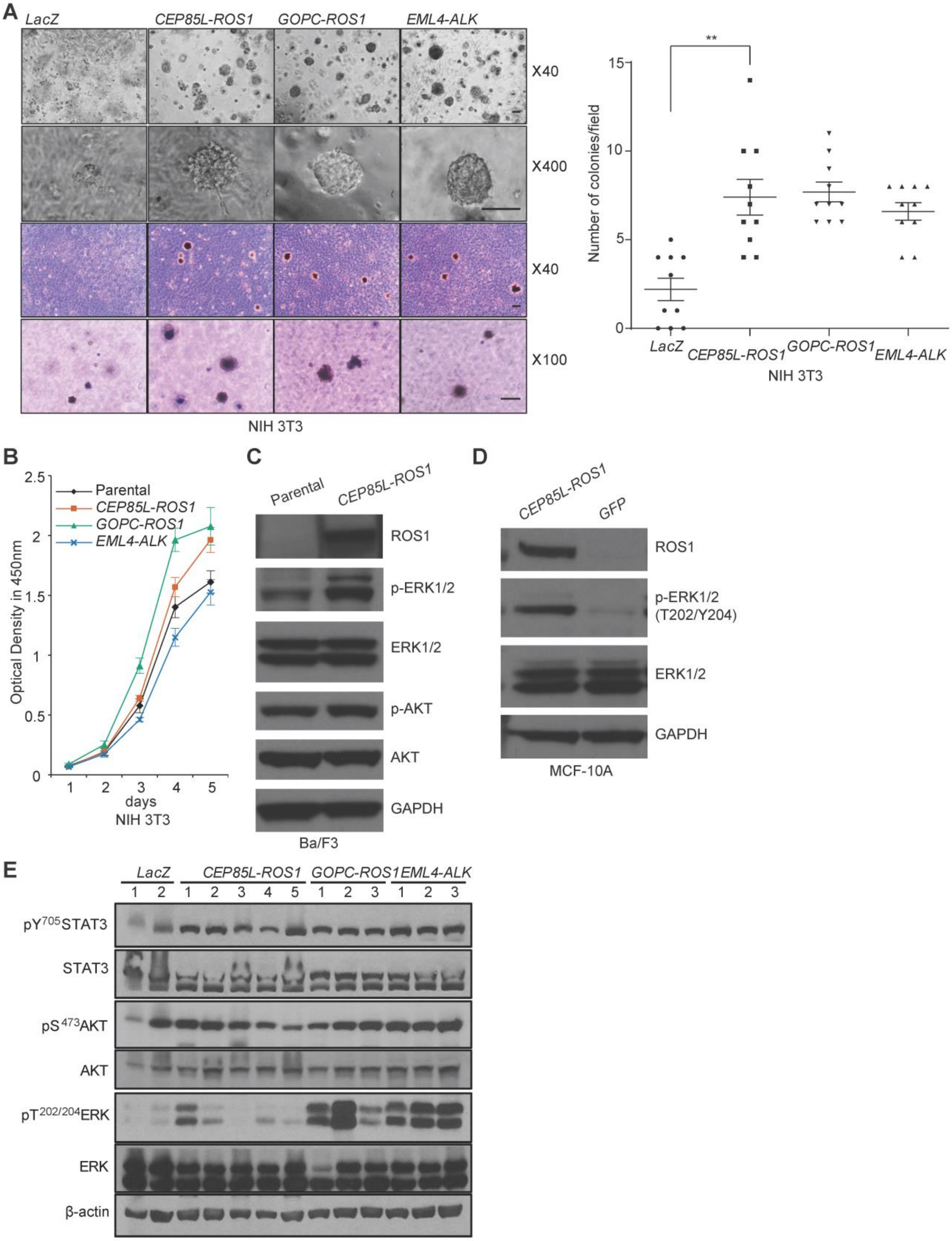


Figure S6. Functional validation for *CEP85L-ROS1* fusion. A, *In vitro* transforming assay of NIH 3T3 cells. Cells expressing the indicated fusion genes were cultured in Matrigel (upper left panels) for 14 days or soft agar (lower left panels) for 21 days. For visualization, some colonies

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.

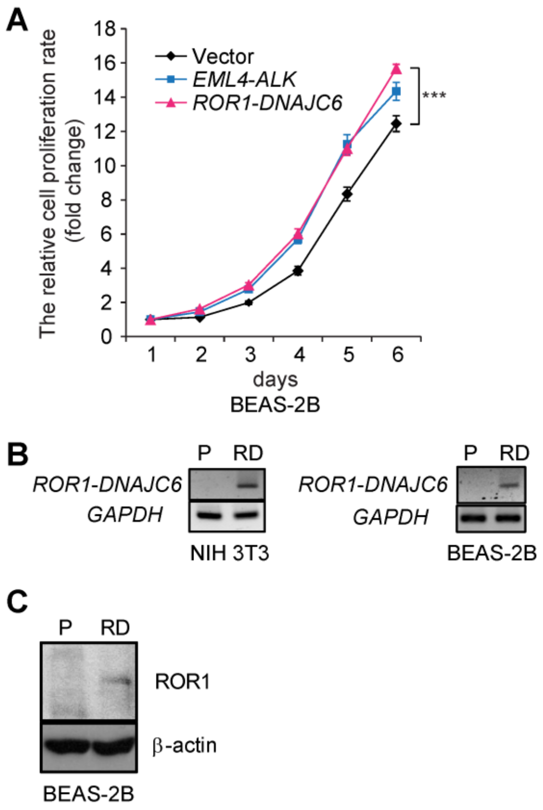


Figure S7. Functional validation for *ROR1-DNAJC6* fusion. A, Growth rate of BEAS-2B cells expressing *ROR1-DNAJC6* fusion measured by optical density at 450 nm (OD450). **B**, Identification of *ROR1-DNAJC6* mRNA expression using RT-PCR in NIH 3T3 and BEAS-2B cells infected with *ROR1-DNAJC6* lentivirus.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

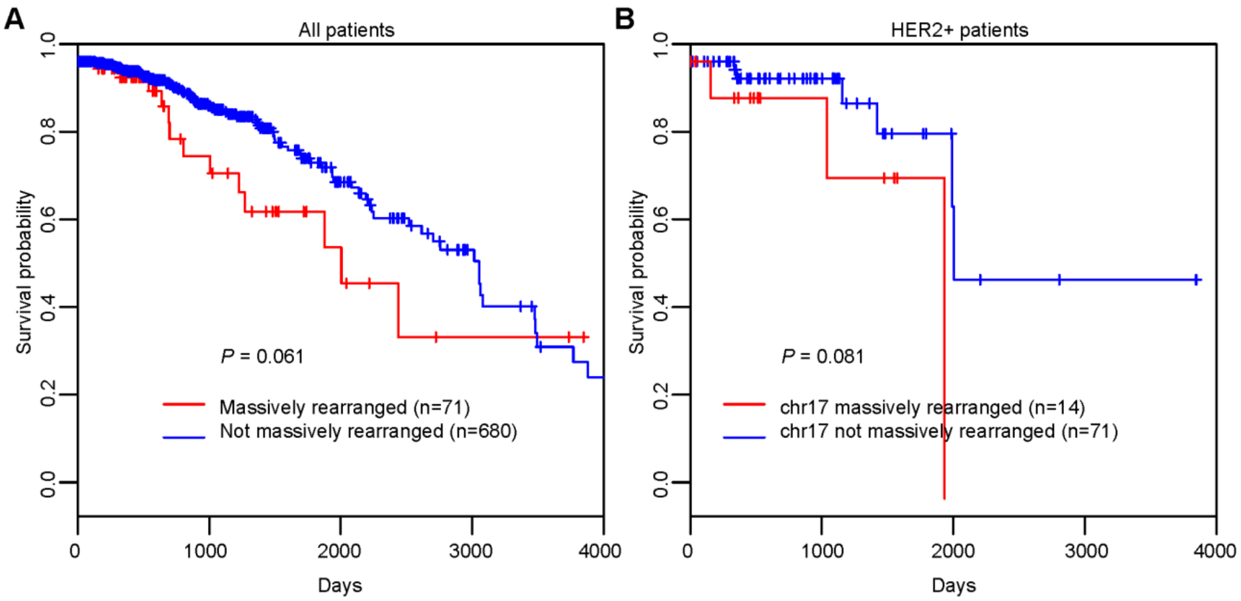
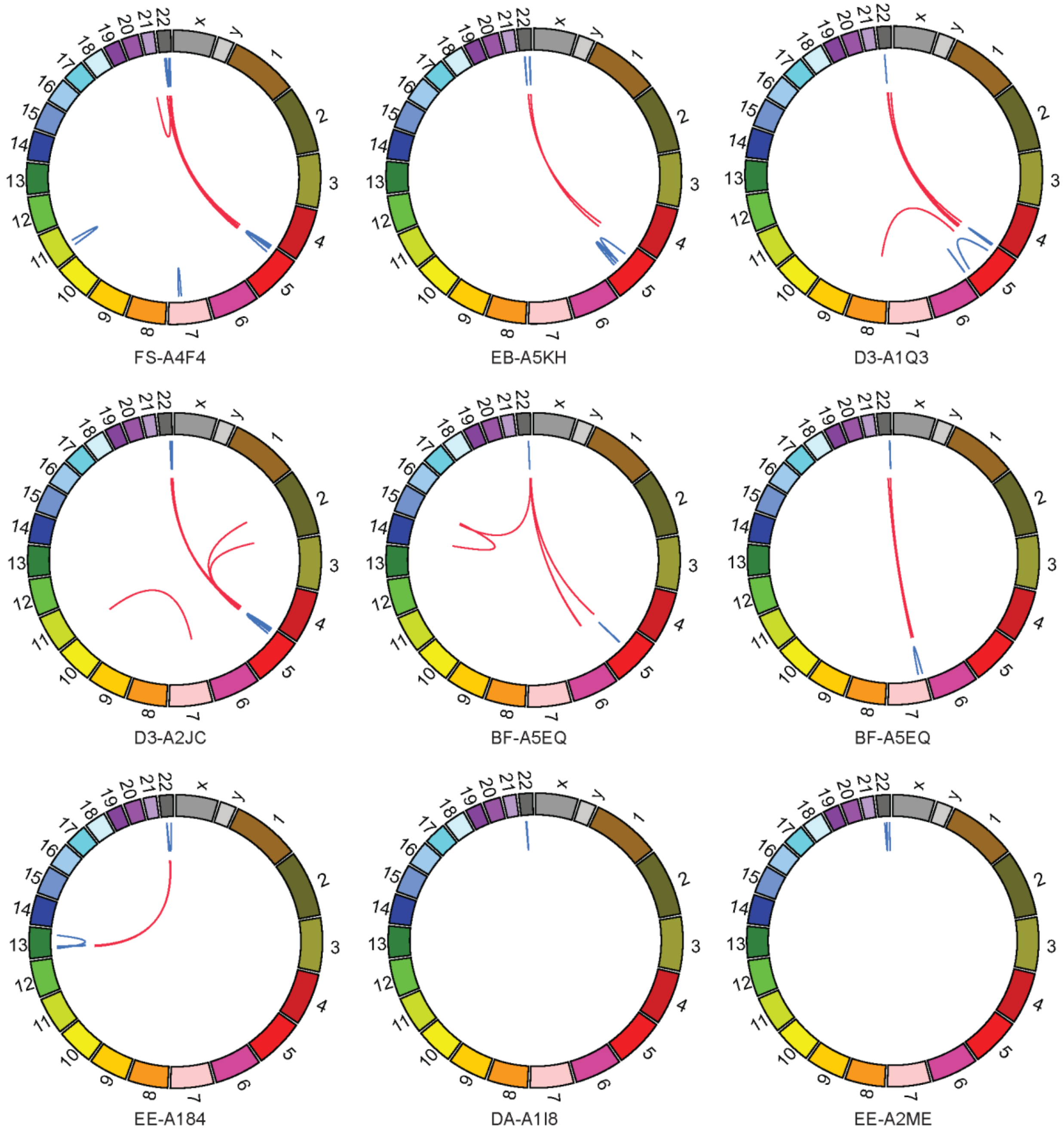


Figure S8. Survival plots for breast cancer patients with and without massive rearrangements. A, All individuals were divided between those with and without rearrangements. **B**, The HER2+ subgroup was divided between those with and without massively-rearranged chromosome 17. The p-values, computed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



FS-A4F4

EB-A5KH

D3-A1Q3

D3-A2JC

BF-A5EQ

BF-A5EQ

EE-A184

DA-A1I8

EE-A2ME

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.

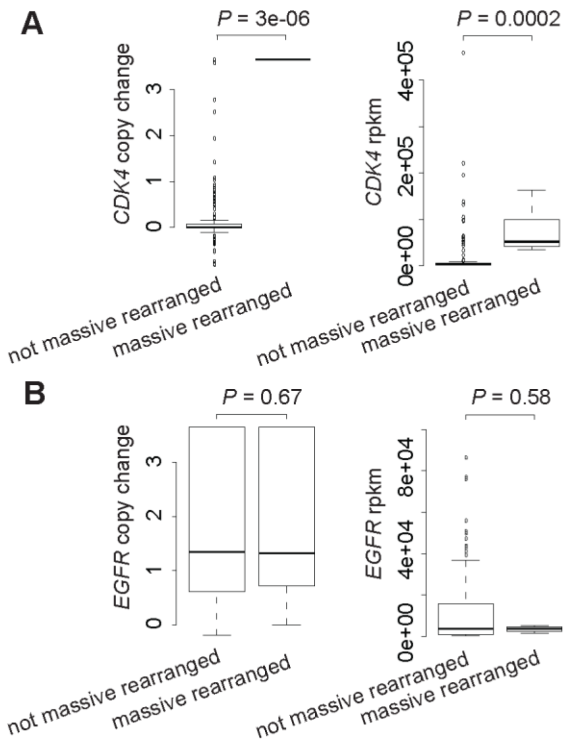


Figure S10. Comparisons of copy numbers and expressions for *CDK4* and *EGFR* in GBM with and without massive rearrangements. A, *CDK4* (chromosome 12). **B**, *EGFR* (chromosome 7). Zero in copy change denotes copy neutral and positive number in copy change denotes copy gain. Wilcoxon one-side rank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