Supplemental Data

Transcriptional Consequences of Genomic Structural Aberrations in Breast Cancer

Koichiro INAKI[1]*, Axel M. HILLMER[2]*, Leena UKIL[1]*, Fei YAO[2, 9], Xing Yi WOO[3], LEAH Vardy[5], Kelson Folkvard Braaten ZAWACK[3], Charlie Wah Heng LEE[3], Pramila Nuwantha ARIYARATNE[3], Yang Sun CHAN[1], Kartiki Vasant DESAI[1], Jonas Bergh[6], Per HALL[7], Thomas Choudary PUTTI[8], Wai Loon ONG[4], Atif SHAHAB[4], Valere CACHEUX-RATABOUL[1], Radha Krishna Murthy KARUTURI[3], Wing-Kin SUNG[3], Xiaoan RUAN[2], Guillaume BOURQUE[3], Yijun RUAN[2], Edison T. LIU[1]

[1]Cancer Biology and Pharmacology, [2]Genome Technology and Biology, [3]Computational and Mathematical Biology, [4]Research Computing, Genome Institute of Singapore, 60 Biopolis Street, Genome, Singapore 138672, Singapore. [5] Translational Regulation in Stem Cells, Institute of Medical Biology, 8A Biomedical Grove, Immunos, Singapore 138648, Singapore. [6]Department of Oncology – Pathology, [7]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, SE-171 77 Stockholm, Sweden. [8]Department of Pathology, National University of Singapore, Singapore 119077, Singapore. [9]Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore.

•contributed equally to this work

Correspondence should be addressed to Edison T. Liu
Genome Institute of Singapore
60 Biopolis, Singapore 138672,
Tel. +65 6478-8007
Fax. +65 6808 8291
E-mail. liue@gis.a-star.edu.sg

Glossary of Terms:

The added complexity of categorization of transcripts arising from a variety of genomic rearrangements presented some challenges to developing a unified classification scheme. Therefore we used a common terminology that describes the structural changes at either the genomic or at the transcript level. However, theoretically a genomic structural mutation of one class may generate a transcript fusion of another class.

Genomic rearrangement descriptors: Different genomic rearrangements such as tandem duplications (TD) or deletions (Del) can be hypothetically projected to generate putative abnormal transcripts. The hypothetical transcriptional output of these genomic rearrangements is the basis of the classification scheme used herein:

Fusion Genes (notation FG): refer to genomic rearrangements that predict for transcripts in which two distinct RefSeq genes are fused together in the same direction/orientation.

3'-Terminus Truncations (notation 3'T): refer to genomic rearrangements that predict for transcripts in which the 3'-terminus portion of a given 5' partner RefSeq gene is truncated and is fused to:
-any segment of the genome not within a RefSeq gene but with some *evidence for a transcript* in current databases. **(Notation: 3'T-E)**
- - any segment of the genome not within a RefSeq gene and where there is *no evidence for any transcript* in current databases. **(Notation: 3'T-N)**

5'-Terminus Truncations (notation 5'T): rearrangements in which the 5'-terminus portion of a 3' partner RefSeq gene is truncated and fused to:
-any segment of the genome not within a RefSeq gene and where there is some *evidence for a transcript* in current databases. **(Notation: 5'T-E)**
- any segment of the genome not within a RefSeq gene and where there is *no evidence for any transcript* in current databases. **(Notation: 5'T-N)**

Intragenic Rearrangements (notation IR): rearrangements in which the genomic abnormalities (deletion, tandem duplication, inversion, or insertion) are located inside the gene body that result in an internal rearrangement or deletion.

Therefore genomic rearrangements can be characterized by a transcript descriptor, e.g. tandem duplications and genomic deletions can be classified as FG if both mutations can hypothetically generate a fusion transcript between two RefSeq gene partners.

Glossary of Terms (2):

Transcript rearrangement descriptors: Each class of abnormal transcripts can also be characterized by the descriptors described above. However, the need to separate the descriptors for the transcripts from those for their cognate genomic rearrangement is because we recognize the possibility that a genomic descriptor will generate a transcript descriptor of a related but different class. For example: an FG genomic rearrangement may actually generate a final fusion transcript that only engages the intron of one gene partner and no exon. Moreover, in this same situation with two RefSeq genes, the 3' fusion partner may contribute sequences from the opposite strand of its gene sequence and therefore would be a novel 3'T transcript and have very different functions that would be predicted by the exon domains in the intact RefSeq gene. Therefore for transcripts, we define the various mutation and fusions only relative to the structure of the validated transcript.

Fusion Gene transcripts (notation FG^R): refer to transcripts in which exons from two distinct RefSeq genes are fused together in the same direction.

3'-Terminus Truncations transcripts (notation $3'T^R$): refer to fusion transcripts in which the 3'-terminus portion of a given 5' partner gene is truncated and is fused to:
- a non-RefSeq but annotated segment that *has evidence for being part of a transcript*. **(Notation: $3'T\text{-}E^R$)**
- any genomic segment that 1) is in the anti-sense strand of any known gene/transcript, or 2) in an unannotated region that has *no evidence for a transcript* in current databases. **(Notation: $3'T\text{-}N^R$)**

5'-Terminus Truncations (notation $5'T^R$): refer to fusion transcripts in which the 5'-terminus portion of a 3' partner RefSeq gene is truncated and fused to:
- a non-RefSeq but annotated segment that *has evidence for being part of a transcript*. **(Notation: $5'T\text{-}E^R$)**
- any genomic segment that 1) is in the anti-sense strand of any known gene/transcript or 2) in an unannotated region that has *no evidence for a transcript* in current databases. **(Notation: $5'T\text{-}N^R$)**

Intragenic Rearrangements (notation IR^R): Aberrant transcripts in which the genomic abnormalities (deletion, tandem duplication, inversion, or insertion) result in an internal rearrangement, insertion, or deletion. The IR transcripts do not include splice variants.

Fusion transcripts: Transcripts that are abnormal or chimeric transcripts but has some sequences derived from a RefSeq gene. In the general case, fusion transcripts refer to FG^R, $3'T^R$, $5'T^R$, and IR^R transcripts. In some cases, the genomic rearrangement predicted for an incorrect fusion transcript upon validation. However, since 5'T rearrangements do not generate a transcript and the IR rearrangements are associated with attenuate expression or silenced genes, we are referring primarily to FG + 3'T. In these cases, we have denoted them as [genomic notation → transcript notation^R]: e.g. $3'T \rightarrow FG^R$, is a genomic rearrangement (3'T) that generated an FG^R fusion transcript because of a splicing event that captured a downstream 3' exon that is part of a RefSeq gene (see Supplemental Fig. 2E).

Glossary of Terms (3):

In some cases, the genomic rearrangement predicted for an incorrect fusion transcript upon validation. In these cases, we have developed a notion that can describe the progress of the discovery: genomic notation → transcript notation, e.g. 3'T→FG describes a genomic rearrangement that generated an FG fusion transcript because of a splicing event that captured a downstream 3' exon that is part of a RefSeq gene.

Novel transcript: Transcripts that are abnormal because they have never been annotated as a transcript or a gene and includes no sequences recognized previously as part of a known transcript.

	Cell lines				Tumors			
	Del	TD	Inv	Total	Del	TD	Inv	Total
Exon rearrangement	52	28	0	80	46	8	0	54
Intron	111	19	1	131	166	10	0	176
Total	163	47	1	211	212	18	0	230

Supplemental Table 1.

Number of intragenic rearrangements causing exon rearrangements in 3 breast cancer cell lines and 5 primary tumors. Del=Deletion; TD=Tandem duplication; Inv=Inversion.

	5' Gene	5' Chr	3' Gene	3' Chr	gPET cluster size	Structure variation		Distance between fusion points (kb)	GIS-PET count		RNA-PET count		Reference
						Connection	Super-Cluster		SHC012	SHC025	IHM098	IHM101	
MCF7	*NAV1*	chr1	*GPR37L1*	chr1	37*	Del	1	255				3	
	MTAP	chr9	*CDKN2BAS*	chr9	26	Del	2	247			5	6	
	MYO9B	chr19	*FCHO1*	chr19	13	Del	1	655				20	
	ARFGEF2	chr20	*SULF2*	chr20	496	U-Inv (complex)	205			1	10	170	Hampton et al. 2009
	VAV3	chr1	*AK123199*	chr1	63*	U-Inv (complex)	205					2	
	RSBN1	chr1	*AK123199*	chr1	40	U-Inv (complex)	205					11	
	NCOA3	chr20	*SULF2*	chr20	27	U-Inv (complex)	205						
	BCAS4	chr20	*ZMYND8*	chr20	6	U-Inv (complex)	6				16	41	
	RPS6KB1	chr17	*VMP1*	chr17	9	TD (complex)	205		18	352	67	101	
	DEPDC1B	chr5	*ELOVL7*	chr5	43	TD	1	127		4	2	3	Hampton et al. 2009
	TOP1	chr20	*CR593014*	chr20	22	TD	1	62				2	
	PTPRN2	chr7	*FAM62B (ESYT2)*	chr7	21	TD	1	392				2	
	SMARCA4	chr19	*CARM1*	chr19	19	TD	1	96	1	3	2	3	
	ATXN7L3	chr17	*FAM171A2*	chr17	18	TD	1	160			2	12	
	PLCG1	chr20	*TOP1*	chr20	18	TD	1	50					
	PNPLA7	chr9	*WDR85*	chr9	17	TD	1	49				5	
	TSPAN9	chr12	*TEAD4*	chr12	17	TD	1	154				11	
	ESR1	chr6	*C6orf97*	chr6	15	TD	2	158					
	GCN1L1	chr12	*MSI1*	chr12	15	TD	1	161					
	CHEK2	chr22	*XBP1*	chr22	11*	TD	3	96				3	
	MYO6	chr6	*SENP6*	chr6	9	TD	1	85				2	
	POP1	chr8	*MATN2*	chr8	9	TD	1	90		4		2	
	ANKS1A	chr6	*UHRF1BP1*	chr6	8	TD	1	81					
	PAPOLA	chr14	*AK7*	chr14	7	TD	1	79				7	
	CXorf15 (TXLNG)	chrX	*SYAP1*	chrX	6	TD	1	61	16	30	7	26	
	GATAD2B	chr1	*NUP210L*	chr1	6	TD	1	169	12	24	3	41	
	BCAS4	chr20	*BCAS3*	chr17	752	Inter (complex)	6		339	133	402	1270	Barlund et al. 2002
	BCAS3	chr17	*ATXN7*	chr3	309*	Inter (complex)	205					9	Bashir et al. 2008
	SULF2	chr20	*PRICKLE2*	chr3	135	Inter (complex)	205					2	
	RAD51C	chr17	*ATXN7*	chr3	129	Inter (complex)	205			2			Hampton et al. 2009
	ATP1A1	chr1	*AK222712 [ZFP64]*	chr20	102	Inter (complex)	205			109		4	
	TAF4	chr20	*BRIP1*	chr17	29	Inter (complex)	205						
	UBE2V1	chr20	*TBX2*	chr17	23	Inter (complex)	205					2	
	PTPRG	chr3	*CCDC129*	chr7	22	Inter (complex)	5						
	TBL1XR1	chr3	*RGS17*	chr6	14*	Inter (complex)	205					8	Hahn et al. 2004
	BCAS3	chr17	*AMPD1*	chr1	13	Inter (complex)	205					2	
	RPS6KB1	chr17	*DIAPH3*	chr13	3	Inter (complex)	205			14	3	11	
	TEX14	chr17	*PTPRG*	chr3	411	Inter	1						
	KCND3	chr1	*PPM1E*	chr17	334	Inter	1						Bashir et al. 2008
	B3GNTL1	chr17	*SLC9A8*	chr20	48	Inter	1						
	ZMYND8	chr20	*USP32*	chr17	25	Inter	1		7				
	ABCA5	chr17	*PPP4R1L*	chr20	23	Inter	3						
	SGPP2	chr2	*ULK4*	chr3	7	Inter	1						

(continued in the next page)

	5' Gene	5' Chr	3' Gene	3' Chr	gPET cluster size	Structure variation		Distance between fusion points (kb)	GIS-PET count		RNA-PET count		Reference
						Connection	Super-Cluster		SHC012	SHC025	IHM098	IHM101	
SKBR3	*PREX1*	chr20	*CPNE1*	chr20	23	Del (complex)	4						
	TRIO	chr5	*FBXL7*	chr5	46	Del	1	1162					
	ATAD5	chr17	*TLK2*	chr17	11	Del	1	31769					
	PBRM1	chr3	*WDR82*	chr3	7	Del	1	393					
	PLDN	chr15	*AKAP13*	chr15	5	Del	1	40274					
	DEPDC1B	chr5	*PDE4D*	chr5	3	Del	1	246					
	WDR67	chr8	*ZNF704*	chr8	17	U-Inv (complex)	7						
	COL14A1	chr8	*MTSS1*	chr8	204	U-Inv	2	4320					
	TAF2	chr8	*COLEC10*	chr8	49	U-Inv	1	696					
	RANBP10	chr16	*PSKH1*	chr16	14	U-Inv	1	160					
	ANKHD1	chr5	*PCDH1*	chr5	10	U-Inv	1	1409					
	VSTM2L	chr20	*CTNNBL1*	chr20	11	TD (complex)	7						
	WDR67	chr8	*AK298294 [SLC30A8]*	chr8	9	TD (complex)	4						
	DHX35	chr20	*ITCH*	chr20	33	TD	1	4613					
	KIAA1303 (RPTOR)	chr17	*AB046838 [RNF213]*	chr17	19	TD	1	490					
	TATDN1	chr8	*GSDMB*	chr17	199	Inter	1						
	RARA	chr17	*PKIA*	chr8	128	Inter	1						
	CBX3	chr7	*C15orf57*	chr15	3	Inter	1						
T47D	*VPS26A*	chr10	*FAM149B1*	chr10	6	Del	1	4054					
	CNNM2	chr10	*DNAJC9*	chr10	5	U-Inv (complex)	8						
	KMO	chr1	*PDE4DIP*	chr1	5	U-Inv	1	96177					
	NBPF1	chr1	*KIAA0445 (CROCC)*	chr1	4	U-Inv	2	257					
	IQGAP2	chr5	*SV2C*	chr5	16	TD	1	295					
	EVI1 (MECOM)	chr3	*TTC18*	chr10	17	Inter (complex)	8						
	KIAA0232	chr4	*CR621911 / BC034612*	chr5	5	Inter	2						
	RERG	chr12	*CBFB*	chr16	4	Inter	1						

Supplemental Table 2A.

List of identified fusion transcripts (FG^R + 3'T-E^R) by RT-PCR in breast cancer cell lines. DNA-PET cluster size shows number of discordant PETs which form the cluster, reflecting the copy number of the rearrangement point. Super-cluster is made by aggregating nearby DNA-PET clusters (Hillmer et al., submitted). High super-cluster values indicate a region with complex structure variations. We defined clusters with a super-cluster size of more than 3 as complex structures. Some transcripts are supported by GIS-PET (SHC012 and SHC025 libraries; Ng et al. 2006; Ruan et al. 2007) and/or RNA-PET (IHM098 and IHM101 libraries) data. Asterisks* denote genomic rearrangements that produce multiple fusion transcripts because of the overlapping gene annotation in the fusion point. For example, a 3' breakpoint that is within an exon of an embedded in that resides in an intron of a larger gene. Del=Deletion; U-Inv=Unpaired-Inversion; Complex-Inter=Inter-chromosomal connections within hot spots of genomic rearrangement (super-cluster size ≥3); TD=Tandem Duplication. Some fusion transcripts were overlapped with reported fusion genes in MCF-7, while we missed 5 fusion genes on the genomic level (MTIF2-PLEKHH2 and KIAA0182-AK023385 in Volik et al. 2006, HYDIN-NBPF1/12 in Raphael et al. 2008, ASTN2-PTPRG and NTNG1- BCAS1 in Bashir et al. 2008).

Sample ID	5' Gene	5' Chr	3' Gene	3' Chr	DNA-PET cluster size	Structure variation		Distance between fusion points (kb)
						Connection	Super-cluster	
Breast Tumor 1	*PITPNB*	chr22	*MN1*	chr22	3	Del	1	124
Breast Tumor 2	*TK2*	chr16	*FTO*	chr16	10	Inv	2	12816
	ZNF341	chr20	*CBFA2T2*	chr20	10	TD	1	120
	LRRC57	chr15	*UBR1*	chr15	6	TD	1	472
Breast Tumor 13	*RNF24*	chr20	*SIGLEC1*	chr20	4	Del	1	287
	BCKDHB	chr6	*IRAK1BP1*	chr6	7	TD	2	1441
	WWOX	chr16	*ZNF571*	chr19	3	Isolated Translocation	1	
Breast Tumor 14	*UHRF2*	chr9	*PDCD1LG2*	chr9	4	TD	3	897

Supplemental Table 2B.

List of identified fusion transcripts by RT-PCR in primary breast tumors. DNA-PET cluster size shows number of discordant PETs which form the cluster. Super-cluster are created by clustering neighboring DNA-PET clusters (Hillmer et al., submitted), indicating the complexity of the environment of each structure variation. We defined clusters with a super-cluster size of more than 3 as complex structures. Del=Deletion; Inv= Inversion; TD=Tandem Duplication.

		5' Gene	5' Chr	3' Gene	3' Chr	DNA-PET cluster size	Structure variation		Distance between fusion points (kb)	GIS-PET count		RNA-PET count		Non-annotated region	Opposite strand gene
							Connection	Super-cluster		SHC012	SHC025	IHM098	IHM101		
MCF-7	3'T-N	*BCAS1*	chr20	NG	chr20	90	Del (complex)	205	5928				6	Inter-genic	
		PREX1	chr20	NG	chr20	27	Del (complex)	205	1191		1	2	16	Inter-genic	
		NCOA3	chr20	NG	chr20	11	Del (complex)	205	5947				2		
		NAV1	chr1	NG	chr1	37*	Del	1	255			14	5	Inter-genic	
		CDKN2A	chr9	NG	chr9	14	Del	2	171				6	Anti-sense	*MTAP* (intron)
		HPS4	chr22	NG	chr22	13	Del	1	27				3	Anti-sense	*ASPHD2* (include exon)
		RSBN1	chr1	NG	chr1	63*	U-Inv (complex)	205					2	Anti-sense	*VAV3* (intron)
		ATP9A	chr20	NG	chr20	14	U-Inv (complex)	205			12	2	9	Inter-genic	
		DMRT2	chr9	NG	chr9	49	U-Inv	1	161				5	Inter-genic	
		TRIM33	chr1	NG	chr1	10	U-Inv	1	11	13	4	3	10	Anti-sense	*TRIM33* (Intron)
		PKP2	chr12	NG	chr12	10	U-Inv	1	15			3	14	Inter-genic	
		GRHL2	chr8	NG	chr8	30	TD	2	90				7	Inter-genic	
		FBRSL1	chr12	NG	chr12	25	TD	1	68			2	6	Inter-genic	
		C10orf30 (BEND7)	chr10	NG	chr10	22*	TD	1	122	4				Anti-sense	*PRPF18* (include exon)
		PRPF18	chr10	NG	chr10	22*	TD	1	122				10	Anti-sense	*BEND7* (intron)
		TRAPPC4	chr11	NG	chr11	19	TD	2	1506				7	Anti-sense	*DSCAML1* (include exon)
		DDHD1	chr14	NG	chr14	17	TD	1	100				2	Inter-genic	
		C1orf144	chr1	NG	chr1	13	TD	1	54	2	31	19	49	Anti-sense	*FBXO42* (intron)
		CHEK2	chr22	NG	chr22	11*	TD	3	96				2	Inter-genic	
		FOXA1	chr14	NG	chr14	3	TD	3	118	38	43	62	187	Anti-sense	*C14orf25* (include exon)
		BCAS3	chr17	NG	chr3	309*	Complex-Inter	205		2	1	5	10	Inter-genic	
		NCOA3	chr20	NG	chr1	79	Complex-Inter	205		2	26	50	63	Inter-genic	
		CADPS	chr3	NG	chr1	73	Complex-Inter	205					3	Inter-genic	
		PHTF1	chr1	NG	chr20	26*	Complex-Inter	205				2		Anti-sense	*SGK2* (intron)
		IFT52	chr20	NG	chr1	26*	Complex-Inter	205				2		Anti-sense	*PHTF1* (intron)
		TRIM37	chr17	NG	chr20	22	Complex-Inter	205			13			Inter-genic	
		TBL1XR1	chr3	NG	chr6	14*	Complex-Inter	205				8	6	Inter-genic	
		TPD52L2	chr20	NG	chr17	7	Complex-Inter	205		3		14	22	Inter-genic	
		AHCYL1	chr1	NG	chr20	6	Complex-Inter	205			8			Inter-genic	
		ATXN7	chr3	NG	chr1	363	Isolated Translocation	1		9		2	42	Inter-genic	
		SLC25A19 (UCP3)	chr17	NG	chr20	25	Isolated Translocation	1				3	3	Inter-genic	
		SPTLC1	chr9	NG	chr1	19	Isolated Translocation	1		5				Anti-sense	*IGSF2* (include exon)
		SMARCC1	chr3	NG	chr4	11	Isolated Translocation	1			8	2		Inter-genic	
		DHX30	chr3	NG	chr4	10	Isolated Translocation	1			1		4	Anti-sense	*EMCN* (include exon)

Supplemental Table 2C.

List of identified 3'T-N^R transcripts in MCF-7. 'Non-annotated region' and 'opposite strand gene' column show whether the 3' part of transcripts are in intergenic regions or on anti-sense strands (in exons or introns) of genes. Asterisk * identifies genomic rearrangements producing multiple fusion transcripts because of the overlapping gene annotation in the fusion point. Del=Deletion; U-Inv=Unpaired-Inversion; TD=Tandem Duplication; Complex-Inter=Inter-chromosomal connection in hot spot of genome break-points (super-cluster size ≥3).

	GO	5'T + IR (625 genes)			3'T + FG (518 genes)			FG (235 genes)		
		Number of genes	Expected	P-value	Number of genes	Expected	P-value	Number of genes	Expected	P-value
Common	**Cellular process**	278	219.9	1.01E-06	228	182.2	2.03E-05	110	82.7	1.59E-04
	Cell communication	211	156.0	5.27E-07	169	129.3	5.34E-05	81	58.7	7.08E-04
	Cell motion	60	35.4	6.16E-05	51	29.4	1.12E-04	26	13.3	9.34E-04
	Cell-cell adhesion	60	29.9	3.79E-07	50	24.8	2.84E-06	22	11.2	2.23E-03
	Signal transduction	202	149.9	1.43E-06	158	124.2	4.20E-04	75	56.4	3.51E-03
Specific to 5'T + IR	**System development**	111	73.7	8.14E-06	75	61.1	3.63E-02	33	28	1.65E-01
	(P-value vs 3'T +FG)			3.21E-03						
	(P-value vs FG)			1.47E-03						
	Nervous system development	83	46.9	3.86E-07	48	38.9	7.81E-02	19	17.6	4.02E-01
	(P-value vs 3'T +FG)			2.24E-04						
	(P-value vs FG)			3.01E-06						
	Ectoderm development	87	53.2	4.62E-06	54	44.1	7.19E-02	20	20.0	5.33E-01
	(P-value vs 3'T +FG)			1.09E-03						
	(P-value vs FG)			2.06E-06						

Supplemental Table 3.

Result of GO analysis done by PANTHER (http://www.pantherdb.org/) for all genes involved in structural mutation categorized by the gene rearrangement classes defined in the glossary. Genes were separated into 5'-terminus truncations (5'T) + Intragenic Rearranged (IR), 3'-terminus truncations (3'T) + Fusion Genes (FG), and FG, and compared with all RefSeq coding genes. GO terms of Biological Process enriched ($P < 0.005$) in common and specific to 5'T + IR categories are shown. Expected number of genes based on the reference (all RefSeq genes) and P-value determined by binomial distribution were provided by PANTHER. GO terms generally involving cell adhesion and cell signaling are commonly found in all the categories, while some development-related terms are enriched in 5'T + IR categories. Note that none of the GO term was specific to 3'T + FG nor FG alone.

		Genomic structural variation									
	Gene rearrangement	Del	TD	U-Inv	Isolated Transloc	Inv	Ins-Intra	Ins-Inter	Complex-Intra	Complex-Inter	Total
Primary tumors	NG	395	38	64	23	6	7	2	23	12	570
	IR	210	15	0	0	0	0	0	2	0	227
	5'T	21	25	28	10	2	3	2	3	1	95
	3'T	29	13	31	22	0	1	2	6	1	105
	FG	14	16	9	3	1	3	5	5	0	56
Cell lines	NG	301	81	85	31	12	7	3	102	46	668
	IR	154	41	0	0	1	0	0	13	0	209
	5'T	53	72	49	33	4	2	1	76	52	342
	3'T	42	63	45	33	4	0	0	78	45	310
	FG	17	24	12	14	0	1	2	18	17	105

Supplemental Table 4.

Number of structure variations in each gene rearrangement category of primary tumors and cell lines. Del=Deletion; TD=Tandem Duplication; U-Inv=Unpaired-Inversion; Isolated Transloc=Isolated Translocation; Inv=Inversion; Ins-Intra=Intra-chromosomal Insertion; Ins-Inter=Inter-chromosomal Insertion; Complex-Intra= Intra-chromosomal connection in hot spot of genome break-points (super cluster size ≥3); Complex-Inter= Inter-chromosomal connection in hot spot of genome break-points (super cluster size ≥3) (Hillmer et al.; submitted). NG= Non-annotated Gene Region; IR= Intragenic Rearrangement; 5'T= 5'-terminus Truncation; 3'T= 3'-terminus Truncation; FG= Fusion Gene.

		Kozak ((A/G)xxATGG)			Kozak (-)		
Translational index		High	Medium	Low	High	Medium	Low
In-frame	Innate ATG	4	1	0	4	1	2
5' UTR	Innate ATG	0	0	0	0	1	0
	de novo ATG	1	1	0	3	0	0
Out-of-frame	Innate ATG (*)	0	0	0	0	0	2
	de novo ATG	0	1	0	1	0	7
Total		5	3	0	8	2	11

Supplemental Table 5.

Presence of Kozak sequence ((A/G)xxATGG) and the relation with translational index in each category of FG^R and 3'T-E^R transcripts. The fraction of transcripts with a high translational index is lower in transcripts without Kozak (8/21 (38%) vs. 5/8 (63%) with Kozak, P=0.013). *ATG codon of largest ORF is used.

5' gene domain	3' gene domain	Frame of fusion	MCF-7-FG	SKBR3-FG	T47D-FG	Tumors-FG	FG-Total	MCF-7-3'T	Total
(+)	(+)	In	2	1	0	1	4	0	4
		Coding exon to 5' UTR	0	2	0	0	2	0	2
		Out	1	2	0	1	4	0	4
	Truncated / (-)	In	4	1	0	1	6	0	6
		Out or NG	3	1	1	2	7	6	13
Truncated / (-)	(+)	In	1	0	1	0	2	0	2
		5'UTR to 5' UTR	2	1	0	0	3	0	3
		5' UTR to Coding exon	0	0	1	0	1	0	1
		Coding exon to 5' UTR	0	2	0	0	2	0	2
		Out	4	3	0	0	7	0	7
	Truncated / (-)		26	5	5	3	39	28	67
Total			43	18	8	8	77	34	111

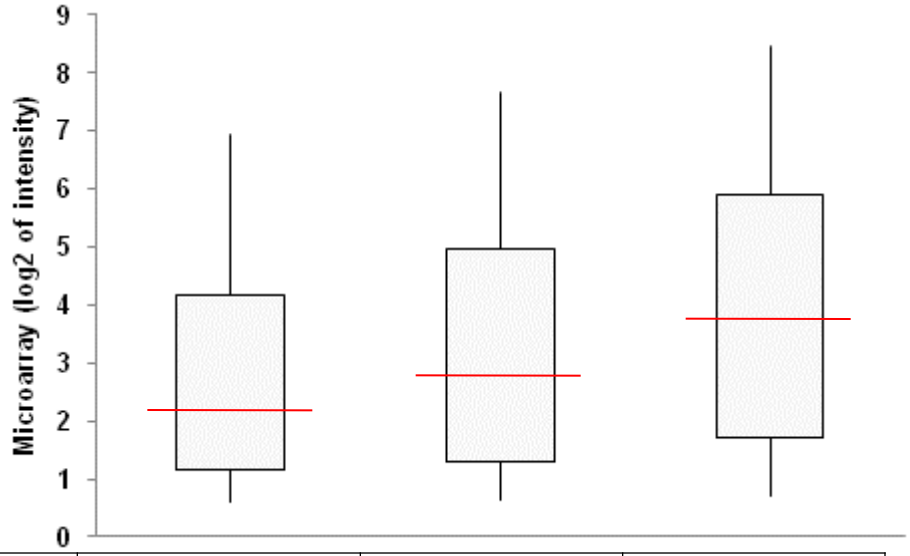
Supplemental Table 6.

Numbers of validated fusion transcripts having predicted biochemical functional domains. Functional domains were defined by SMART and Pfam domains. FG=Fusion Gene; 3'T=3'-terminus truncations; NG=fused with non-annotated gene region.

Family	Pfam ID	P value	Fusions
WD40	PF00400	2.04E-06	***WDR67****-AK298294 [SLC30A8],* ***WDR67****-ZNF704, PBRM1-****WDR82***
RhoGEF	PF00621	2.63E-05	***PREX1****-NG****, TRIO****-FBXL7, PLDN-****AKAP13****,* ***VAV3****-AK123199*
DEP	PF00610	4.51E-05	***PREX1****-NG,* ***DEPDC1B****-ELOVL7,* ***DEPDC1B****-PDE4D*
Pkinase	PF00069	8.53E-05	***TEX14****-PTPRG, RANBP10-****PSKH1****, ATAD5-****TLK2***

Supplemental Table 7.

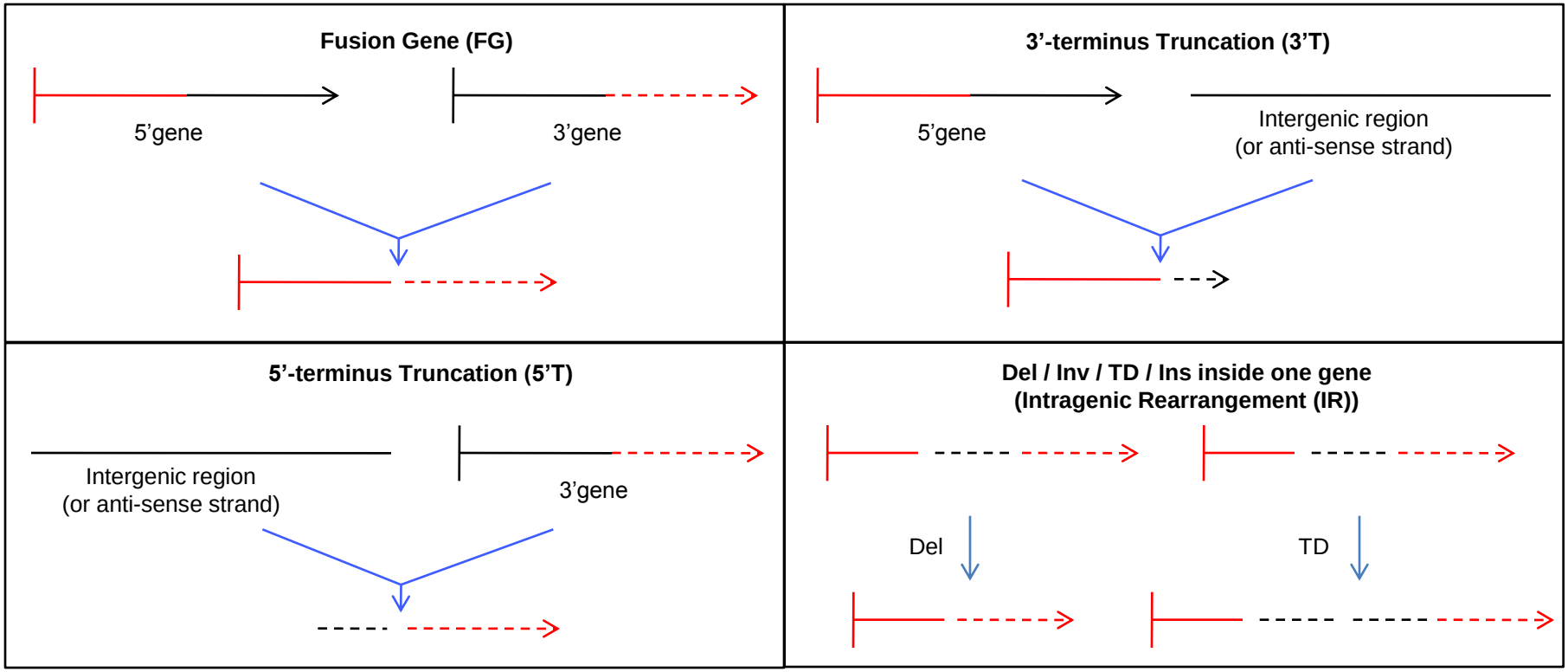
Enriched Pfam (http://pfam.sanger.ac.uk/) functional domains present in the validated fusion transcripts (P<5E-4). Total of 76 domains (57 IDs) in fusion transcripts were compared with 41,994 NCBI all Pfam domains (4,258 IDs) (http://spock.genes.nig.ac.jp/~genome/gtop.html).



Copy number	Reduced	Normal	Amplified
Number of genes	1373	9967	4568
Median	2.15	2.72	3.69
P-value (vs Reduced)		2.63 E-16	2.12 E-67
P-value (vs Normal)			1.80 E-55

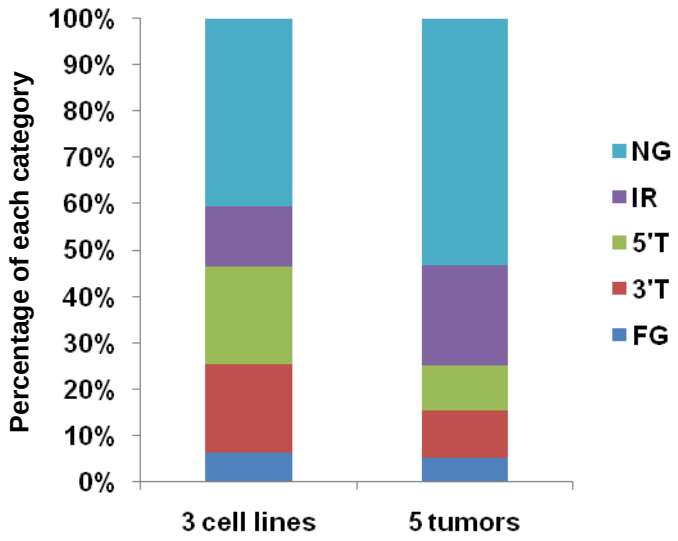
Supplemental Figure 1.

Expression level of genes in amplified and copy number-reduced regions as ascertained by tag counts. Relative expression levels (microarray signal $\log_2$ intensity) of 15,908 genes in MCF-7 were compared between genes in normal copy number (=2 copies), reduced ($\leq$1 copy), and amplified ($\geq$3 copies) regions. Copy number was determined by uniquely mapped PET tag counts (see Methods). Box shows 0.25 and 0.75 percentiles, vertical lines show 0.05 to 0.95 percentiles, and red lines show medians. The results show that sequence based copy number counts correlated with levels of gene expression.



Supplemental Figure 2A.

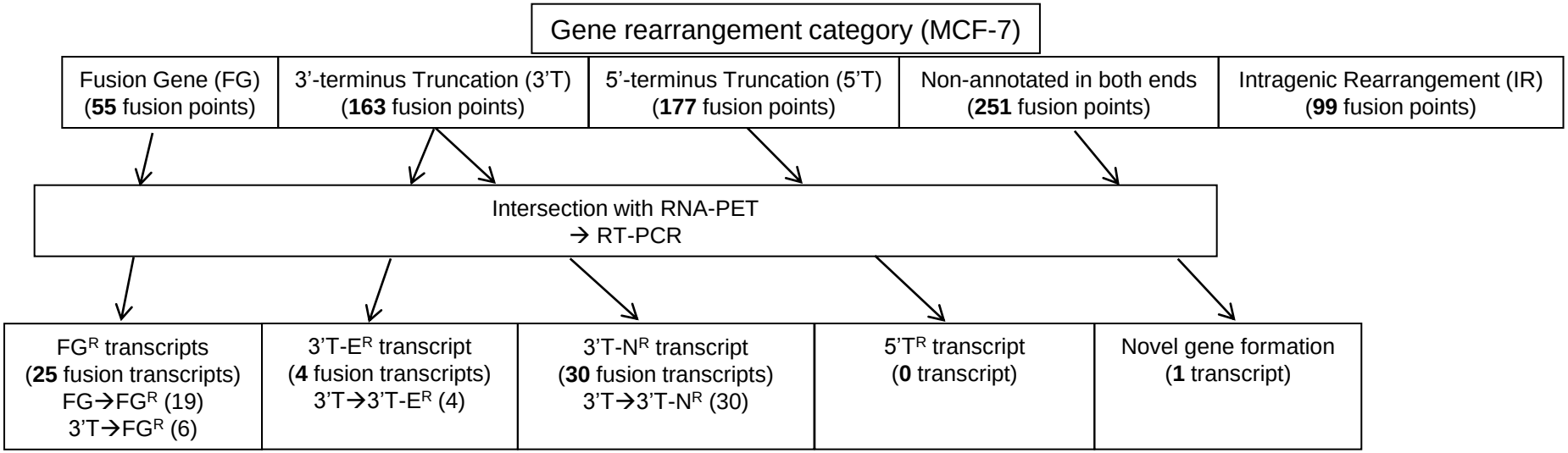
Schematic of the classes of genome rearrangements. The location of the genomic fusion points and the assessment of the directionality to gene components allowed us to categorize the gene rearrangement into four classes. Del=deletion; Inv= inversion; TD=tandem duplication; Ins= insertion.



		FG	3'T	5'T	IR	NG
Number	3 cell lines	105	310	342	209	668
	5 primary tumors	56	105	95	227	570
%	3 cell lines	6.4	19.0	20.9	12.8	40.9
	5 primary tumors	5.3	10.0	9.0	21.6	54.1

Supplemental Figure 2B.

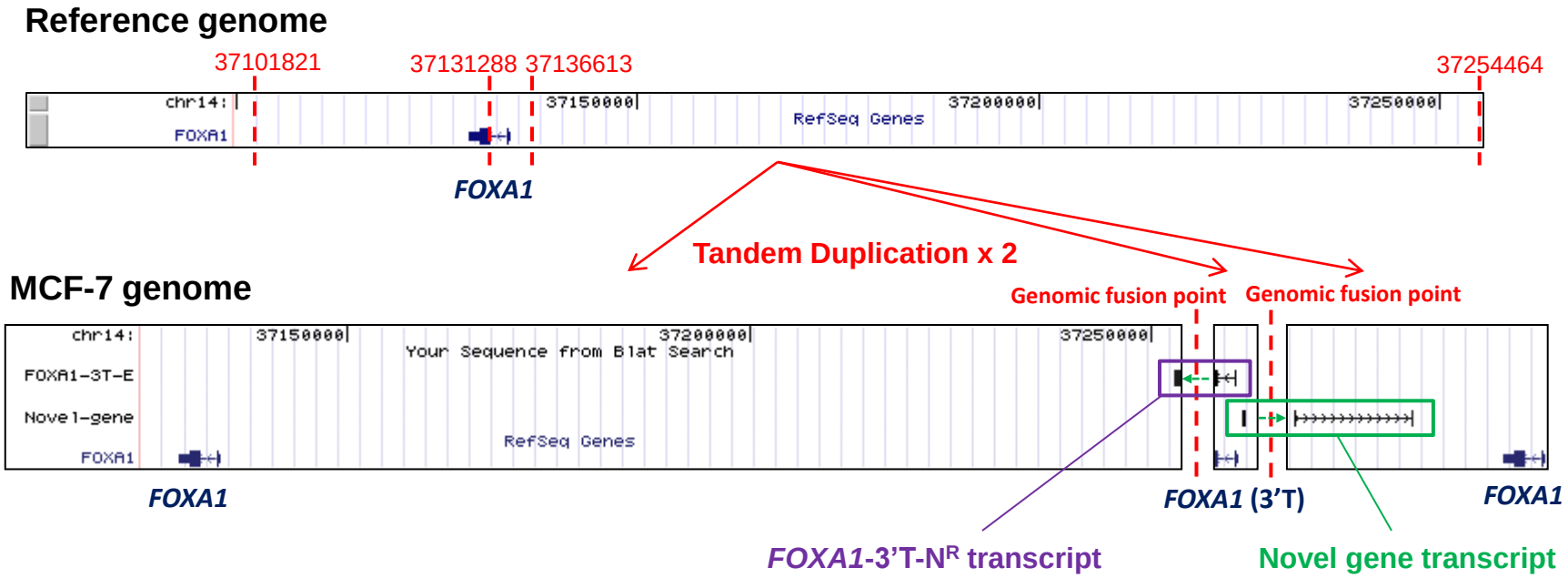
Number and Fraction of each gene rearrangement category in the cell lines and primary tumors studied. NG=Non-Genic region rearrangements; IR=Intragenic Rearrangement; 5'T=5'-terminus Truncation; 3'T=3'-terminus Truncation; FG=Fusion Gene. Primary tumors have proportionately more rearrangements in non-genic and intragenic regions than cell lines.



Library	FG^R	3'T-E^R	3'T-N^R	5'T^R	Novel gene formation
IHM098	11 (11)	0 (0)	17 (18)	0 (1)	1 (2)
IHM101	25 (25)	4(4)	27 (29)	0 (1)	1 (2)

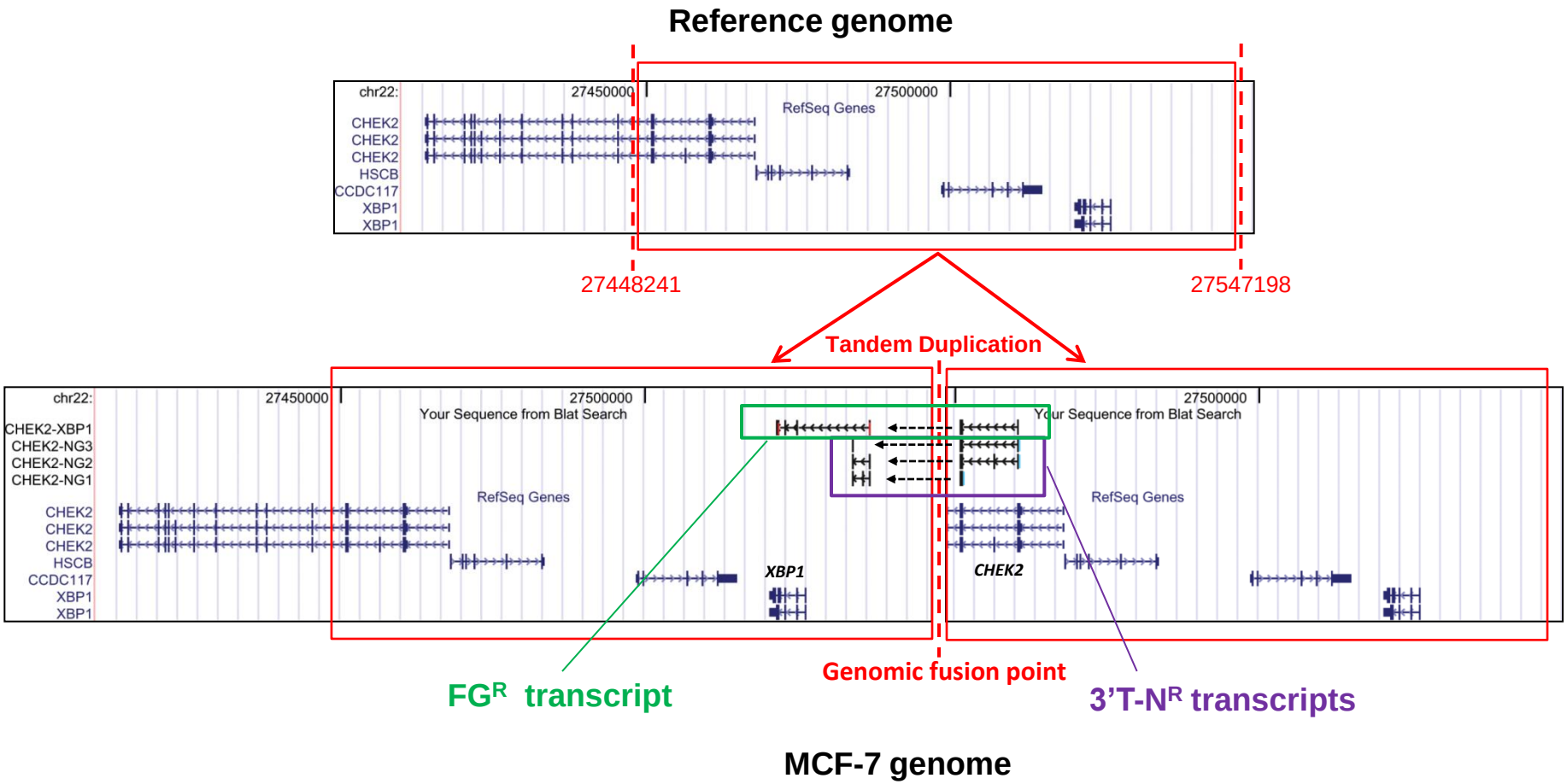
Supplemental Figure 2C.

Overview of the experimental framework to identify aberrant transcripts by the intersection of DNA-PET and RNA-PET in MCF-7 cells. Numbers of gene rearrangements in each class predicted by DNA-PET data are shown on the top. Numbers of aberrant transcripts which are predicted by the intersection with RNA-PET libraries and validated by RT-PCR in each transcript categories are shown in the middle. Based on the 3' part of fusion structure, 3'T^R transcript is separated in 2 sub-categories as shown in the glossary. The relationship between the genomic rearrangement and the validated transcript configuration of the resultant transcript is in the fusion transcript boxes. 3'T → FG^R represents a genomic rearrangement that was a 3'T structural mutation (5' partner is a RefSeq gene, and a 3' fusion point that, in this case, was outside the boundaries of any known gene or transcript, but that gave rise to an FG^R transcript (involving exons from 2 RefSeq genes) because it engaged the RefSeq gene adjacent to the genomic breakpoint. Numbers of validated transcripts in each RNA-PET libraries are shown in the bottom table and the numbers of candidate transcripts predicted by the intersection with DNA-PET and RT-PCR tested are given in the parentheses.



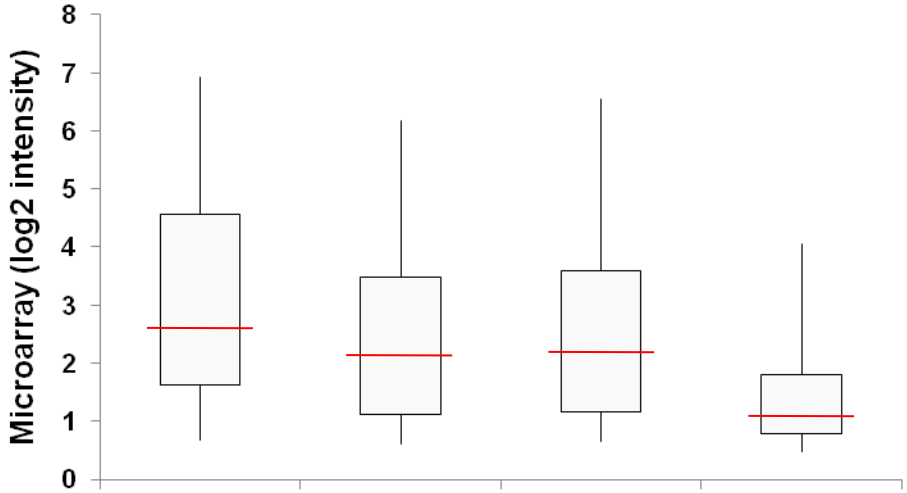
Supplemental Figure 2D.

FOXA1-3'T-N^R transcript (purple box) and novel gene transcript formation driven by bidirectional promoter of *FOXA1* (green box) caused by two tandem duplications. Genomic fusion points generated by the tandem duplications were determined by genomic PCR and shown as red dashed lines with coordinates (37254464 / 37131288 and 37136613 / 37101821). Tandem duplications cause local amplification of *FOXA1* gene as well as generating abnormal transcripts. Note that the novel gene is directed in the opposite orientation from *FOXA1*.



Supplemental Figure 2E.

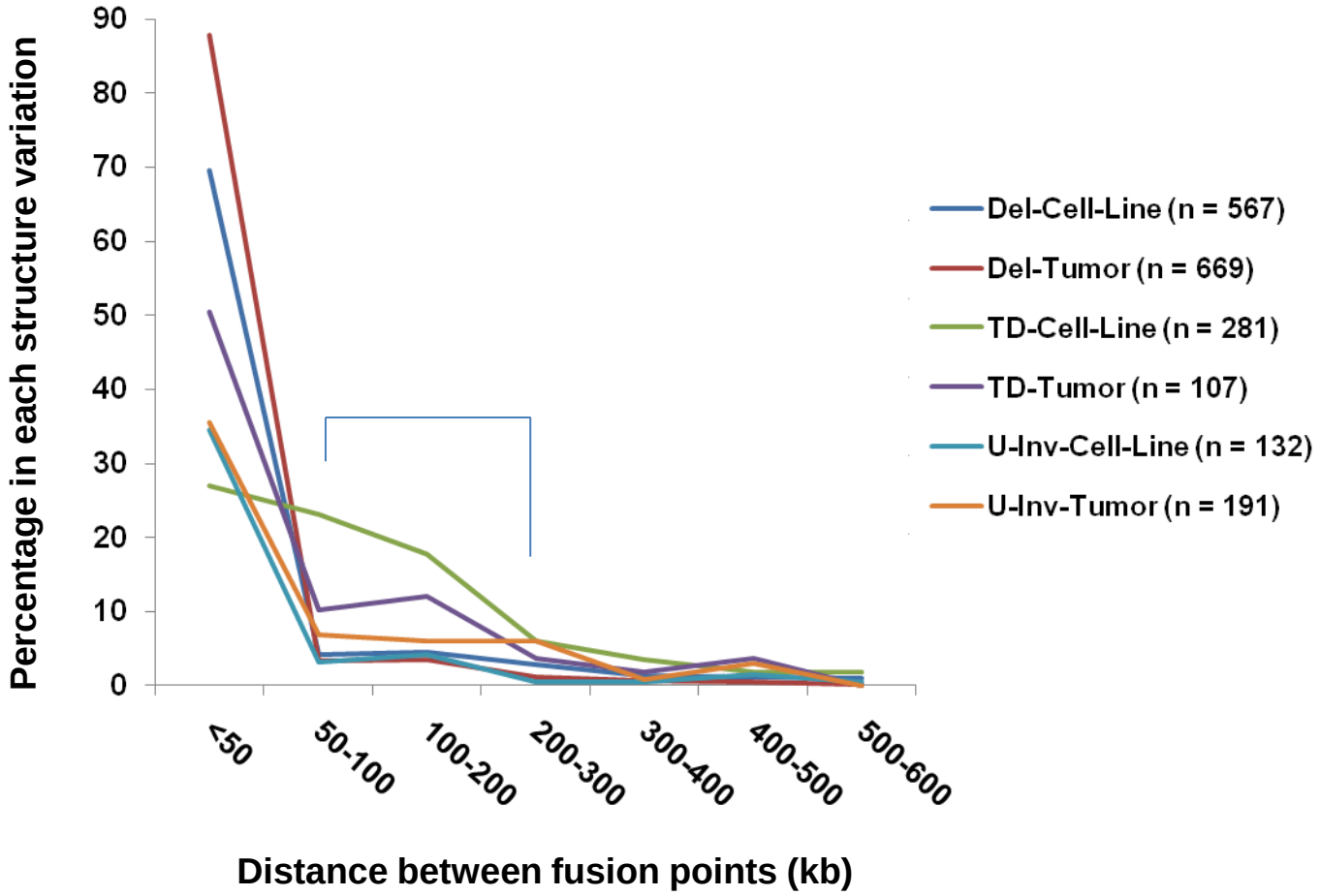
Example of FG^R and 3'T-N^R transcripts at the *CHEK2-XBP1* locus. Genomic fusion points produced by tandem duplication (duplication of red box) are determined by genomic PCR and shown as red dashed lines with coordinates. The FG^R transcript (green box) starts from *CHEK2* and creates a fusion to one exon in an intergenic region which is fused to exon 2 of *XBP1* (across its TSS). 3'T-N^R (purple box) start at a similar position but terminate in an intergenic region. We identified three splice variants in the intergenic region.



	FG	3'T	5'T	IR
Number	105	191	198	98
Median	2.58	2.15	2.21	1.12
P-value (vs IR)	1.10E-09	1.86E-07	8.12E-09	
P-value (vs 5'T)	5.70E-02	4.52E-01		
P-value (vs 3'T)	1.28E-02			

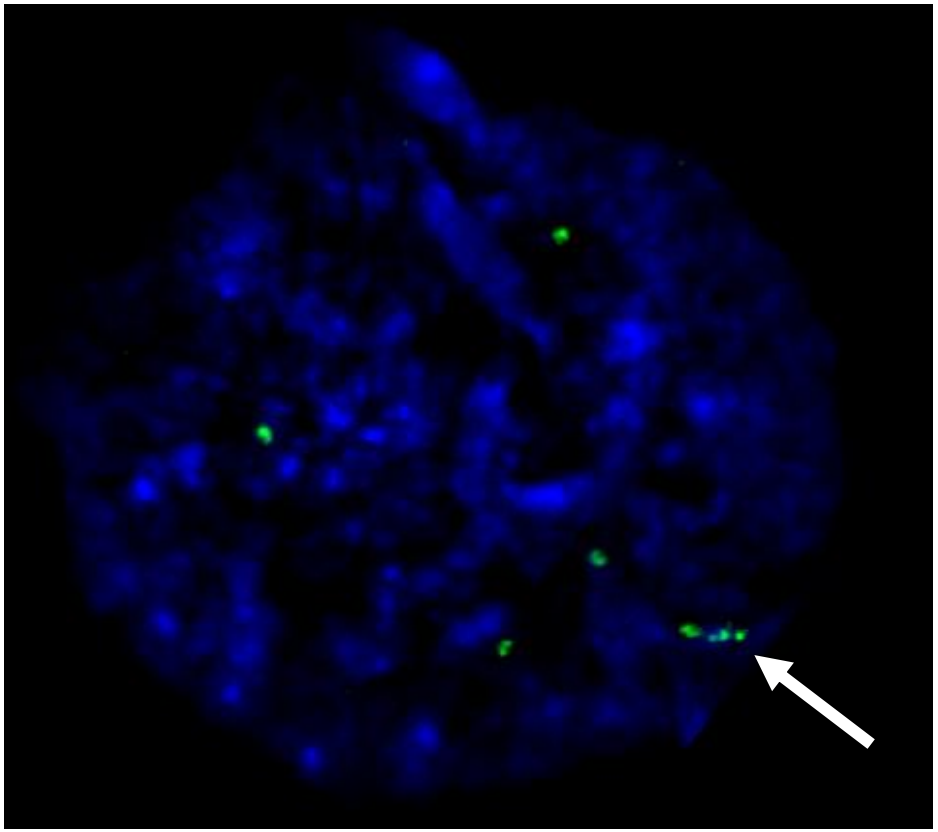
Supplemental Figure 2F.

Expression level (microarray intensity) of the genes (Refseq coding) truncated by each gene rearrangement category (IR=Intragenic Rearrangement, 5'T=5'-terminus truncation, 3'T=3'-terminus truncation, FG=Fusion Gene) in MCF-7. Fusion partners of FG show higher expression levels compared with IR, as well as 3'T. 3'T and 5'T genes also show higher expression levels compared with IR. Boxes show 0.25 and 0.75 percentiles, vertical lines show 0.05 to 0.95 percentiles, and red lines show medians.



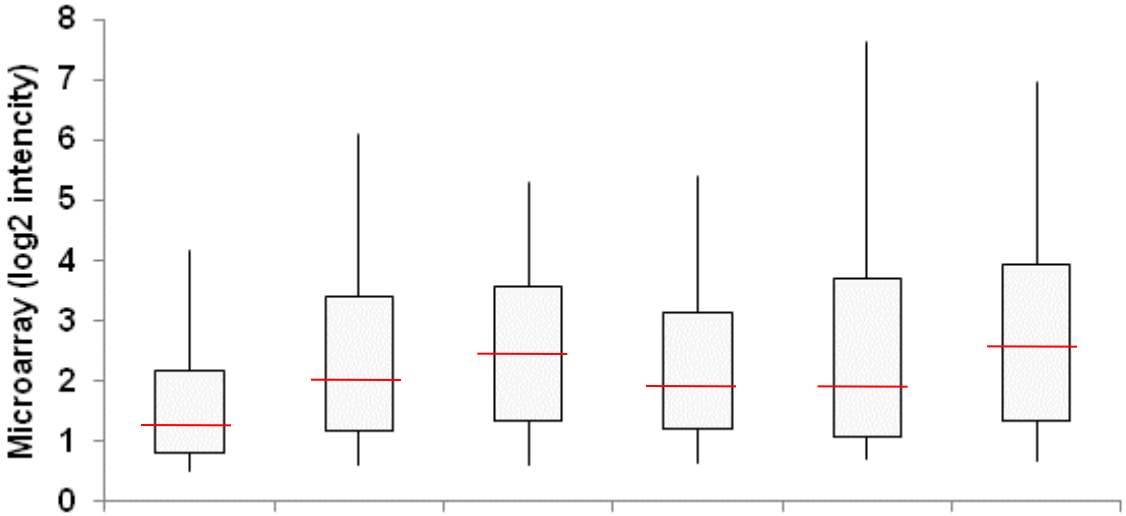
Supplemental Figure 2G.

Distance between fusion breakpoints in intra-chromosomal structure variations (SVs) indicating tendency to create fusion genes from neighboring genes. Percentages of the individual SV classes falling in the indicated windows of distances between the fusion points on the reference genome are shown. Tandem duplications are enriched (47% in cell lines, P=9.18E-34; 26% in tumors, P=2.62E-03) in the middle range (50-300kb windows) of the distances, compared with all intra-chromosomal structure variations (16%). Del=Deletion; TD=Tandem Duplication; U-Inv=Unpaired-Inversion.



Supplemental Figure 2H.

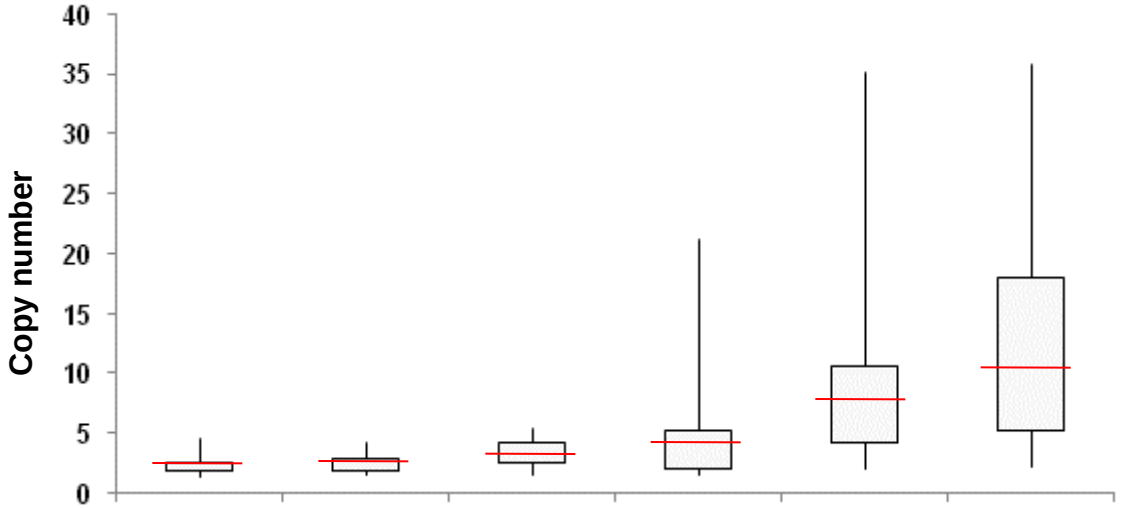
FISH analysis for *FOXA1* locus in MCF-7. Inter-phase of MCF-7 cell was hybridized with fosmid probe (G248P82033A5) covering the tandem duplication. Local amplification is indicated by the arrow.



	Deletion	TD	U-Inv	Isolated Translocation	Complex-Intra	Complex-Inter
Number	96	177	40	53	108	108
Median	1.26	1.99	2.45	1.92	1.92	2.58
P-value (vs Deletion)		3.75E-04	5.58E-03	2.17E-02	7.41E-05	8.94E-07
P-value (vs TD)			7.29E-01	6.61E-01	1.53E-01	2.41E-02
P-value (vs U-Inv)				5.24E-01	4.01E-01	1.56E-01
P-value (vs Transloc)					1.20E-01	2.66E-02
P-value (vs Complex-Intra)						5.73E-01

Supplemental Figure 2I.

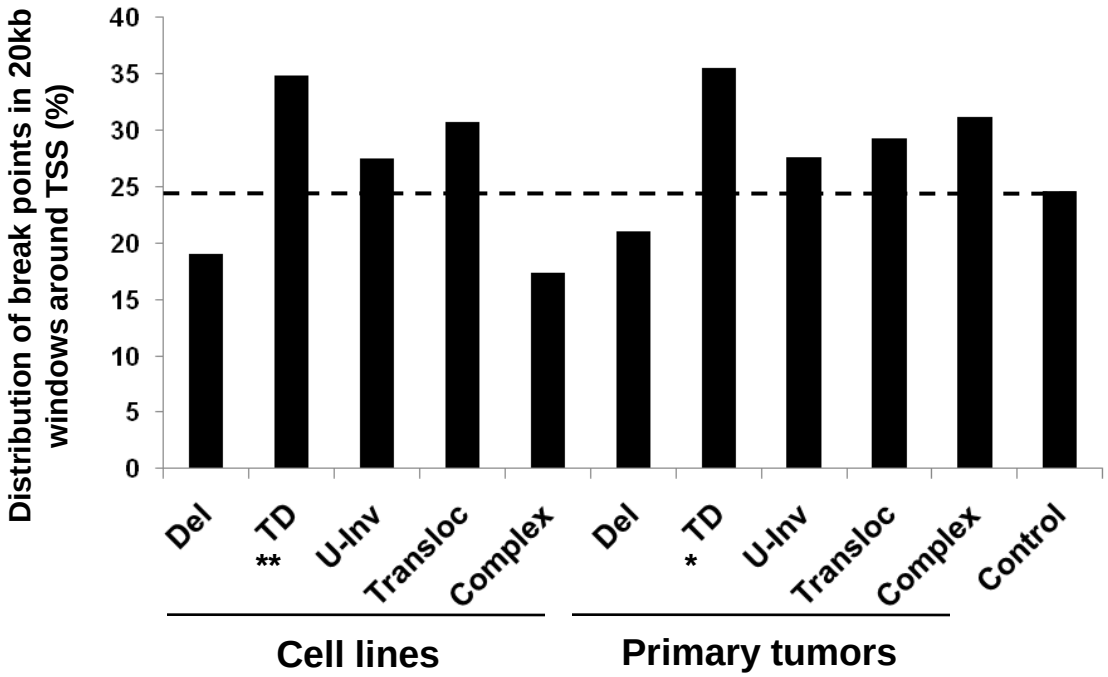
Expression level (microarray intensity) of the genes which are truncated by a structure variation in MCF-7. Genes which are affected by complex inter-chromosomal rearrangements (Complex-Inter), complex intra-chromosomal rearrangements (Complex-Intra), isolated translocations (Transloc), unpaired inversions (U-Inv), and tandem duplications (TD) show higher expression levels than genes affected by deletions. Boxes show 0.25 and 0.75 percentiles, vertical lines show 0.05 to 0.95 percentiles, and red lines show medians.



	Deletion	TD	U-Inv	Isolated Translocation	Complex-Intra	Complex-Inter
Number	95	178	40	53	108	108
Median	2.3	2.42	2.96	3.98	7.55	10.29
P-value (vs Deletion)		9.79E-01	1.76E-02	9.22E-04	1.14E-13	1.05E-18
P-value (vs TD)			3.69E-03	6.81E-04	5.74E-14	7.16E-19
P-value (vs U-Inv)				1.69E-02	2.30E-11	1.54E-16
P-value (vs Transloc)					1.93E-04	2.29E-08
P-value (vs Complex-Intra)						2.74E-02

Supplemental Figure 2J.

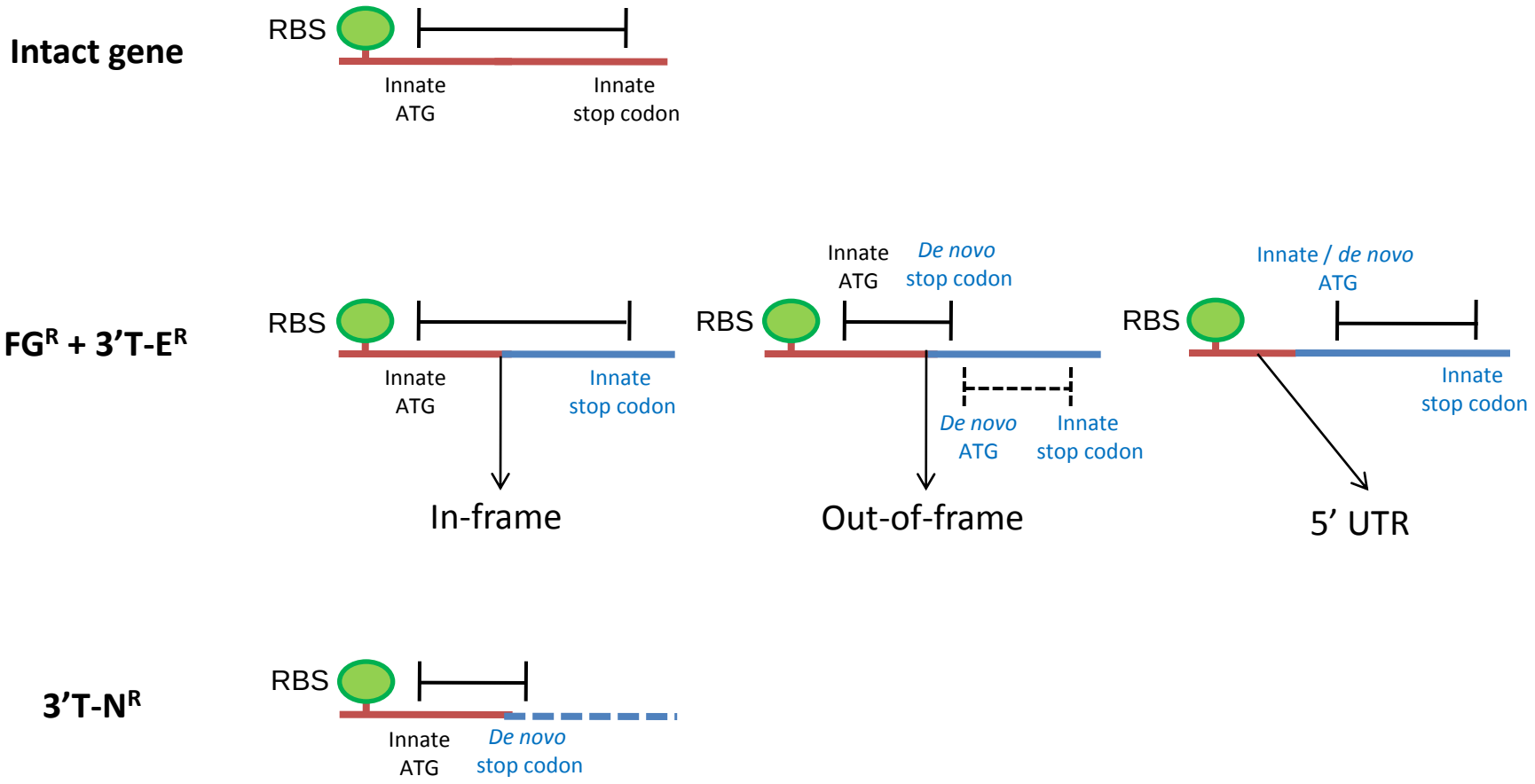
Copy number of the genes truncated by structure variations in MCF-7. The genomic copy numbers of genes which are affected by complex inter-chromosomal rearrangements (Complex-Inter), complex intra-chromosomal rearrangements (Complex-Intra), isolated translocations (Transloc), and unpaired inversions (U-Inv) are higher than those by tandem duplications (TD) and deletions. Note that the copy numbers for genes which are affected by deletions and TDs do not show significant differences, which is in contrast to the gene expression levels. This shows that instability of the genomic architecture is associated with gene amplification. Boxes show 0.25 and 0.75 percentiles, vertical lines show 0.05 to 0.95 percentiles, and red lines show medians.



		Cell lines					Primary tumors				
		Del	TD	U-Inv	Isolated Translocation	Complex	Del	TD	U-Inv	Isolated Translocation	Complex
Number of break points	Total	1134	562	382	218	894	1338	214	264	116	106
	TSS 20kb windows	216	196	105	67	155	281	76	73	34	33
P-value (vs control)		2.43E-07	6.00E-08	2.59E-02	1.01E-02	6.40E-09	5.36E-05	1.87E-04	3.50E-02	4.78E-02	3.14E-02

Supplemental Figure 2K.

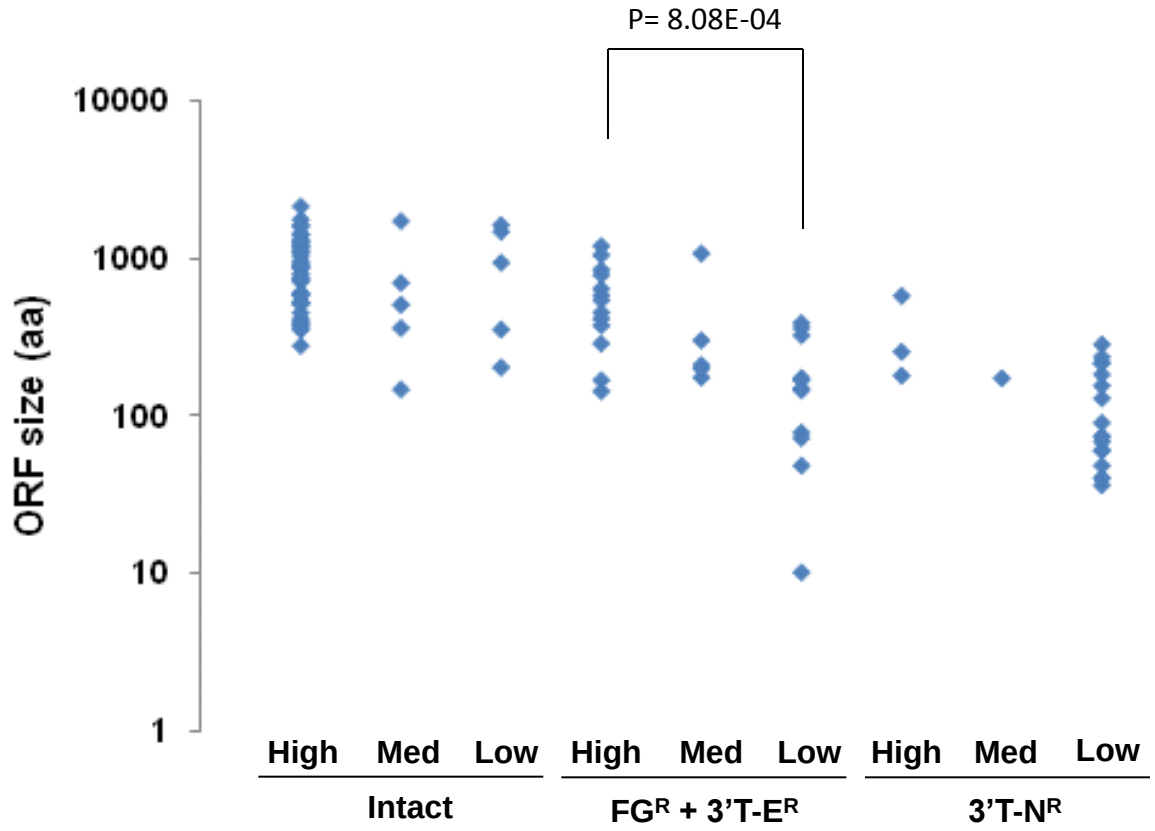
Distribution of genome break points in 20kb windows around transcription start sites (TSS) for each structure variation category. The expectation (horizontal dashed line) is calculated by the cumulative length of all 20kb TSS windows of all RefSeq coding genes relative to the size of the genome. Significant enrichments (P<0.005) of breakpoints around TSS are shown in TDs of cell lines (** P = 6.00E-8) and primary tumors (* P = 1.87E-4). This suggests that TDs may be selected for the disruption of regulatory regions. Del=Deletion; TD=Tandem Duplication; U-Inv=Unpaired-Inversion; Transloc=Isolated Translocation; Complex=intra- and inter-chromosomal connections in hot spot of genome break-points (super cluster size ≥3).



Supplemental Figure 3A.

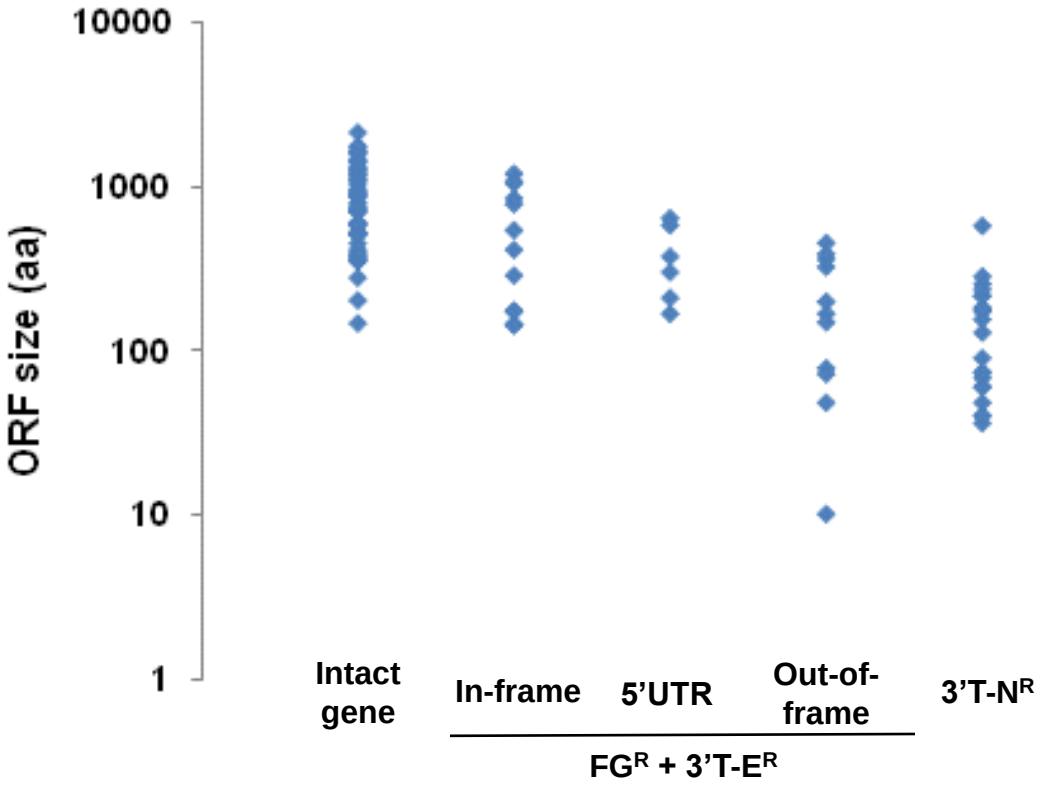
ORF structure of intact and fusion transcripts. FG^R and 3'T-E^R transcripts are separated into 3 categories:

1) the innate ORF of the 3' fusion partner is in-frame to the 5' ORF; 2) the innate ORF of the 3' fusion partner is out-of-frame to the 5' ORF or 3) the 5'UTR of the 5' fusion partner is fused to the 3' partner where the putative translation will start from the innate or *de novo* ATG codon. The 3'T-N^R structures usually have shorter ORFs by encountering stop codons within the non-transcribed and non-annotated fusion segments. RBS=Ribosomal Binding Site.



Supplemental Figure 3B.

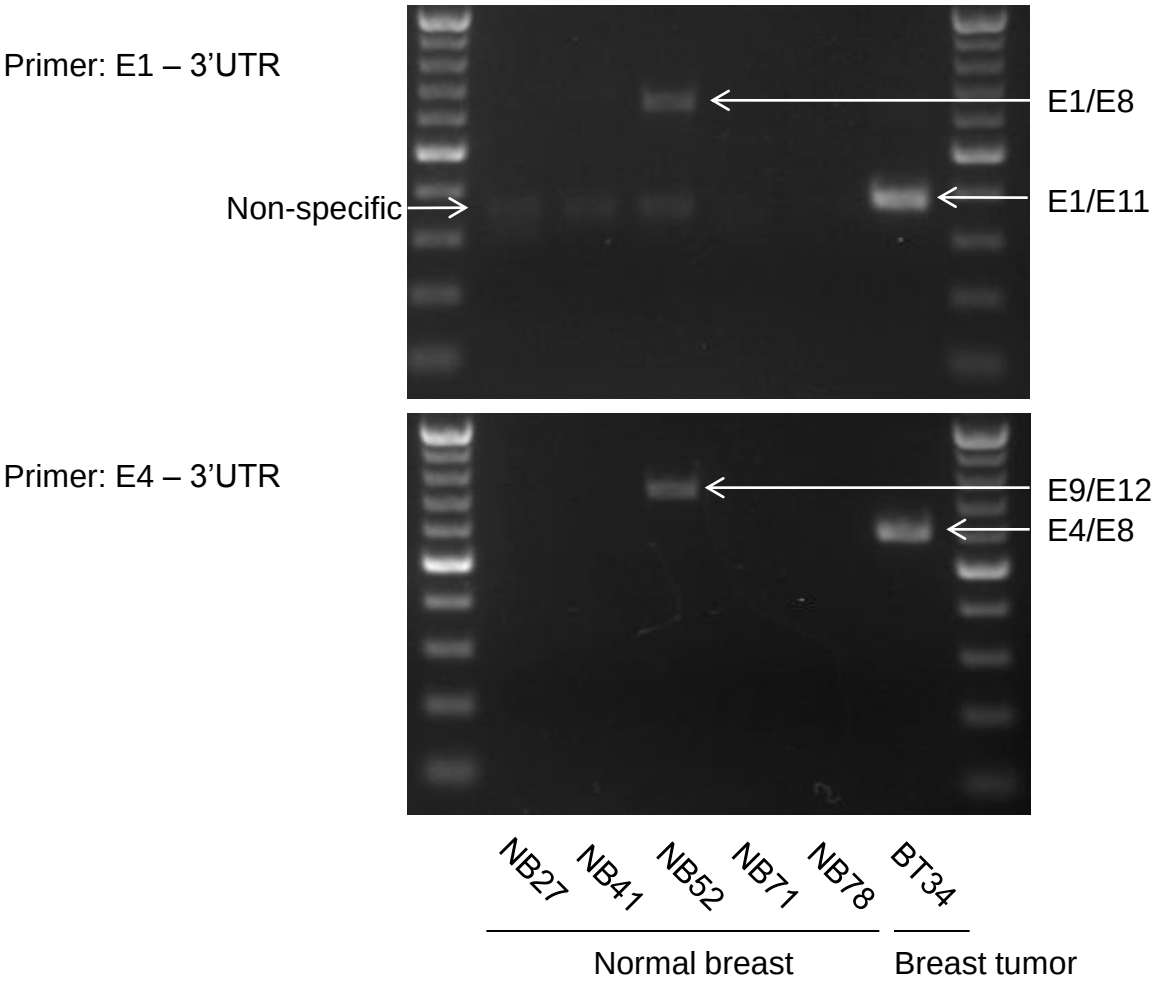
Comparison of ORF size between genes with High, Medium, or Low translational index (TI) in intact gene and fusion transcripts. Predicted ORFs are explained in Supplemental Figure 3A. For out-of-frame category, first ORF from 5' gene is used. FG^R + 3'T-E^R transcripts showed significant difference in ORF size between High and Low TI genes (P= 8.08E-04), while Low genes of 3'T-N^R showed a tendency to be smaller ORF but not significant. One limitation of polysomal fractionation might be that transcripts with small ORFs would engage smaller numbers of ribosomes even if translated. However, smallest ORF size of a High Translational Index gene is 143 and even when we separate out smaller ORFs than143, most genes in Out-of-frame and 3'T-N^R still show Low Index (71% and 60%, respectively) in contrast to In-frame and 5'UTR (18% and 0%, respectively). Thus we speculate that ORF size is not sole determinant of lower TI and the difference of 3'UTR structure (using *de novo* stop codons) also affects the translation in Out-of-frame and 3'T-N transcripts.



P-Value (vs. In-frame)	3.32 E-02			
P-Value (vs. 5'UTR)	2.95 E-04	1.96 E-01		
P-Value (vs. Out-of-frame)	3.90 E-11	1.11 E-02	9.16 E-02	
P-Value (vs. 3'T-N)	4.23 E-15	4.33 E-03	3.40 E-02	3.41 E-01

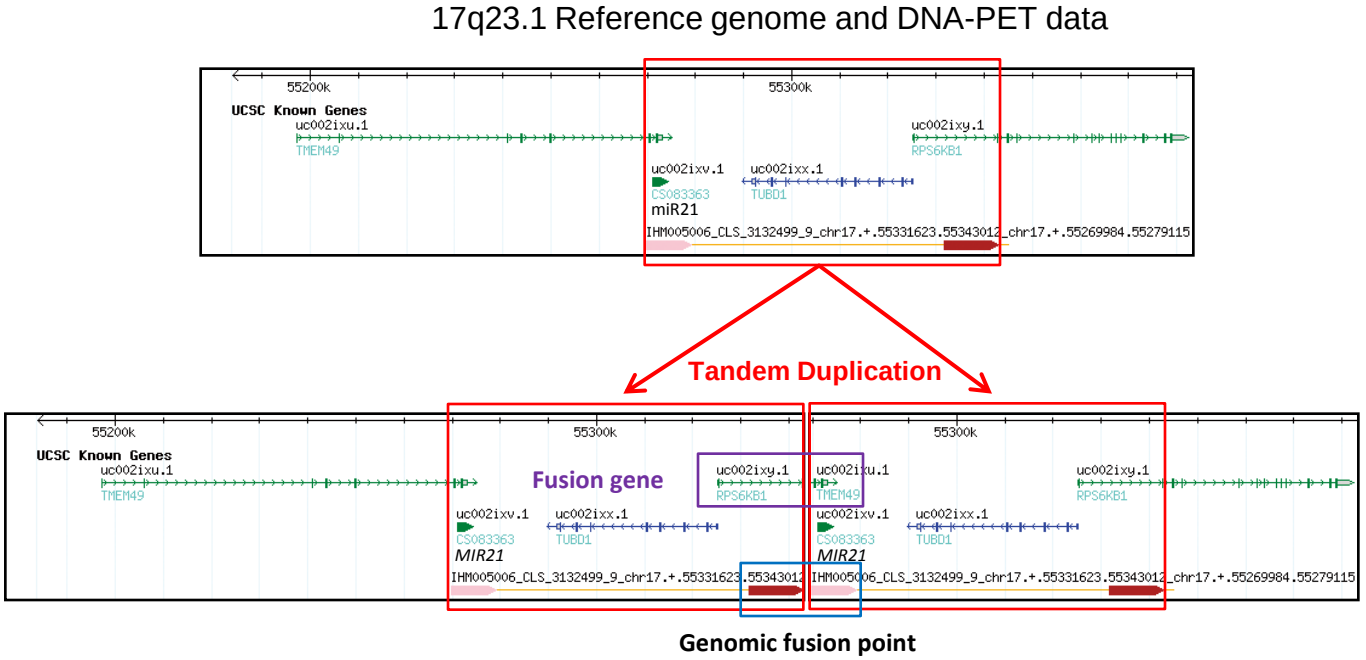
Supplemental Figure 3C.

Comparison of ORF size between different ORF categories. Intact gene transcripts show larger ORF than fusion transcripts, while ORFs of In-frame are larger than Out-of-frame and 3'T-N^R, and those of 5'UTR are larger than 3'T-N^R. Thus smaller ORFs are enriched in Out-of-frame and 3'T-N^R transcripts.



Supplemental Figure 6.

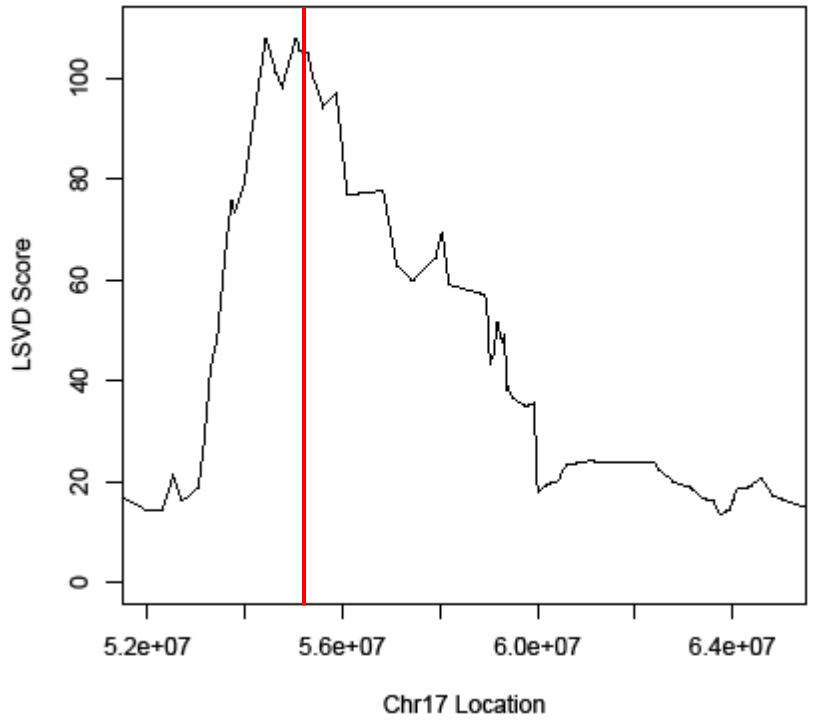
Expression of *RPS6KB1-VMP1* fusion transcript in normal breast RNA. Low level expressions of E1/E8 (fusion between exon 1 of *RPS6KB1* and exon 8 of *VMP1*) and E9/E12 fusion transcripts were observed in a normal breast (NB52), while strong expressions of E1/E11 and E4/E8 fusion transcripts were observed in a breast tumor serving as positive control (BT34). Note that sequencing analysis revealed the non-specific band in some lanes. Primers used are: 5' – AGACAGGGAAGCTGAGGACA - 3' and 5' - AACAGGAGCAAATACTGGGA - 3' match to exon 1 and exon 4 of *RPS6KB1* ,respectively, and 5' – CAGAACCCATCCACTCCAAT - 3' match to 3'UTR of *VMP1*. The primer matched to 3'UTR was used to deny any contaminations.



Genomic structure of *RPS6KB1-VMP1* fusion locus in MCF-7 genome

Supplemental Figure 4.

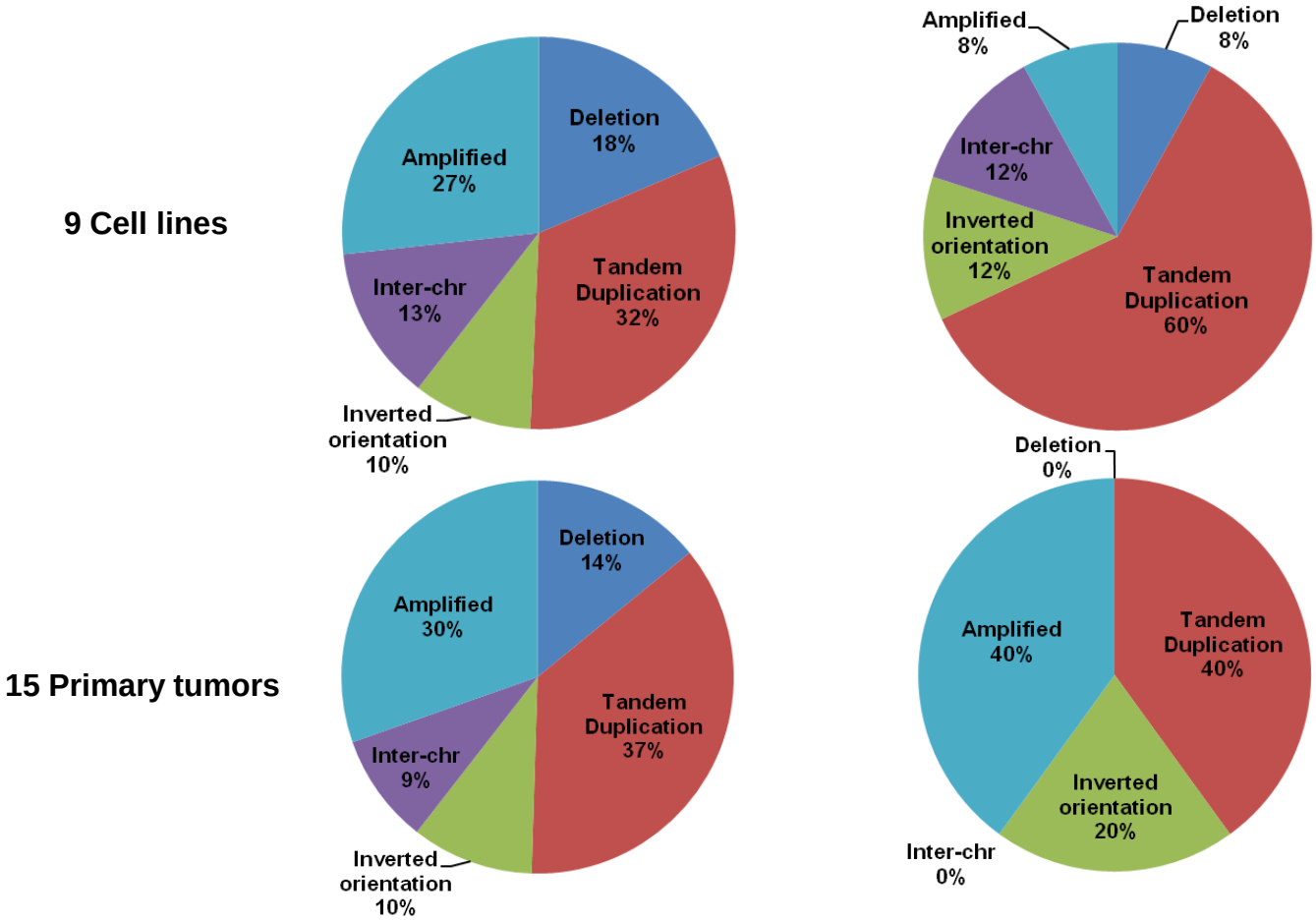
Local amplification of *MIR21* and *TUBD1* and creation of *RPS6KB1-VMP1 (TMEM49)* fusion gene (purple box) induced by a tandem duplication in MCF-7. In the top panel, brown and pink arrows indicate the coordinates of 5' and 3' anchors of a PET cluster on the reference genome, respectively, with directional information. The cluster information indicates the genomic fusion structure in MCF-7 (blue box in lower panel), showing a tandem duplication of the genomic region (red box) including *MIR21* and *TUBD1*, creating an FG fusion gene.



	NUH cohort		
	Fusion (+)	Fusion (-)	Total
Number of tumors	21	48	69
LSVD (+) tumors	16 (76%)	12 (25%)	28 (41%)

Supplemental Figure 5.

Local singular value decomposition (LSVD) analysis on 17q22-24 in 737 tumors of 4 cohorts (Zhang et al. 2009), indicating a coexpression of adjacent genes and a possible presence of an amplicon (top). Red line shows the position of the tandem duplication which results in the *RPS6KB1-VMP1* fusion. This states that this TD is the likely driver for the amplicon (Zhang et al. 2009). Number of LSVD (+) tumors in S6K-fusion (+) and (-) tumors of the Singapore cohort appear in the bottom of the figure. The fusion (+) tumors show higher fraction of LSVD (+) tumors (76% (16/21), P= 8.15E-04 vs. total.



Rearrangements	Deletion	Tandem Duplication	Inverted orientation	Inter-chromosomal	Amplified	Total
Cell lines	214	370	113	147	308	1152
Primaries	143	369	102	92	308	1014

Fusion gene transcripts	Deletion	Tandem Duplication	Inverted orientation	Inter-chromosomal	Amplified	Total
Cell lines	2	15	3	3	2	25
Primaries	0	4	2	0	4	10

Supplemental Figure 7.

Categorization of structure variations using the classes presented in this report of all somatic rearrangements and rearrangements generating fusion gene transcripts in the breast cancer genome data by Stephens et al (Stephens et all. 2009). The comparison reveals an overrepresentation of tandem duplication in fusion gene transcripts in both 9 cell lines and 15 primary tumors. These results suggest that tandem duplication is also a favored mechanism in the generation of fusion transcripts.