SUPPORTING ONLINE MATERIAL

An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer

Xiao-Song Wang[1, 2, 3], John R. Prensner[1, 3, 8], Guoan Chen[4, 8], Qi Cao[1, 3], Bo Han[1, 3], Saravana M Dhanasekaran[1, 3], Rakesh Ponnala[1], Xuhong Cao[1, 3], Sooryanarayana Varambally[1, 3, 7], Dafydd G. Thomas[3], Thomas J. Giordano[3], David G. Beer[4], Nallasivam Palanisamy[1, 3], Maureen A. Sartor[2], Gilbert S. Omenn[2#], Arul M. Chinnaiyan[1, 2, 3, 5, 6, 7#]

1. Michigan Center for Translational Pathology, Ann Arbor, MI, 48109, USA,
2. National Center for Integrative Biomedical Informatics, CCMB, MI, 48109, USA,
3. Department of Pathology, University of Michigan, Ann Arbor, MI, 48109, USA,
4. Department of Surgery, University of Michigan, Ann Arbor, MI, 48109, USA
5. Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, MI, 48109, USA
6. Department of Urology, University of Michigan, Ann Arbor, MI, 48109, USA,
7. Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI, 48109, USA
8. These author contributed equally

#Address correspondence and requests for reprints to: Gilbert S. Omenn, M.D., Ph.D. & Arul M. Chinnaiyan, M.D., Ph.D.

This PDF contains:

Nature Biotechnology: doi: 10.1038/nbt.1584

Supplementary Discussion

The functional significance of fusion-interaction hubs and their potential as drug targets

To explore the functional role of most significant shared interacting genes with the connecting fusion genes, we resolved the fusion-interaction network by setting the p value to 10^{-7}. After spring-embedded relaxing using the VisANT program [1], the FI network was spread into six major clusters, which join gene fusions from similar tumor entities (**Figure 1d)**. The shared interacting genes with the greatest statistical significance in each subset of connected fusions were designated as fusion-interaction hubs in each cluster.

In cluster i, this approach identified *GATA3* as the hub for a subset of T cell receptor fusion partners -- *LYL1, TAL1, TAL2, LMO1* and *LMO2*, which are normally transcriptional cofactors for *GATA3*[2]. Whereas *MEIS1* was the center of another subset of *NUP98* fusion partners, the *HOX* family genes, which were reported to collaborate with *MEIS1* in AML transformation[3]. In clusters ii and iii, *CDK6* and *CTNNB1* join two distinct subsets of immunoglobulin fusion partners, exemplified by cyclins and *BCL* genes. Interestingly, *CDK6* is known to form a complex with cyclins, that phosphorylates and inhibits Rb; whereas *CTNNB1* is a known upstream protein of the *BCL* genes[4]. In cluster v, *PIK3R1* connects most of *BCR* and *ETV6* fusion partners. This suggested the central role of *PIK3R1* in mediating the signaling of these fusion proteins. *HDAC1*, the hub for a subset of *RUNX1* fusions in cluster iv, is normally a co-repressor of the *RUNX1* partners -- *RUNX1T1, CBFA2T3, MDS1 and EVI1*[5]. Moreover, *ERG* and *RPS6KA5* (*MSK1*) appear to be the hubs of *ESWR1* fusions in cluster vi; the latter was reported to regulate *EWSR1* partners -- *CREB1, ETV1* and *ATF1*[6]. The consistent functional relationship between a FI hub and a fusion partner family suggests this hub as a functional factor joining this family. This implies the role of FI hubs in mediating the function of the fusion proteins, and the potential benefits of mining for drug targets from shared interacting hubs to block multiple fusions. For example, HDAC1 is recruited by 3' partners of RUNX1 resulting in dominant-negative effect over wild-type RUNX1[7, 8], thus may be a potential target to block RUNX1 fusions (**Figure 1d**).

The significance of the FI hubs in molecular targeting could be verified by the therapeutic effect of their ablating drugs on the specific tumor entities where the connected gene fusions occur. We therefore investigated the literature for the hubs that have blocking reagents, *PIK3R1*, *CDK6*, and *HDAC1*, according to the drug target database[9]. A most recent report revealed that *PIK3R1* and *PIK3R2* are the specific PI3K isoforms that mediate transformation of *BCR-ABL1* positive pre–B-ALL, and *PI3K* ablation by PI-103 can block *BCR-ABL* leukemogenesis in mice[10]. Moreover, the *HDAC1-3* inhibitor, Vorinostat, was reported to be effective in AML patients by a recent study[11], whereas Flavopiridol, an inhibitor of *CDKs*, was under clinical trial for the treatment of relapsed or refractory B cell lymphoma[12].

The application of ConSig technology to deep sequencing data analysis

The ConSig technology preferentially identifies biologically important genes in cancer. This is particularly useful in the analysis of a large number of putative chimeras generated by next generation sequencing data to filter secondary fusions. Of note, in this application, the main theme is to evaluate the biological relevance of putative chimeras, in stead of distinguishing fusions and mutations, whereby the radial ConSig will be more informative. This is especially important for evaluating the genes involved in both fusion and point mutations (mixed type cancer genes), for example, a fusion involving *EGFR* gene will be considered as biologically important because of the prior knowledge of *EGFR* point mutation in cancer.

Moreover, the 3' fusion partners display more distinctive signature concepts than the 5' partners (**Supplementary Fig. 4**), therefore the ConSig technology will be more discriminative in evaluating 3' genes. In

practice, we usually first rate the 3' partners of fusion chimeras by *r*ConSig scores, and then rate the 5' partners by *r*ConSig score to supplement this analysis.

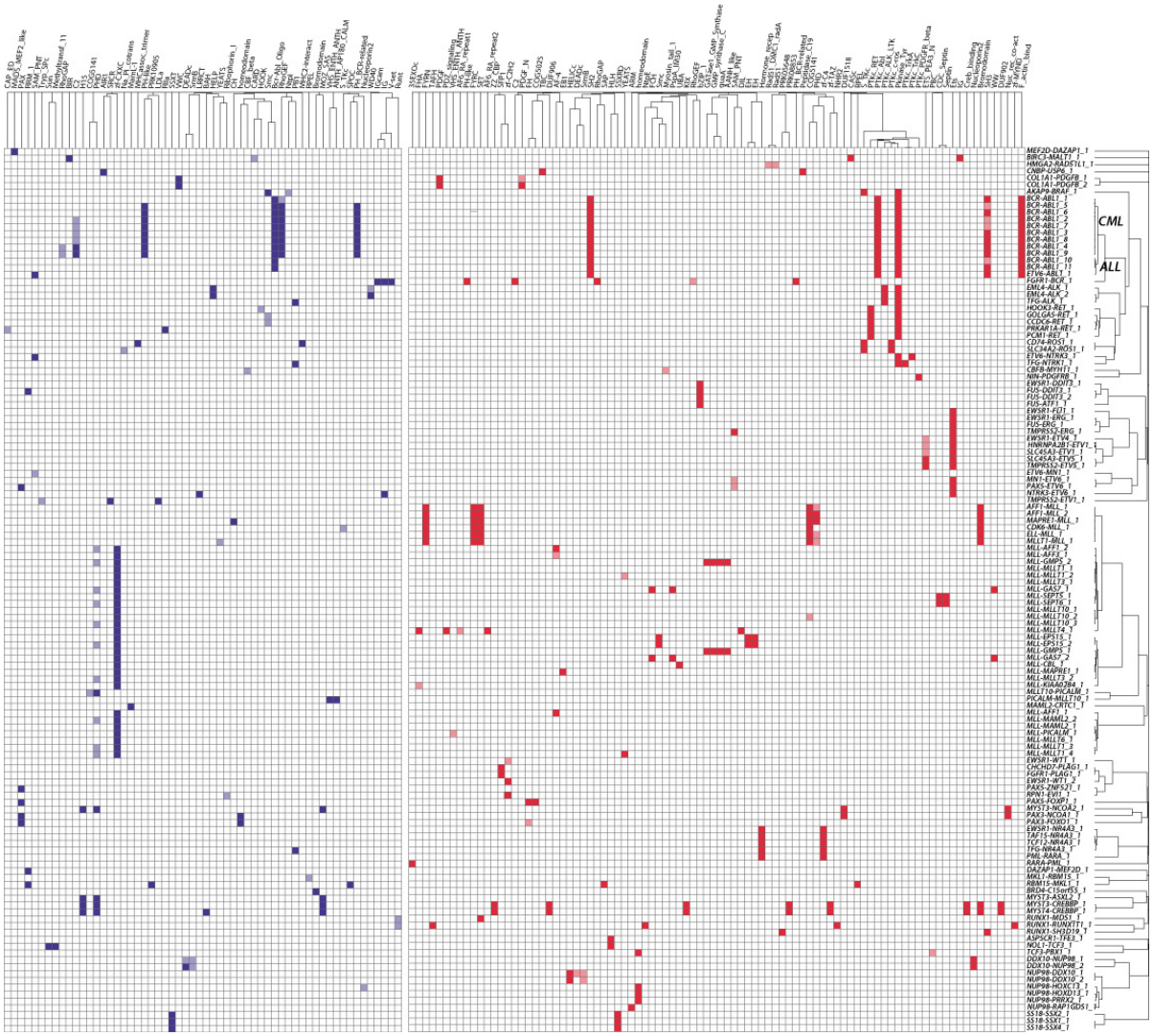
Confirmation of the fusion breakpoint principle

While the fusion breakpoint principle can be inferred based on conventional cytogenetics analysis, it should be stressed that the net output of high-throughput genomic measurement was different from G-banding and FISH, where the balanced genomic relocation information was lost. **Supplementary Fig. 6** demonstrates the possible complex chromosome rearrangements generating contradictory cases to the breakpoint principle on array CGH, but not on FISH data. For example, the THP-1 cell line harbors a MLL-AF9 fusion with duplication of the 3' MLL gene that deviates from the principle. Studying public spectral karyotyping data revealed complex translocations between chr9 and chr10 resulting in possible three-way fusions involving the MLL gene 13. For this reason, extensive evidence from large numbers of malignancies is required to confirm the prevalence of this principle on high-throughput genomic data.

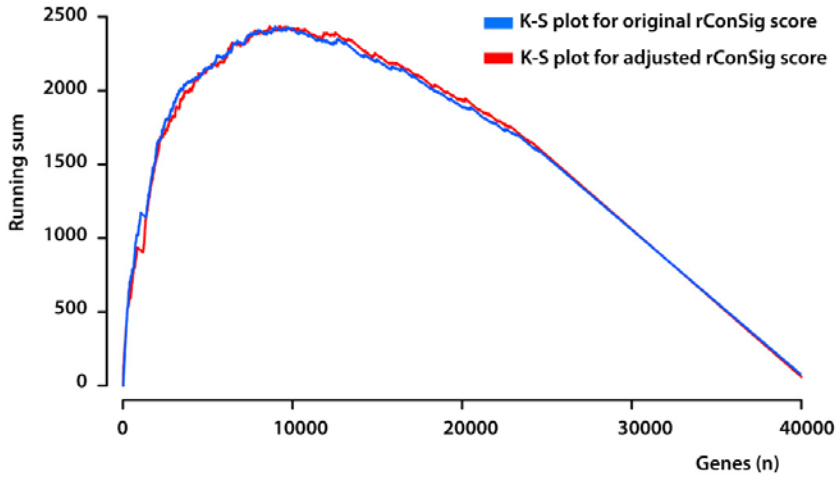
To confirm this principle, we did a large-scale meta-analysis of unbalanced gene fusions based on high-resolution array CGH/SNP datasets annotated with gene fusions, as well as literature curation (**Table 2, Supplementary Fig. 5b**). We started with analyses of four independent leukemia datasets and two lymphoma datasets. Of 32 samples with 10 different unbalanced fusions in these datasets, 31 follow the principle. Analyses of the known fusions in mesenchymal and epithelial tumors also yielded strong supporting evidences. In four sarcoma datasets, three unbalanced fusions were identified in 23 samples, including *EWSR1-FLI1* in Ewing's sarcoma, *COL1A1-PDGFB* fusion in dermatofibrosarcoma protuberans (DFSP), and *ASPSCR1-TFE3* in alveolar soft part sarcoma (ASPS). None of these contradict the principle. In salivary adenoma, a *FGFR1-PLAG1* fusion is found with interstitial duplications[14], whereas *TMPRSS2-ERG* fusion in prostate cancer is frequently reported having heterozygous deletions[15, 16]. This finding is noteworthy because both are intra-chromosome gene fusions, but the genomic placements of genes in the two fusions are clearly opposite. This clearly demonstrated the inferred principle. The same pattern was also observed in the recurrent *KIAA1549-BRAF* fusion in astrocytoma (AST)[17]. Furthermore, we analyzed the reports for all unbalanced intra-chromosome fusions from the Mitelman database[18], and confirmed that the inferred principle is the adherent genetic factor that determines the nature of genomic imbalances associated with these fusions (**Supplementary Table 9**).

To further test this principle on prostate cancer, where unbalanced fusions are less studied by conventional cytogenetics, we reviewed all fluorescence *in situ* hybridization that our lab has performed on prostate cancer for *ETS* family gene fusions with break-apart probes. A total of 60 samples with 5 unbalanced *ETS* fusions were found from 238 prostate cancer samples, including *TMPRSS2-ERG*, *TMPRSS2-ETV4*, *C15orf21-ETV1*, *HNRP-ETV1*, and *CANT1-ETV4*; no contradictory case was identified (**Supplementary Fig. 5c**, and **Supplementary Table 10**).

Supplementary Figures

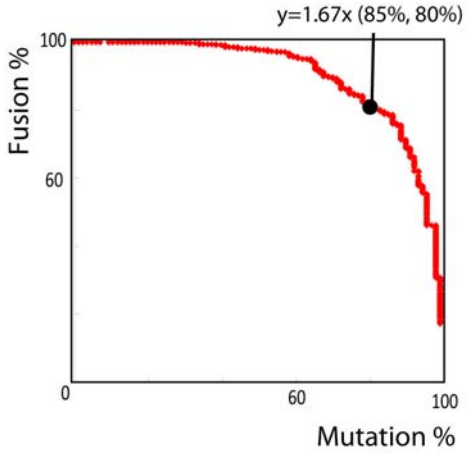


Supplementary Figure 1. The domain architectures of known gene fusions. Domains for known fusion genes were clustered according to their sequence similarity (columns), while the gene fusions were clustered according to their domain similarity (rows). The domains from the 5' partners are colored blue (left panel), and domains from 3' partners red (right panel). The proportion of each domain retained in the fusion protein is indicated by the saturation of the colors (ranging from 0~100%). The different domain patterns from the same gene fusion originated from different translocation breakpoints were labeled with different numbers, e.g. "BCR-ABL1_1". The sequences associated with each domain pattern were listed in Supplementary Table 1.

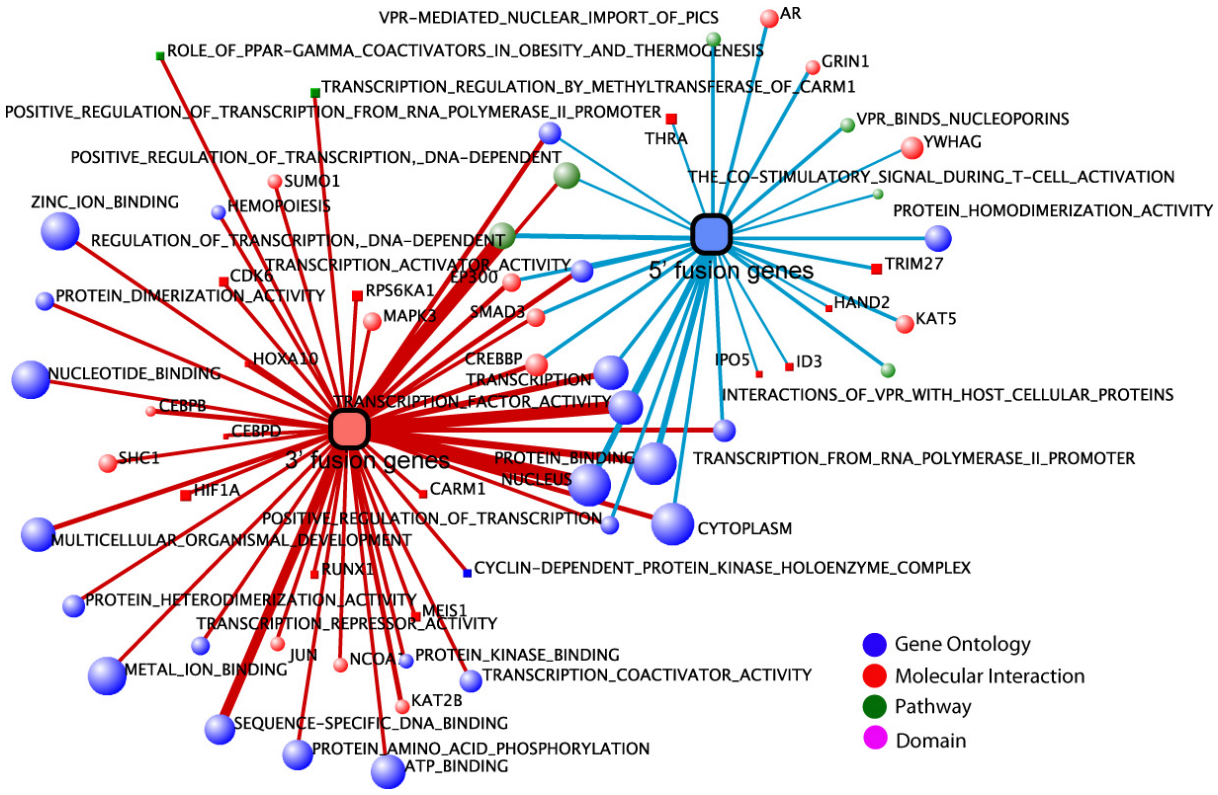


Supplementary Figure 2. Kolmogorov-Smirnov (K-S) analysis for the known cancer genes based on the rConSig-score with or without the pathways significantly overlapping with the molecular interactions ($p<0.01$).

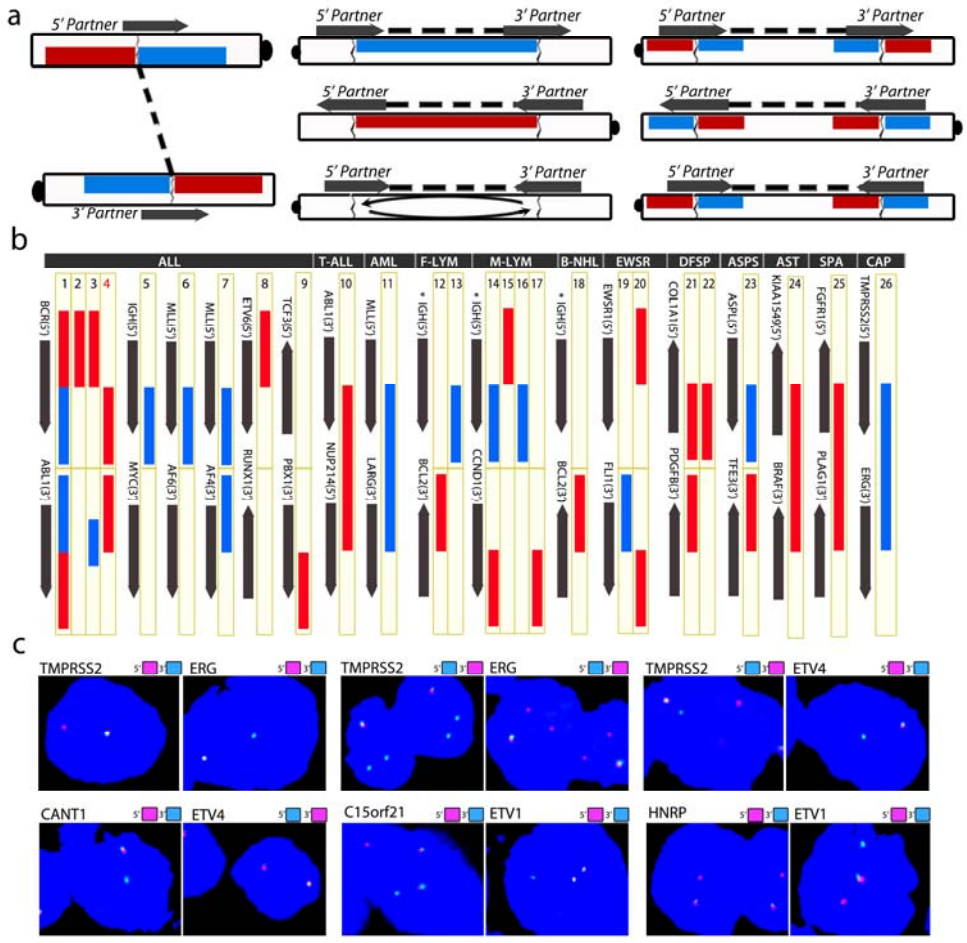
Nature Biotechnology: doi: 10.1038/nbt.1584



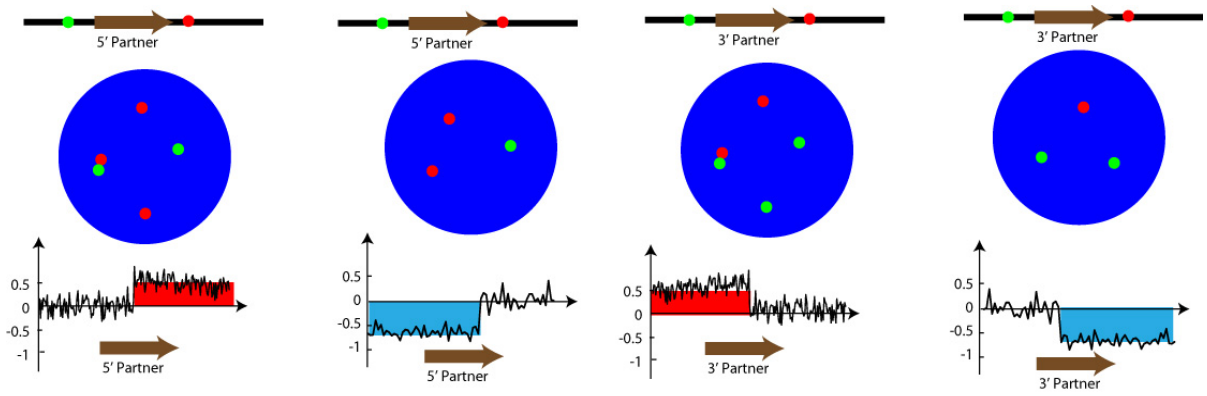
Supplementary Figure 3. The D-line y=*k*x (*k*=1.67) of the fusion-mutation ConSig plot was determined by setting optimal separation capacity. The D-line separates 85% of mutation genes from 80% of fusion genes. This plot was generated by calculating the separation capacity by increasing the k value from 0 to 20.



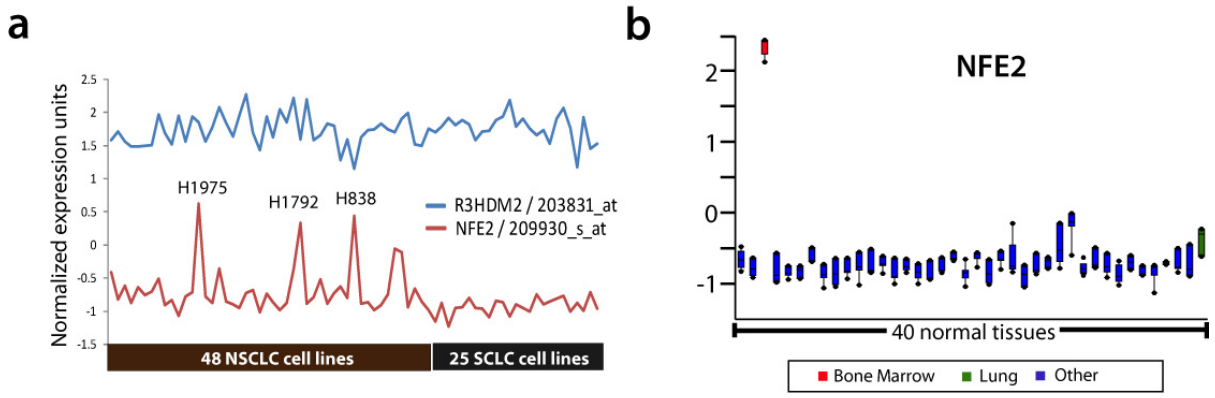
Supplementary Figure 4. The signature molecular concepts for 5' and 3' fusion genes. The enrichment analysis of 5' or 3' fusion genes against all molecular concepts was done by Fisher's exact tests. As a result, 3' fusion genes demonstrated much more enriched "signature concepts". The p value cutoff was $p< 10^{-5}$ for 5' fusion genes, and $p< 10^{-6}$ for 3' fusion genes.



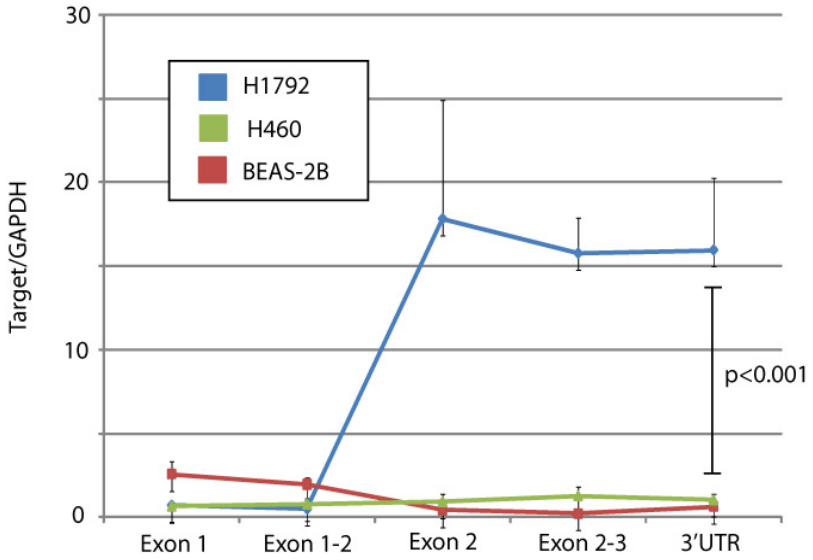
Supplementary Figure 5. The fusion breakpoint principle and the confirming evidence. (a) The fusion breakpoint principle. The left panel illustrates the application of the principle to inter-chromosomal translocations generating unbalanced gene fusions; the middle panel illustrates the inferred pattern of genomic imbalances for intra-chromosome gene fusions resulting from a single chromosome rearrangement with three different gene placements; the right panel shows the complex pattern of genomic imbalances for intra-chromosome translocations resulting from multiple rearrangements. Copy number increases (red bars) and deletions (blue bars) are represented as half the width of the chromosome, indicating the possibilities that the breakpoints are balanced. Genes are indicated by grey arrows (right arrow, "+" strand; left arrow, "-" strand). (b) Evidence for the principle from analysis of independent datasets and literature curation. The fusion gene partners are positioned at the left side of each breakpoint pattern (numbered in the top), with the DNA strands demonstrated by up ("-" strand) or down ("+" strand) arrows. A contradictory pattern is indicated with a red number. Copy number increases (red bars) and deletions (blue bars) are aligned inside the chromosomes (yellow bars) having these aberrations. The samples associated with each numbered breakpoint pattern can be found in Supplementary Table 7. Note: the fusion genes labeled with * are presented with low resolution probes in the corresponding datasets, and should be interpreted with caution. (c) Representative FISH results of the unbalanced ETS transcription factor fusions in 238 prostate cancer patients (University of Michigan Cohort). Split probes for each fusion partner were used to detect unbalanced translocations. The probe information is summarized in Supplementary Table 11.



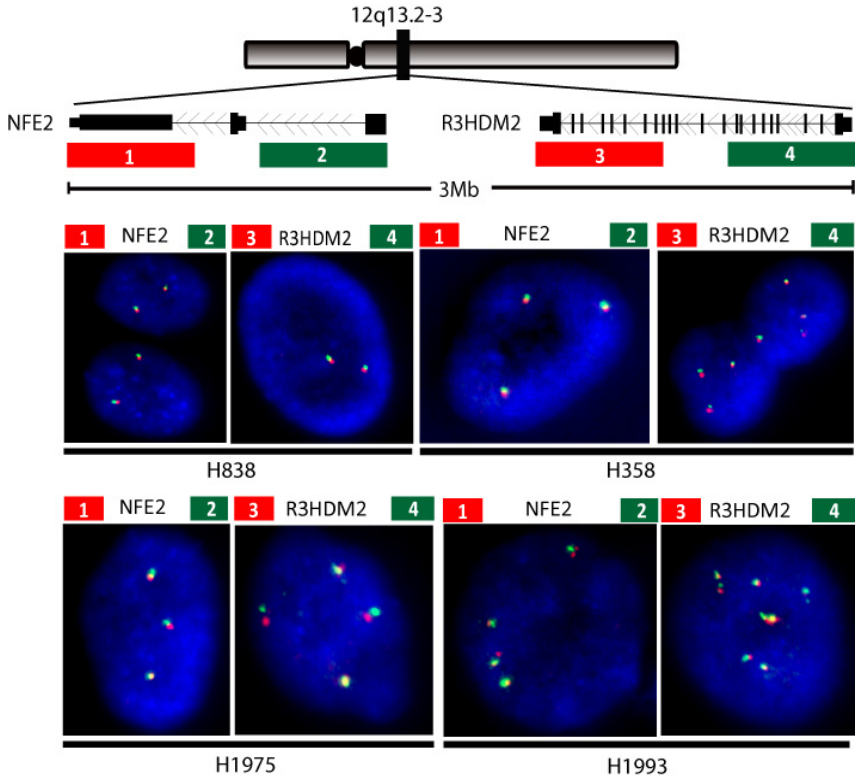
Supplementary Figure 6. Complex chromosome rearrangements generate contradictory cases to the breakpoint principle on array CGH, but not on FISH data. Upper panel shows the location of FISH probes (green and red dots) on the genomic loci of 5' or 3' partner genes (grey arrows). The middle panel shows the FISH appearance of complex chromosome rearrangements resulting in a balanced fusion (split signal) and an unbalanced translocation (from left to right: 3' duplication, 5' deletion, 5' duplication, 3' deletion). The lower panel shows the relative quantification of DNA copy number data generated by microarray CGH analysis from the genomic regions 1Mb apart from the fusion genes. The x axis indicates the physical position of the genomic aberrations. The fusion partners are indicated by grey arrows.



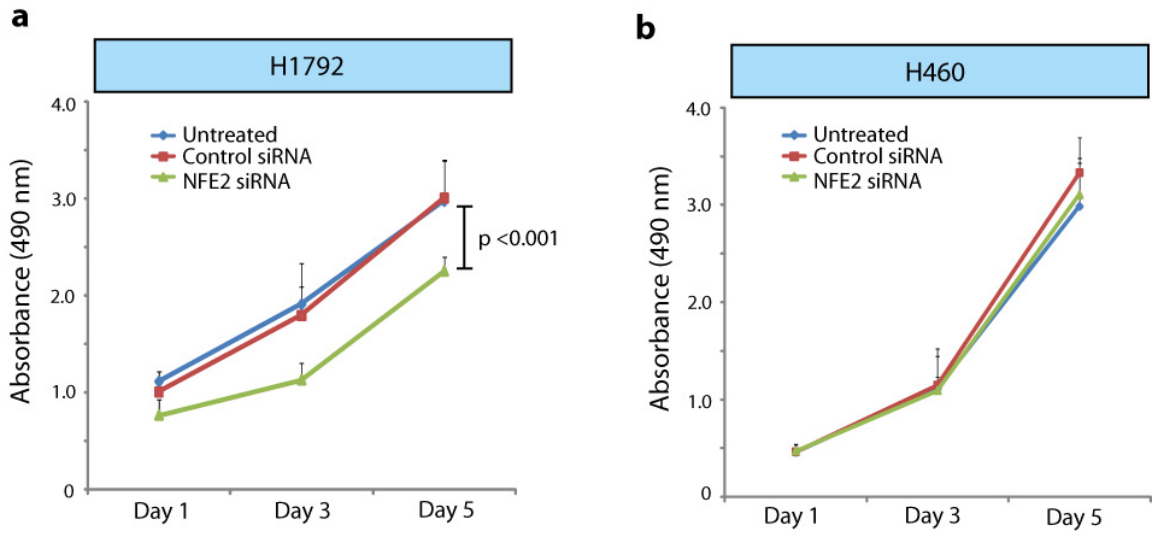
Supplementary Figure 7. Gene expression profile of *R3HDM2* and *NFE2*. (a) Microarray expression data of *R3HDM2* and *NFE2* on lung cancer cell lines. Figure shows the normalized expression units in all profiled lung cancer cell lines (Richard Wooster et al. gene expression study[19]). Visualization tools incorporated in Oncomine[20] were used to generate graphical displays. The cell lines that have marked over-expression of *NFE2* are indicated. (b) The expression of *NFE2* in 40 distinct normal tissues using Oncomine (Dataset: Both_Normal; probe: 209930_s_at). See Supplementary Table 12 for tissue classes. Expression (in normalized expression units) in 40 distinct normal tissues is shown (normal lung, green; bone marrow, red). Box and whisker plots show median +/- 90th%/10th%.



Supplementary Figure 8. Exon-walking RT-PCR reveals specific overexpression of *NFE2* coding exons. Exon-walking qRT-PCR with primer pairs corresponding to the indicated exons were used to evaluate exon-level expression changes in *NFE2*. The *R3HDM2-NFE2* fusion is predicted to join untranslated promoter sequences of *R3HDM2* with the coding exons of *NFE2* (exons 2-3, with exon 1 being an untranslated sequence in *NFE2*), resulting in the indicated overexpression of only *NFE2* coding exons. The H460 lung cancer cell line and BEAS-2B normal cells showed low endogenous *NFE2* levels. Primer pairs used are listed in **Supplementary Table 3**.



Supplementary Figure 9. FISH analysis of *NFE2* and *R3HDM2* loci by split probes strategy on selected lung cancer cell lines. Upper, the genomic organizations of *NFE2* and *R3HDM2* loci were shown in the schematic, with red and green bars indicating the location of BAC clones. Lower, interphase FISH analysis with *NFE2* and *R3HDM2* split probes showing normal co-localizing signals on H1975, H838, H358, and H1993 cell lines. H1975, H838 and H358 are three additional cell lines with *NFE2* over-expression; H1993 is a cell line that has relatively low *NFE2* expression. BAC clone probes: 1, RP11-621J12; 2, RP11-753H16; 3, RP11-799O6; 4, RP11-258J5.



Supplementary Figure 10. WST-1 assay shows inhibited cell proliferation after the *NFE2* knockdown on H1792 cell line, but not on the control cell line H460. (a). *NFE2* knockdown inhibits cell proliferation on H1792 cell line expressing the *R3HDM3-NFE2* fusion by a WST-1 assay (absorbance at 490nm was measured). (b) *NFE2* knockdown on H460 cell line with low level *NFE2* expression did not have significant effect on cell proliferation.

Supplementary Tables

Supplementary Table 1. Genbank sequences suggesting distinct domain patterns of known gene fusions. Each domain pattern was defined as the unique domain architecture generated by the fusion of two wild-type proteins; fusion of the same gene partners could generate different domain patterns due to the difference of fusion breakpoints.

Fusion	Pattern No.	Genbank Accessions
AFF1-MLL	1	AF487906;AF492831;
AFF1-MLL	2	AF177238;AF177239;
AKAP9-BRAF	1	AY803272;
ASPSCR1-TFE3	1	AY034077;
BCR-ABL1	1	AM491362 (e6a2);
BCR-ABL1	2	EU236680(e14a3);S72478(e14a3);
BCR-ABL1	3	EU216071(e14a2,Y5);M25946(e14a2,K562,CML);M30829(e14a2,K562); M30832(e14a2, EM2,CML);
BCR-ABL1	4	AF487522(e18a2,CML);
BCR-ABL1	5	AM491359(e13a3);AY043457(e13a3,CML);
BCR-ABL1	6	AJ131467(e13a2);EF158045(e13a2, SCA and CML);EU216066(e13a2, CML);
BCR-ABL1	7	AM491360(e14a3);
BCR-ABL1	8	AJ131466(e14a2);M13096(e14a2,K562);
BCR-ABL1	9	AM491363(e19a2);
BCR-ABL1	10	AM491361(e1a3);S72479(e1a3,ALL);
BCR-ABL1	11	AF113911(e1a2);M17541(e1a2,ALL);M19730(e1a2,ALL); X06418(e1a2,ALL);X07537(e1a2, ALL)
BIRC3-MALT1	1	AF123094;
BRD4-C15orf55	1	AY166680;
CBFB-MYH11	1	AF249897;AF249898;
CCDC6-RET	1	D90075;
CD74-ROS1	1	EU236945;
CDK6-MLL	1	AF492830;
CHCHD7-PLAG1	1	DQ478931;DQ478932;
CNBP-USP6	1	AY624556;
COL1A1-PDGFB	1	X98709;X98710;Y15913;Y15917;Y15918;Y15919;Y15921;
COL1A1-PDGFB	2	X98707;X98708;Y08643;Y15914;Y15915;Y15916;Y15920;Y16346;
DAZAP1-MEF2D	1	AY678451;
DDX10-NUP98	1	AB001342;
DDX10-NUP98	2	AB001343;
ELL-MLL	1	DQ437655;
EML4-ALK	1	AB274722;
EML4-ALK	2	AB275889;

ETV6-ABL1	1	Z35761;
ETV6-MN1	1	X85024;X85026;
ETV6-NTRK3	1	AF041811;AF125808;
EWSR1-DDIT3	1	X92120;
EWSR1-ERG	1	S72621;S72622;S72865;
EWSR1-ETV4	1	U35622;
EWSR1-FLI1	1	AF327066;S62665;S72620;
EWSR1-NR4A3	1	AF524261;S81242;
EWSR1-WT1	1	S74529;
EWSR1-WT1	2	S79672;
FGFR1-BCR	1	AJ298917;
FGFR1-PLAG1	1	EF525168;EF525169;
FUS-ATF1	1	AJ295163;
FUS-DDIT3	1	AJ301611;
FUS-DDIT3	2	AJ301612;S62138;S75762;S75763;X71427;
FUS-ERG	1	S77574;
GOLGA5-RET	1	X15786;
HMGA2-RAD51L1	1	AY138857;AY138858;AY138859;
HNRNPA2B1-ETV1	1	EF632110;
HOOK3-RET	1	DQ104207;
MAML2-CRTC1	1	AY186998;
MAPRE1-MLL	1	AY752859;
MEF2D-DAZAP1	1	AY675556;
MKL1-RBM15	1	AF364036;
MLL-AFF1	1	AF024541;AF177236;AF177237;DQ451148;
MLL-AFF1	2	AF031404;AF487905;AF492832;S67825;
MLL-AFF3	1	AF422798;
MLL-CBL	1	AY125965;
MLL-EPS15	1	AF331760;
MLL-EPS15	2	AY187922;
MLL-GAS7	1	AF231998;AF231999;
MLL-GAS7	2	AF231995;AF231996;AF231997;
MLL-GMPS	1	AF297746;AF297748;
MLL-GMPS	2	AF297747;AF297749;
MLL-KIAA0284	1	AM422012;
MLL-MAML2	1	AJ972402;DQ084494;DQ886023;
MLL-MAML2	2	DQ886024;
MLL-MAPRE1	1	AY752858;
MLL-MLLT1	1	DQ224341;
MLL-MLLT1	2	AF331759;AY040555;
MLL-MLLT1	3	DQ224342;
MLL-MLLT1	4	AY187921;

MLL-MLLT10	1	AF272375;AF272383;
MLL-MLLT10	2	AY187923;
MLL-MLLT10	3	AF272376;AF272384;AF272385;
MLL-MLLT3	1	EF406122;
MLL-MLLT3	2	S82034;
MLL-MLLT4	1	DQ387206;
MLL-MLLT6	1	S72604;
MLL-PICALM	1	AF477006;
MLL-SEPT5	1	AF061154;
MLL-SEPT6	1	AF450279;AF512943;AF512944;AF512945;AF512946;
MLLT10-PICALM	1	AF060927;AF060930;AF060931;
MLLT1-MLL	1	AF373587;
MN1-ETV6	1	X85025;X85027;
MYST3-ASXL2	1	AB084281;
MYST3-CREBBP	1	AJ251843;
MYST3-NCOA2	1	EF374064;
MYST4-CREBBP	1	AJ299261;
NIN-PDGFRB	1	AY764156;
NOL1-TCF3	1	EU155120;
NTRK3-ETV6	1	AF125809;
NUP98-DDX10	1	AB000267;
NUP98-DDX10	2	AB000268;
NUP98-HOXC13	1	AJ438986;
NUP98-HOXD13	1	AB038155;
NUP98-PRRX2	1	AY662674;
NUP98-RAP1GDS1	1	AF133331;AF133332;
PAX3-FOXO1	1	AF178854;BC008826;U02308;U02368;
PAX3-NCOA1	1	AY633656;
PAX5-ETV6	1	DQ841178;
PAX5-FOXP1	1	DQ845346;
PAX5-ZNF521	1	DQ845345;
PCM1-RET	1	AJ297349;
PICALM-MLLT10	1	EF051633;
PML-RARA	1	M73779;S50916;
PRKAR1A-RET	1	L03357;
RARA-PML	1	M82827;
RBM15-MKL1	1	AJ303089;
RPN1-EVI1	1	AF310158;
RUNX1-MDS1	1	S69002;
RUNX1-RUNX1T1	1	AX813476;AX813478;D13979;D14822;D14823;S78158;S78159;
RUNX1-SH3D19	1	EU093086;EU093087;
SLC34A2-ROS1	1	EU236946;EU236947;

SLC45A3-ETV1	1	EF632109;
SLC45A3-ETV5	1	EU314932;
SS18-SSX1	1	S79325;
SS18-SSX2	1	X79200;
SS18-SSX4	1	AF114234;
TAF15-NR4A3	1	AF162670;AJ243810;AJ245932;
TCF12-NR4A3	1	AF289510;
TCF3-PBX1	1	AY311345;M31522;
TFG-ALK	1	AF125093;AF143407;AF390893;
TFG-NR4A3	1	AY532911;
TFG-NTRK1	1	X85960;
TMPRSS2-ERG	1	DQ204773;DQ831522;EU090248;
TMPRSS2-ETV1	1	DQ204770;
TMPRSS2-ETV5	1	EU314929;EU314930;EU314931;

Supplementary Table 2. Top hub genes shared by 3'fusion partner families as revealed by significant overlapping statistics. "x, k, N" correspond to the variables demonstrated in the algorithm of hypergeometric statistics (Figure 1a)

5'Partner	3'partners(x)	Hub Genes	3'partners binding the hubs (k)	Total interacting genes of the hub gene (N)	P value
NUP98	20	*MEIS1*	6	21	3.77E-16
IGL@	11	*CDK6*	4	20	1.52E-11
MYST3	3	*CARM1*	3	11	1.57E-11
IGH@	36	*RUNX1*	5	25	2.36E-11
IGL@	11	*RUNX1*	4	25	3.96E-11
IGL@	11	*DMTF1*	3	4	6.26E-11
BCR	4	*PIK3R1*	4	126	9.54E-11
TRD@	8	*GATA3*	3	7	1.86E-10
MYST3	3	*HIF1A*	3	29	3.47E-10
TRB@	15	*LMO1*	3	6	8.63E-10
EWSR1	13	*ERG*	3	7	9.49E-10
IGL@	11	*CDKN1A*	4	55	1.06E-09
MYST3	3	*NCOA2*	3	44	1.26E-09
IGH@	36	*CTNNB1*	6	122	1.30E-09
IGL@	11	*AKAP8*	3	9	1.31E-09
TRB@	15	*GATA3*	3	7	1.51E-09

Supplementary Table 3. PCR primers used in this study.

Primer	Accession Number	Refseq (ucsc)	Type	Sequence (5'->3')
R3HDM2-NFE2 fusion (exon 2 to exon 2)	NM_014925-NM_006163	uc001snt-uc001sfq	Forward	ACTCATGGAGGCTGAGCATT
R3HDM2-NFE2 fusion (exon 2 to exon 2)	NM_014925-NM_006163	uc001snt-uc001sfq	Reverse	AGCTCGGTGATGGACATGAT
NFE2 (exon 1)	NM_006163	uc001sfq	Forward	AGCAGGGTGACCCCTGATGTTGCCC
NFE2 (exon 1)	NM_006163	uc001sfq	Reverse	ACTCCCCCAAACTGTTTTCCTGGCT
NFE2 (exon 1 - exon 2)	NM_006163	uc001sfq	Forward	AGCAGGGTGACCCCTGATGTTGCCC
NFE2 (exon 1 - exon 2)	NM_006163	uc001sfq	Reverse	TGGTCCAGGTTCCCGGAAAGCCCA
NFE2 (exon 2)	NM_006163	uc001sfq	Forward	TGGCCCAGTAGGATGTCCCCGTGT
NFE2 (exon 2)	NM_006163	uc001sfq	Reverse	GTGGACAGCTGTATCACCCTGTTCCT
NFE2 (exon 2 - exon 3)	NM_006163	uc001sfq	Forward	TCCCCAGCAGAGCAGGAACAGGGTGA
NFE2 (exon 2 - exon 3)	NM_006163	uc001sfq	Reverse	AAGGTATGGAGCTGGGGCTTGGGGCT
NFE2 (3'UTR)	NM_006163	uc001sfq	Forward	CTGAATCTCTTGAGCTGGAGG
NFE2 (3'UTR)	NM_006163	uc001sfq	Reverse	GCTGGCAAGGTATAGTTGGAGT
GAPDH	NM_002046	-	Forward	TGCACCACCAACTGCTTAGC
GAPDH	NM_002046	-	Reverse	GGCATGGACTGTGGTCATGAG

Supplementary Table 4. The summary of gene fusions reported in the leukemia array SNP dataset (GSE9113)

5' Partner	5' Partner Cytoband	3' Partner	3' Partner Cytoband	Total Fusion Samples (n)	Unbalanced Fusions (n)	Unbalanced/ Total %
BCR	22q11.23	*ABL1*	9q34.12	43 ALL 23 CML	9 ALL 5 CML	21.2
ETV6	12p13.2	*RUNX1*	21q22.12	48	17	35.4
MLL	11q23.3	*Multiple*		22	5	22.7
TCF3	19p13.3	*PBX1*	1q23.3	17	16	94.1
PAX5	9p13.2	*ETV6*	12p13.2	2	2	100
PAX5	9p13.2	*FOXP1*	3p14.1	1	1	100
PAX5	9p13.2	*ZNF521*	18q11.2	1	1	100

Supplementary Table 5. Analysis of the genomic imbalances associated with each gene fusion identified from the leukemia array SNP dataset (GSE9113). Note: “N”, no change; “amp”, amplification; “del”, deletion; “T”, telomere; “C”, centromere; “->” denotes the other end of the segmental deletion or amplification not generating the fusion. The case contradicting the fusion breakpoint principle was marked with bold and italic.

GEO accession	SAMPLE_ID	FUSION	Copy number abberation at 5' gene locus	Copy number abberation at 3' gene locus
GSM235572	#1	*BCR-ABL*	N	N
GSM235734	#10	*BCR-ABL*	N	N
GSM235735	#11	*BCR-ABL*	5'amp->T;3'del->Rgr	5'del->PPP2R4;3'amp->T
GSM235736	#12	*BCR-ABL*	N	N
GSM235738	#13	*BCR-ABL*	N	N
GSM235740	#14	*BCR-ABL*	5'amp->T	3'amp->T
GSM235743	#15	*BCR-ABL*	N	N
GSM235747	#16	*BCR-ABL*	3'del->LOC649264	N
GSM235751	#17	*BCR-ABL*	N	N
GSM235753	#18	*BCR-ABL*	N	N
GSM235801	#19	*BCR-ABL*	N	N
GSM235617	#2	*BCR-ABL*	5'amp->T	3'amp->T
GSM235809	#20	*BCR-ABL*	N	N
GSM235865	#21	*BCR-ABL*	N	N
GSM235810	#22	*BCR-ABL*	N	N
GSM235713	#23	*BCR-ABL*	N	N
GSM235714	#24	*BCR-ABL*	N	N
GSM235715	#25	*BCR-ABL*	N	N
GSM235716	#26	*BCR-ABL*	N	N
GSM235717	#27	*BCR-ABL*	N	N
GSM235718	#28	*BCR-ABL*	N	N
GSM235719	#29	*BCR-ABL*	N	N
GSM235645	#3	*BCR-ABL*	N	N
GSM235720	#30	*BCR-ABL*	N	N
GSM235721	#31	*BCR-ABL*	N	N
GSM235722	#32	*BCR-ABL*	N	N
GSM235723	#33	*BCR-ABL*	N	N
GSM235724	#34	*BCR-ABL*	5'amp->MAPK1	3'amp->Chr9:133236046
GSM235725	#35	*BCR-ABL*	N	N
GSM235726	#36	*BCR-ABL*	N	N
GSM235727	#37	*BCR-ABL*	N	N
GSM235728	#38	*BCR-ABL*	N	N
GSM235729	#39	*BCR-ABL*	N	N
GSM235664	#4	*BCR-ABL*	N	N
GSM235730	#40	*BCR-ABL*	5'amp->T	3'amp->T

GSM235731	#41	*BCR-ABL*	3'del->UPB1	5'del->ZER1
GSM235732	#42	*BCR-ABL*	N	N
GSM235733	#43	*BCR-ABL*	N	N
GSM235680	#5	*BCR-ABL*	N	5'del->5'region
GSM235693	#6	*BCR-ABL*	3'del->T	N
GSM235702	#7	*BCR-ABL*	N	N
GSM235767	#8	*BCR-ABL*	5'amp->T	3'amp->T
GSM235812	#9	*BCR-ABL*	N	N
GSM236531	#10-AP	*CML(BCR-ABL1)*	N	N
GSM236532	#10-CP	*CML(BCR-ABL1)*	N	N
GSM236534	#11-AP	*CML(BCR-ABL1)*	N	N
GSM236533	#11-CP	*CML(BCR-ABL1)*	N	N
GSM236536	#12-AP	*CML(BCR-ABL1)*	N	N
GSM236535	#12-CP	*CML(BCR-ABL1)*	N	N
GSM236537	#12-CP2	*CML(BCR-ABL1)*	N	N
GSM236538	#13-CP	*CML(BCR-ABL1)*	N	N
GSM236539	#13-CP2	*CML(BCR-ABL1)*	N	N
GSM236540	#14-BC	*CML(BCR-ABL1)*	N	N
GSM236541	#14-Rem	*CML(BCR-ABL1)*	N	N
GSM236544	#15-BC	*CML(BCR-ABL1)*	N	N
GSM236542	#15-CP	*CML(BCR-ABL1)*	N	N
GSM236543	#15-CP2	*CML(BCR-ABL1)*	N	N
GSM236547	#16-BC	*CML(BCR-ABL1)*	N	N
GSM236548	#16-BC-GL	*CML(BCR-ABL1)*	N	N
GSM236545	#16-CP	*CML(BCR-ABL1)*	N	N
GSM236546	#16-CP2	*CML(BCR-ABL1)*	N	N
GSM236550	#17-AP	*CML(BCR-ABL1)*	5'amp->T	3'amp->T
GSM236549	#17-CP	*CML(BCR-ABL1)*	N	N
GSM236551	#18-BC	*CML(BCR-ABL1)*	N	N
GSM236553	#19-BC	*CML(BCR-ABL1)*	5'amp->T	3'amp->T
GSM236554	#19-BC-GL	*CML(BCR-ABL1)*	N	N
GSM236552	#19-CP	*CML(BCR-ABL1)*	N	N
GSM236511	#1-BC	*CML(BCR-ABL1)*	N	N
GSM236510	#1-CP	*CML(BCR-ABL1)*	N	N
GSM236556	#20-AP	*CML(BCR-ABL1)*	N	N
GSM236557	#20-BC	*CML(BCR-ABL1)*	N	N
GSM236555	#20-CP	*CML(BCR-ABL1)*	N	N
GSM236558	#21-CP	*CML(BCR-ABL1)*	N	N
GSM236559	#21-CP2	*CML(BCR-ABL1)*	N	N
GSM236561	#22-BC	*CML(BCR-ABL1)*	5'amp->T;3'del->IGLL1	5'del->CDK9;3'amp->T
GSM236562	#22-BC-GL	*CML(BCR-ABL1)*	N	N
GSM236560	#22-CP	*CML(BCR-ABL1)*	3'del->IGLL1	5'del->CDK9

GSM236564	#23-BC	*CML(BCR-ABL1)*	N	N
GSM236565	#23-BC-GL	*CML(BCR-ABL1)*	N	N
GSM236563	#23-CP	*CML(BCR-ABL1)*	N	N
GSM236512	#2-CP	*CML(BCR-ABL1)*	N	N
GSM236513	#2-CP2	*CML(BCR-ABL1)*	N	N
GSM236514	#3-AP	*CML(BCR-ABL1)*	N	N
GSM236515	#3-BC	*CML(BCR-ABL1)*	N	3'amp->T
GSM236518	#4-BC	*CML(BCR-ABL1)*	N	N
GSM236516	#4-CP	*CML(BCR-ABL1)*	N	N
GSM236517	#4-Rem	*CML(BCR-ABL1)*	N	N
GSM236521	#5-BC	*CML(BCR-ABL1)*	N	N
GSM236520	#5-BC-GL	*CML(BCR-ABL1)*	N	N
GSM236519	#5-Rem	*CML(BCR-ABL1)*	N	N
GSM236524	#6-BC	*CML(BCR-ABL1)*	5'amp->T	3'amp?->T
GSM236523	#6-BC-GL	*CML(BCR-ABL1)*	N	N
GSM236522	#6-CP	*CML(BCR-ABL1)*	N	N
GSM236526	#7-BC	*CML(BCR-ABL1)*	N	N
GSM236525	#7-CP	*CML(BCR-ABL1)*	N	N
GSM236528	#8-AP	*CML(BCR-ABL1)*	N	N
GSM236527	#8-CP	*CML(BCR-ABL1)*	N	N
GSM236529	#9-BC	*CML(BCR-ABL1)*	N	N
GSM236530	#9-BC-GL	*CML(BCR-ABL1)*	N	N
GSM235579	#1	*E2A-PBX1*	5'del->T	3'amp->T
GSM235776	#10	*E2A-PBX1*	5'del->T	3'amp->T
GSM235789	#11	*E2A-PBX1*	5'del?->T	3'amp->T
GSM235804	#12	*E2A-PBX1*	5'del->T	3'amp->T
GSM235813	#13	*E2A-PBX1*	5'del->T	3'amp->T
GSM235820	#14	*E2A-PBX1*	5'del->T	3'amp->T
GSM235854	#15	*E2A-PBX1*	5'del->T	3'amp->T
GSM235859	#16	*E2A-PBX1*	5'del->T	3'amp->T
GSM235861	#17	*E2A-PBX1*	5'del->T	3'amp->T
GSM235602	#2	*E2A-PBX1*	5'del->T	3'amp->T
GSM235620	#3	*E2A-PBX1*	5'del->T	3'amp->T
GSM235625	#4	*E2A-PBX1*	5'del?->T	3'amp->T
GSM235632	#5	*E2A-PBX1*	5'del?->T	3'amp->T
GSM235641	#6	*E2A-PBX1*	5'del->T	3'amp->T
GSM235650	#7	*E2A-PBX1*	5'del->T	3'amp->T
GSM235668	#8	*E2A-PBX1*	N	N
GSM235701	#9	*E2A-PBX1*	5'del?->T	3'amp->T
GSM235670	#10	*PAX5-ETV6*	PAX5:3'del->T	ETV6:5'del->T
GSM235611	#3	*PAX5-ZNF521*	PAX5:3'del->T	ZNF521:5'del->T
GSM235631	#1	*MLL-*	N	N

GSM235846	#10	*MLL-*	N	N
GSM235866	#11	*MLL-*	N	N
GSM235563	#12	*MLL-*	N	N
GSM235578	#13	*MLL-*	N	N
GSM235627	#15	*MLL-*	N	N
GSM235673	#16	*MLL-*	N	MLLT3 5'del->C
GSM235712	#17	*MLL-*	N	N
GSM235742	#18	*MLL-*	N	N
GSM235746	#19	*MLL-*	N	N
GSM235633	#2	*MLL-*	3'del->HYOU1	AFF1 int del
GSM235818	#20	*MLL-*	3'del->BCL9L	N
GSM235847	#21	*MLL-*	N	N
GSM235855	#22	*MLL-*	N	N
GSM235869	#23	*MLL-*	N	N
GSM235652	#3	*MLL-*	3'del->CBL	N
GSM235662	#4	*MLL-*	N	N
GSM235768	#5	*MLL-*	N	N
GSM235780	#6	*MLL-*	N	N
GSM235851	#7	*MLL-*	N	N
GSM235834	#8	*MLL-*	N	AFF1 int del
GSM235837	#9	*MLL-*	N	N
GSM235828	#14	*PAX5-FOXP1*	PAX5:3'del->PTPRD	FOXP1:N
GSM235561	#1	*PAX5-ETV6*	PAX5:3'del->T	ETV6:5'del->T
GSM235566	#1	*TEL-AML1*	3'del->3'region	5'del->5'region
GSM235601	#10	*TEL-AML1*	N	N
GSM235603	#11	*TEL-AML1*	int del	N
GSM235605	#12	*TEL-AML1*	N	N
GSM235616	#13	*TEL-AML1*	N	N
GSM235634	#14	*TEL-AML1*	N	N
GSM235642	#15	*TEL-AML1*	N	N
GSM235653	#16	*TEL-AML1*	N	N
GSM235654	#17	*TEL-AML1*	N	N
GSM235658	#18	*TEL-AML1*	N	N
GSM235661	#19	*TEL-AML1*	N	N
GSM235567	#2	*TEL-AML1*	N	N
GSM235672	#20	*TEL-AML1*	3'del->chr12:028840592	3'amp->JAM2
GSM235674	#21	*TEL-AML1*	int del	N
GSM235684	#22	*TEL-AML1*	3'del->3'region	N
GSM235685	#23	*TEL-AML1*	N	N
GSM235688	#24	*TEL-AML1*	int del	N
GSM235696	#25	*TEL-AML1*	N	N
GSM235699	#26	*TEL-AML1*	int del	N

GSM235704	#27	*TEL-AML1*	N	N
GSM235705	#28	*TEL-AML1*	N	N
GSM235708	#29	*TEL-AML1*	N	N
GSM235570	#3	*TEL-AML1*	N	5'del->TCC3
GSM235754	#30	*TEL-AML1*	N	5'del->5'region
GSM235760	#31	*TEL-AML1*	N	N
GSM235761	#32	*TEL-AML1*	N	N
GSM235762	#33	*TEL-AML1*	N	N
GSM235764	#34	*TEL-AML1*	3'del->chr12:017224159	N
GSM235770	#35	*TEL-AML1*	N	N
GSM235772	***#36***	***TEL-AML1***	***5'del->T***	***5'amp->T***
GSM235774	#37	*TEL-AML1*	N	N
GSM235779	#38	*TEL-AML1*	3'del->C	N
GSM235788	#39	*TEL-AML1*	3'del->CDKN1B	N
GSM235571	#4	*TEL-AML1*	N	N
GSM235794	#40	*TEL-AML1*	N	N
GSM235800	#41	*TEL-AML1*	3'del->LRP6	N
GSM235816	#42	*TEL-AML1*	N	N
GSM235821	#43	*TEL-AML1*	N	N
GSM235827	#44	*TEL-AML1*	N	N
GSM235835	#45	*TEL-AML1*	3'del->BICD1	3'amp->T
GSM235857	#46	*TEL-AML1*	N	N
GSM235862	#47	*TEL-AML1*	N	N
GSM235845	#48	*TEL-AML1*	N	int del
GSM235576	#5	*TEL-AML1*	N	N
GSM235581	#6	*TEL-AML1*	N	N
GSM235582	#7	*TEL-AML1*	int del	N
GSM235585	#8	*TEL-AML1*	N	N
GSM235597	#9	*TEL-AML1*	N	N

Supplementary Table 6. The summary of the genomic imbalances associated with gene fusions identified from the 36 leukemia cell lines. Note: “N”, no change; “amp”, amplification; “del”, deletion; “T”, telomere; “C”, centromere; “->” denotes the other end of the segmental amplification or deletion not generating the fusion. The case contradicting the principle was marked with bold and italic.

GEO accession	Sample ID	Fusion	Copy number abberation at 5' gene locus	Copy number abberation at 3' gene locus
GSM236815	#BV173	*BCR-ABL1*	5'amp->IGL@	5'del->C
GSM236820	#K-562	*BCR-ABL1*	5'amp->T	3'amp->NUP214
GSM236836	#OP1	*BCR-ABL1*	N	N
GSM236840	#SD1	*BCR-ABL1*	N	5'del->5'region
GSM236842	#SUPB-15	*BCR-ABL1*	N	N
GSM236843	#THP-1	*BCR-ABL1*	N	N
GSM236844	#TOM-1	*BCR-ABL1*	N	N
GSM236824	#ME-1	*CBFB-MYH11*	N	N
GSM236846	#UOCB1	*E2A-HLF*	N	N
GSM236814	#AT1	*ETV6-RUNX1*	N	N
GSM236823	#KG-1	*FGFR1OP2-FGFR1*	5'amp->chr12:025768792; 3'del->AMN1	3'amp->chr8:036399366
GSM236812	#380	*IGH-BCL2*	int del	N
GSM236839	#RS4_11	*MLL-AF4*	N	N
GSM236832	#MV4-11	*MLL-AF4(AFF1)*	N	N
GSM236827	#ML-2	*MLL-AF6(MLLT4)*	3'del->T	5'del->ESR1
GSM236835	#NOMO-1	*MLL-AF9*	N	5'del->chr9:032018982
GSM236843	***#THP-1***	***MLL-AF9***	***3'amp->T***	***3'amp->T***
GSM236831	#Mono-mac-6	*MLL-AF9(MLLT3)*	N	N
GSM236812	#380	*MYC-IGH*	N	int del
GSM236845	#U-937	*PICALM-AF10(MLLT10)*	N	N
GSM236834	#NB4	*PML-RARA*	N	N
GSM236821	#Kasumi-1	*RUNX1-RUNX1T1*	N	N
GSM236841	#SKNO-1	*RUNX1-RUNX1T1*	N	N
GSM236816	#CCRF-CEM	*SIL(STIL)-SCL(TAL1)*	N	N
GSM236813	#697	*TCF3-PBX1*	N	3'amp->T
GSM236822	#Kasumi-2	*TCF3-PBX1*	3'del->T	3'amp->T
GSM236826	#MHH-CALL-3	*TCF3-PBX1*	3'del->T	3'amp->T
GSM236838	#Reh	*TEL-AML1*	whole gene del	5'amp->T

Supplementary Table 7. Analysis and curation results of the public array CGH/array SNP/tiling CGH data with gene fusions associated from publications. ASPS, Alveolar soft part sarcoma; DFSP, Dermatofibrosarcoma Protruberans; AML, Acute Myelogenous Leukemia; ALL, Acute Lymphoblastic Lymphoma; EWS, Ewings' sarcoma; NHL, non-hodgekin lymphoma; AST, Brain Astrocytoma; LUG, Lung Carcinoma; CaP, Prostate Adenocarcinoma; SPA, Salivary Pleomorphic Adenoma.

GEO accession/ Pubmed ID	Cancer	GSM accession	Sample ID	Fusion	DNA Breakpoint Pattern ID	Copy number abberation at 5' gene locus	Copy number abberation at 3' gene locus
GSE9611	ALL	GSM243107	#1	*BCR-ABL1*	--	N	N
GSE9611	ALL	GSM243108	#2	*BCR-ABL1*	#1	5'amp->T; 3'del->LOC51233	5'del->CCBL1; 3'amp->OBP2B
GSE9611	ALL	GSM243109	#3	*BCR-ABL1*	#2	5'amp->T	N
GSE9611	ALL	GSM243110	#4	*BCR-ABL1*	--	N	N
GSE9611	ALL	GSM243111	#5	*BCR-ABL1*	#2	5'amp->T	N
GSE9611	ALL	GSM243112	#6	*BCR-ABL1*	#3	5'amp->T	int del
GSE9611	ALL	GSM243113	#7	*BCR-ABL1*	--	N	N
GSE9611	***ALL***	***GSM243114***	***#8***	***BCR-ABL1***	***#4***	***3'amp->T***	***5'amp->C***
GSE9611	ALL	GSM243115	#9	*BCR-ABL1*	--	N	N
GSE9611	ALL	GSM243116	#10	*BCR-ABL1*	--	N	N
GSE9611	ALL	GSM243119	#13	*IGH-MYC*	#5	3'del->T	N
GSE9611	ALL	GSM243120	#14	*IGH-MYC*	--	N	N
GSE7255	ALL	GSM174868	9348 (#1)	*MLL-AF6*	#6	3'del->CBL	N
GSE7255	ALL	GSM174860	9225(#2)	*MLL-AF4*	#7	3'del->DDX6	5'del->LOC442777
GSE7255	ALL	GSM174830	9256(#3)	*ETV6-RUNX1*	--	N	N
GSE7255	ALL	GSM174846	9418(#4)	*ETV6-RUNX1*	#8	5'amp->T	N
GSE7255	ALL	GSM174851	9367(#5)	*ETV6-RUNX1*	--	N	N
GSE7255	ALL	GSM174852	9393(#6)	*E2A-PBX1*	#9	N	3'amp->T
GSE8918	NHL	GSM226057	FL# 1	*lgH-BCL2(90%)*	--	N	N
GSE8918	NHL	GSM226058	FL# 2	*lgH-BCL2(90%)*	--	N	N
GSE8918	NHL	GSM226059	FL# 3	*lgH-BCL2(90%)*	--	N	N
GSE8918	NHL	GSM226060	FL# 4	*lgH-BCL2(90%)*	--	N	N
GSE8918	NHL	GSM226061	FL# 5	*lgH-BCL2(90%)*	--	N	N
GSE8918	NHL	GSM226062	FL# 6	*lgH-BCL2(90%)*	--	N	N
GSE8918	NHL	GSM226063	FL# 7	*lgH-BCL2(90%)*	--	N	N
GSE8918	NHL	GSM226064	FL# 8	*lgH-BCL2(90%)*	--	N	N
GSE8918	NHL	GSM226065	FL# 9	*lgH-BCL2(90%)*	--	N	N
GSE8918	NHL	GSM226066	FL# 10	*lgH-BCL2(90%)*	--	N	N
GSE8918	NHL	GSM226067	FL# 11	*lgH-BCL2(90%)*	--	N	N
GSE8918	NHL	GSM226068	FL# 12	*lgH-BCL2(90%)*	#12	N	3'amp->T
GSE8918	NHL	GSM226069	FL# 13	*lgH-BCL2(90%)*	#13	3'del?->T	N
GSE8918	NHL	GSM226070	FL# 14	*lgH-BCL2(90%)*	--	N	N
GSE8918	NHL	GSM226071	FL# 15	*lgH-BCL2(90%)*	--	N	N

GSE8918	NHL	GSM226088	MCL# 31	*lgH-CCND1(95%)*	#14	3'del?->T	3'amp->CENTD2
GSE8918	NHL	GSM226089	MCL# 32	*lgH-CCND1(95%)*	--	N	N
GSE8918	NHL	GSM226090	MCL# 33	*lgH-CCND1(95%)*	#15	5'amp?->JAG2	N
GSE8918	NHL	GSM226091	MCL# 34	*lgH-CCND1(95%)*	--	N	N
GSE8918	NHL	GSM226092	MCL# 35	*lgH-CCND1(95%)*	--	N	N
GSE8918	NHL	GSM226093	MCL# 36	*lgH-CCND1(95%)*	#16	3'del?->T	N
GSE8918	NHL	GSM226094	MCL# 37	*lgH-CCND1(95%)*	--	N	N
GSE8918	NHL	GSM226095	MCL# 38	*lgH-CCND1(95%)*	#17	N	3'amp->UVRAG
GSE8918	NHL	GSM226096	MCL# 39	*lgH-CCND1(95%)*	--	N	N
GSE8918	NHL	GSM226097	MCL# 40	*lgH-CCND1(95%)*	--	N	N
GSE8918	NHL	GSM226098	MCL# 41	*lgH-CCND1(95%)*	--	N	N
GSE8918	NHL	GSM226099	MCL# 42	*lgH-CCND1(95%)*	--	N	N
GSE8918	NHL	GSM226100	MCL# 43	*lgH-CCND1(95%)*	--	N	N
GSE8918	NHL	GSM226101	MCL# 44	*lgH-CCND1(95%)*	--	N	N
GSE8918	NHL	GSM226111	LPL# 54	*IgH-PAX5(50%)*	--	N	N
GSE8918	NHL	GSM226112	LPL# 55	*IgH-PAX5(50%)*	--	N	N
GSE8918	NHL	GSM226113	LPL# 56	*IgH-PAX5(50%)*	--	N	N
GSE8918	NHL	GSM226114	LPL# 57	*IgH-PAX5(50%)*	--	N	N
GSE8918	NHL	GSM226115	LPL# 58	*IgH-PAX5(50%)*	--	N	N
GSE8918	NHL	GSM226116	LPL# 59	*IgH-PAX5(50%)*	--	N	N
GSE8918	NHL	GSM226117	LPL# 60	*IgH-PAX5(50%)*	--	N	N
GSE8918	NHL	GSM226118	LPL# 61	*IgH-PAX5(50%)*	--	N	N
GSE8918	NHL	GSM226119	LPL# 62	*IgH-PAX5(50%)*	--	N	N
GSE8918	NHL	GSM226120	LPL# 63	*IgH-PAX5(50%)*	--	N	N
GSE8918	NHL	GSM226136	MALT# 79	*BIRC3-MALT1(30%)**	--	N	N
GSE8918	NHL	GSM226137	MALT# 80	*BIRC3-MALT1(30%)**	--	N	N
GSE8918	NHL	GSM226138	MALT# 81	*BIRC3-MALT1(30%)**	--	N	N
GSE8918	NHL	GSM226139	MALT# 82	*BIRC3-MALT1(30%)**	--	N	N
GSE8918	NHL	GSM226140	MALT# 83	*BIRC3-MALT1(30%)**	--	N	N
GSE8918	NHL	GSM226141	MALT# 84	*BIRC3-MALT1(30%)**	--	N	N
GSE8918	NHL	GSM226142	MALT# 85	*BIRC3-MALT1(30%)**	--	N	N
GSE8918	NHL	GSM226143	MALT# 86	*BIRC3-MALT1(30%)**	--	N	N
GSE8918	NHL	GSM226144	MALT# 87	*BIRC3-MALT1(30%)**	--	N	N
GSE8398	EWS	GSM207892	# 1	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207893	# 2	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207894	# 3	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207895	# 4	*EWSR1-FLI1*	#19	N	5'del->TMEM135
GSE8398	EWS	GSM207896	# 5	*EWSR1-FLI1*	#20	5'amp->T	3'amp->T
GSE8398	EWS	GSM207897	# 6	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207898	# 7	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207899	# 8	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207900	# 9	*EWSR1-FLI1*	--	N	N

GSE8398	EWS	GSM207901	# 10	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207902	# 12	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207903	# 13	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207904	# 14	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207905	# 15	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207906	# 16	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207907	# 17	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207908	# 18	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207909	# 19	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207910	# 20	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207911	# 21	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207912	# 22	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207913	# 23	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207914	# 24	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207915	# 25	*EWSR1-FLI1*	--	N	N
GSE8398	EWS	GSM207916	# 26	*EWSR1-FLI1*	--	N	N
16193090	B-NHL	NA	581/90(#1)	*IGH-BCL2*	#18	?	3'amp
16193090	B-NHL	NA	364/86(#2)	*IGH-BCL2*	#18	?	3'amp
16193090	B-NHL	NA	436/91(#3)	*IGH-BCL2*	#18	?	3'amp
16193090	B-NHL	NA	176/88(#4)	*IGH-BCL2*	#18	?	3'amp
16193090	B-NHL	NA	472/90(#5)	*IGH-BCL2*	#18	?	3'amp
16193090	B-NHL	NA	287/88(#6)	*IGH-BCL2*	#18	?	3'amp
16193090	B-NHL	NA	21/87(#7)	*IGH-BCL2*	#18	?	3'amp
16193090	B-NHL	NA	130/92(#8)	*IGH-BCL2*	--	?	N
16193090	B-NHL	NA	377/83(#9)	*IGH-BCL2*	--	?	N
16193090	B-NHL	NA	311/89(#10)	*IGH-BCL2*	--	?	N
16193090	B-NHL	NA	41/88(#11)	*IGH-BCL2*	--	?	N
16193090	B-NHL	NA	64/89(#12)	*IGH-BCL2*	--	?	N
16193090	B-NHL	NA	34/90(#13)	*IGH-BCL2*	--	?	N
16193090	B-NHL	NA	381/88(#14)	*IGH-BCL2*	--	?	N
16193090	B-NHL	NA	140/90(#15)	*IGH-BCL2*	--	?	N
16193090	B-NHL	NA	345/87(#16)	*IGH-BCL2*	--	?	N
16193090	B-NHL	NA	140/90(#17)	*IGH-BCL2*	--	?	N
16193090	B-NHL	NA	345/87(#18)	*IGH-BCL2*	--	?	N
15361874	T-ALL	NA	#1	*NUP214-ABL1*	#10	5'amp->ABL1	3'amp->NUP214
15361874	T-ALL	NA	#2	*NUP214-ABL1*	#10	5'amp->ABL1	3'amp->NUP214
15361874	T-ALL	NA	#3	*NUP214-ABL1*	#10	5'amp->ABL1	3'amp->NUP214
15361874	T-ALL	NA	#4	*NUP214-ABL1*	#10	5'amp->ABL1	3'amp->NUP214
15361874	T-ALL	NA	#5	*NUP214-ABL1*	#10	5'amp->ABL1	3'amp->NUP214
15361874	T-ALL	NA	#6	*NUP214-ABL1*	#10	5'amp->ABL1	3'amp->NUP214
10681437	AML	NA	#1	*MLL-LARG*	#11	3'del->LARG	5'del->MLL
GSE3930	DFSP	GSM89915	STT154(#1)	*COL1A1-PDGFB*	#21	5'amp	3'amp

GSE3930	DFSP	GSM89903	STT154(#2)	*COL1A1-PDGFB*	#21	5'amp	3'amp
GSE3930	DFSP	GSM89916	STT491(#3)	*COL1A1-PDGFB*	#21	5'amp	3'amp
GSE3930	DFSP	GSM89911	STT491(#4)	*COL1A1-PDGFB*	#21	5'amp	3'amp
GSE3930	DFSP	GSM89919	STT1984(#5)	*COL1A1-PDGFB*	#21	5'amp	3'amp
GSE3930	DFSP	GSM89904	STT1984(#6)	*COL1A1-PDGFB*	#21	5'amp	3'amp
GSE3930	DFSP	GSM89931	STT1971(#7)	*COL1A1-PDGFB*	#22	5'amp	N
17124411	DFSP	NA	7 cases	*COL1A1-PDGFB*	#21	5'amp in 5 cases	3'amp in 3 cases
11244503	ASPS	NA	12 cases	*ASPSCR1-TFE3*	#23	3'del in 9 cases	3'amp in 9 cases
18974108	AST	NA	29 cases	*KIAA1549-BRAF*	#24	5’amp in 29 cases	3’amp in 29 cases
18059337	SPA	NA	11 cases	*FGFR1-PLAG1*	#25	5'amp in 10 cases	3'amp in 10 cases
17654723 16951139	CaP	NA	106 cases	*TMPRSS2-ERG*	#26	3’del in 54 cases	5’amp in 54 cases

* In MALT, besides BIRC3-MALT1 reported in 30% cases, there were also IgH-MALT1 reported in 15-20% cases, IgH-FOXP1 in 10% cases and IgH-BCL10 in 5% cases.

Supplementary Table 8. Meta-analysis of unbalanced gene fusions in multiple human cancers to test the fusion breakpoint principle. Abbreviations: ALL, acute lymphoblastic leukemia; CML, chronic mylogenous leukemia; AML, acute mylogenous leukemia; NHL, non-Hodgkin's lymphoma; EWS, Ewing's sarcoma; DFSP, dermatofibrosarcoma protruberans; ASPS, Alveolar soft part sarcoma; AST, astrocytoma; SPA, Salivary Pleomorphic Adenoma; CaP, prostate cancer.

Cancer type	First Author	Citation	Platform	Total Samples with fusions	Unbalanced	Follow the principle
ALL/CML	Mulighan CG	Nature 2007;446:758	Affymetrix 500K aSNP	185	68	66
ALL	Paulsson K	PNAS 2008;105:6708	Affymetrix 500K aSNP	13	6	5
ALL	Kuiper RP	Leukemia 2007;21:1258	Affymetrix 100K aSNP	6	4	4
T-ALL	Graux C	Nat Genet 2004;36:1084	CGH	6	6	6
AML	Kourlas PJ	PNAS 2000;97:2145	CGH	1	1	1
NHL	Ferreira BI	Haematologica 2008;93:670	Agilent 44B aCGH	48	8	8
B-NHL	Galteland E	Leukemia 2005;19:2313	BAC aCGH	18	7	7
EWS	Ferreira BI	Oncogene 2008;27:2084	Agilent 44B aCGH	25	2	2
DFSP	Linn SC	Am J Pathol 2003;163:2383	Standford aCGH	7	7	7
DFSP	Kaur S	Cytogenet Genome Res. 2006	Agilent 13k aCGH	7	5	5
ASPS	Ladanyi M	Oncogene 2001;20:48	FISH	12	9	9
AST	Jones DT	Cancer Res 2008:68:8673	MHP 1Mb aCGH	29	29	29
SPA	Persson F	Oncogene 2008;27:3072	Agilent 44B aCGH	11	10	10
CaP	Liu W	Gene Chrom Canc 2007;46:972	Affymetrix 500K aSNP	41	16	16
CaP	Pertner S	Cancer Res 2006;66:8337	FISH	65	38	38
CaP	Wang XS	This study	FISH	104	60	60
			TOTAL:	578	276	273(98.9%)

Supplementary Table 9. Curation of the experimental data suggesting the genomic abberations for all intra-chromosome gene fusions from the Mitelman database. "No.": the number of publications reporting the gene fusions. Note: Amp, segmental amplification, Del: interstitial deletion, inv: inversion. APL, Acute promyelocytic leukemia; AUL, Acute undifferentiated leukemia; CEL, Chronic eosinophilic leukemia; MYE, Myeloproliferative disease; SAR, Sarcoma; LPB, Lipoblastoma; LYM, Lymphoma; F-LYM, Follicular Lymphoma; T-LYM, T-cell Lymphoma; B-LYM, B-cell Lymphoma, M-LYM, Mantle Cell Lymphoma; NLD, Nonneoplastic lymphatic disorder; LPL, Lymphoplasmacytic Lymphoma; THY, Thyroid Adenocarcinoma.

Gene Fusion	Cancer Type	Genomic Distance between fusion partners (kb)	Predicted genomic imbalance	Total No. of reports	Pubmed ID of informative reports	Experimental Methods	Curation Results *
MLL/ARHGEF17	AML	45055	amp	1	--	--	no information
TPM3/TPR	THY	32118	amp	1	--	--	no information
RPN1/EVI1	AML	40433	amp	2	--	--	no information
NUP214/ ABL1	ALL	237	amp	1	15361874	Tiling CGH	interstitial amplicon
PRKAR1A/RARA	APL	28252	amp	1	17712046	FISH	no unbalance info
TCEA1/PLAG1	SAL	2138	amp	2	16736500	FISH	no unbalance info
MLL/DCPS	AML	7778	del	1	--	--	no information

TFRC/BCL6	B-LYM	8315	del	1	--	--	no information
HAS2/PLAG1	LPB	65408	del	2	10987300	FISH	bac not locatable
TMPRSS2/ERG	CaP	2803	del	10	16951139	aCGH	interstitial deletion
HNRPA2B1/ETV1	CaP	12203	del	1	17671502	FISH	del 3' HNRPA2B1;del 5' ETV1
MLL/CBL	AML	681	del	1	12696071	FISH	del 3'MLL
MLL/ARHGEF12	AML	1812	del	2	10681437	inference	del 3'MLL; del 5'LARG
FIP1L1/PDGFRA	CEL	770	del	4	12660384, 14973504	FISH	del internal BAC at CHIC2 locus
SET/NUP214	AML	2492	del	2	17296573	aCGH	interstitial deletion
STIL/TAL1	ALL	20	del	2	8459224	citation	interstitial deletion
GOPC/ROS1	GBM	134	del	1	12661006	sequencing	interstitial deletion
MLL/TIRAP	AML	7757	del	1	15626757	RT-PCR	no reciprocal fusion
RET/NCOA4	THY	8297	del	4	--	--	no information
LPP/BCL6	B-LYM	467	inv	1	--	--	no information
MLL/BCL9L	ALL	371	inv	1	--	--	no information
MLL/MAML2	AML	22096	inv	1	--	--	no information
BCL11B/TRD@	ALL	76700	inv	1	15668700	FISH	balanced inversion
TRA@/TCL1A	NLD	73155	inv	1	7662982	FISH	dup of 5'& 3' TCL1A
RET/CCDC6	THY	18274	inv	5	1542652	Southern blot	reciprocal fusion
EML4/ALK	LUG	12252	inv	3	18083107	FISH	del 5'ALK
EWSR1/PATZ1	SAR	2025	inv	1	10949935	RT-PCR	no reciprocal fusion
MLL/PICALM	AML;ALL	32355	inv	2	12461747	RT-PCR	no reciprocal fusion
CHCHD7/PLAG1	SAL	0	inv	1	16736500	FISH	no unbalanced info
AKAP9/BRAF	THY	48503	inv	1	15630448	RT-PCR	reciprocal fusion
TPM3/NTRK1	THY	2621	inv	2	7590742	RT-PCR	reciprocal fusion
AFF1/ELF2	ALL	51917	inv	1	17410185	RT-PCR	three way balanced
DSCAML1/MLL	ALL	639	inv	1	17410185	RT-PCR	three way balanced
FXYD6/MLL	ALL	560	inv	1	17410185	RT-PCR	three way balanced

* Amplifications/deletions fit to the prediction from the inferred principle as well as translocations without genomic imbalances are considered as following the inferred principle.

Supplementary Table 10. The summary of FISH findings for unbalanced ETS gene fusions in 171 prostate cancer cases (UM cohort)

Case (n)	3' ETS Gene	FISH finding	5' fusion partner	FISH finding
44	*ERG*	5'deletion	*TMPRSS2*	3' deletion
4	*ERG*	split	*TMPRSS2*	3' deletion
3	*ERG*	5' deletion	*TMPRSS2*	split
5	*ERG*	5'deletion; 3'duplication	*TMPRSS2*	3' deletion; 5' duplication
1	*ETV1*	split	*HNRPA2B1*	3' deletion

1	*ETV1*	5’deletion	*C15ORF21*	split
1	*ETV4*	5’deletion	*TMPRSS2*	split
1	*ETV4*	5’deletion	*CANT1*	3’ deletion

Supplementary Table 11. The split-apart probes used for fluorescence in situ hybridization detecting ETS gene rearrangements in prostate cancer.

Gene	Chromosome band	5' region	3' region
ERG	21q22.2	RP11-95I21	RP11-476D17
ETV1	7p21.2	RP11-703A4	RP11-124L22
ETV4	17q21.31	RP11-436J4	RP11-100E5
ETV5	3q27.2	RP11-379C23	RP11-1144N13
TMPRSS2	21q22.3	RP11-35C4	RP11-120C17
SLC45A3	1q32.1	RP11-1089F13	RP11-1143H2
C15ORF21	15q21.1	RP11-474E1	RP11-626F7
HERV-K_22q11.23	22q11.23	RP11-947A12	RP11-61P17
HNRPA2B1	7p15.2	RP11-379M24	RP11-11F13
FLJ35294	17p13.1	RP11-1099M24	RP11-55C13
CANT1	17q25.3	RP11-52K16	RP11-46K10
KLK2	19q13.33	CTC-771P3	RP11-26P14
DDX5	17q24.1	RP11-81D7	RP11-315N9

Supplementary Table 12. Normal tissues from the "Both_Normal" dataset analyzed using Oncomine database (www.oncomine.org) [20]. The gene expression data of NFE2 across the 40 normal tissue types indicated in this table are displayed in Supplementary Fig. 9.

No.	Tissue Type	Sample Number (n)
1	Adipose	10
2	Adrenal Gland Cortex	4
3	Bone Marrow	5
4	Bronchus	3
5	Cervix	4
6	Colon Cecum	3
7	Coronary Artery	3
8	Dorsal Root Ganglia	8
9	Endometrium	4
10	Esophagus	4
11	Heart Atrium	4
12	Heart Ventricle	3
13	Kidney Cortex	4
14	Kidney Medulla	4
15	Liver	4
16	Lymph Nodes	4
17	Mammary Gland	3
18	Myometrium	5
19	Nipple Cross-Section	4
20	Nodose Nucleus	8
21	Oral Mucosa	4
22	Ovary	4
23	Pharyngeal Mucosa	4
24	Pituitary Gland	8
25	Prostate Gland	3
26	Salivary Gland	4
27	Saphenous Vein	3
28	Skeletal Muscle	5
29	Spleen	4
30	Stomach	11
31	Testes	3
32	Thyroid Gland	4
33	Tongue	8
34	Tonsil	3
35	Trachea	3
36	Trigeminal Ganglia	8
37	Urethra	3
38	Vagina	4
39	Vulva	4
40	Lung	3

References

1. Hu, Z. et al. VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Res* **35**, W625-632 (2007).
2. Ono, Y., Fukuhara, N. & Yoshie, O. TAL1 and LIM-only proteins synergistically induce retinaldehyde dehydrogenase 2 expression in T-cell acute lymphoblastic leukemia by acting as cofactors for GATA3. *Mol Cell Biol* **18**, 6939-6950 (1998).
3. Thorsteinsdottir, U., Kroon, E., Jerome, L., Blasi, F. & Sauvageau, G. Defining roles for HOX and MEIS1 genes in induction of acute myeloid leukemia. *Mol Cell Biol* **21**, 224-234 (2001).
4. Sampietro, J. et al. Crystal structure of a beta-catenin/BCL9/Tcf4 complex. *Mol Cell* **24**, 293-300 (2006).
5. Senyuk, V. et al. The leukemia-associated transcription repressor AML1/MDS1/EVI1 requires CtBP to induce abnormal growth and differentiation of murine hematopoietic cells. *Oncogene* **21**, 3232-3240 (2002).
6. Janknecht, R. Regulation of the ER81 transcription factor and its coactivators by mitogen- and stress-activated protein kinase 1 (MSK1). *Oncogene* **22**, 746-755 (2003).
7. Maki, K., Yamagata, T. & Mitani, K. Role of the RUNX1-EVI1 fusion gene in leukemogenesis. *Cancer Sci* **99**, 1878-1883 (2008).
8. Hart, S.M. & Foroni, L. Core binding factor genes and human leukemia. *Haematologica* **87**, 1307-1323 (2002).
9. Wishart, D.S. et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **34**, D668-672 (2006).
10. Kharas, M.G. et al. Ablation of PI3K blocks BCR-ABL leukemogenesis in mice, and a dual PI3K/mTOR inhibitor prevents expansion of human BCR-ABL+ leukemia cells. *J Clin Invest* **118**, 3038-3050 (2008).
11. Garcia-Manero, G. et al. Phase 1 study of the histone deacetylase inhibitor vorinostat (suberoylanilide hydroxamic acid [SAHA]) in patients with advanced leukemias and myelodysplastic syndromes. *Blood* **111**, 1060-1066 (2008).
12. Dunleavy, K. Phase I/II Study of Flavopiridol in Patients With Refractory or Recurrent Mantle Cell Lymphoma or Diffuse Large B-Cell Lymphoma. *NCI - Center for Cancer Research: NCI-07-C-0081* (2007-2009).
13. Odero, M.D., Zeleznik-Le, N.J., Chinwalla, V. & Rowley, J.D. Cytogenetic and molecular analysis of the acute monocytic leukemia cell line THP-1 with an MLL-AF9 translocation. *Genes Chromosomes Cancer* **29**, 333-338 (2000).
14. Persson, F. et al. High-resolution array CGH analysis of salivary gland tumors reveals fusion and amplification of the FGFR1 and PLAG1 genes in ring chromosomes. *Oncogene* (2007).
15. Perner, S. et al. TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer. *Cancer Res* **66**, 8337-8341 (2006).
16. Liu, W. et al. Multiple genomic alterations on 21q22 predict various TMPRSS2/ERG fusion transcripts in human prostate cancers. *Genes Chromosomes Cancer* **46**, 972-980 (2007).
17. Jones, D.T. et al. Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas. *Cancer Res* **68**, 8673-8677 (2008).
18. Mitelman F, J.B.a.M.F.E. Mitelman Database of Chromosome Aberrations in Cancer. (2007).
19. Richard, W. https://array.nci.nih.gov/caarray/project/woost-00041.
20. Rhodes, D.R. et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**, 166-180 (2007).

Nature Biotechnology: doi: 10.1038/nbt.1584