D994–D1004 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 8 November 2018*
doi: 10.1093/nar/gky1067

FusionGDB: fusion gene annotation DataBase

Pora Kim[1] and Xiaobo Zhou[1,2,3,*]

[1]Center for Computational Systems Medicine, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA, [2]McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA and [3]School of Dentistry, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Received August 09, 2018; Revised October 05, 2018; Editorial Decision October 17, 2018; Accepted November 01, 2018

ABSTRACT

Gene fusion is one of the hallmarks of cancer genome via chromosomal rearrangement initiated by DNA double-strand breakage. To date, many fusion genes (FGs) have been established as important biomarkers and therapeutic targets in multiple cancer types. To better understand the function of FGs in cancer types and to promote the discovery of clinically relevant FGs, we built FusionGDB (Fusion Gene annotation DataBase) available at https://ccsm.uth.edu/FusionGDB. We collected 48 117 FGs across pan-cancer from three representative fusion gene resources: the improved database of chimeric transcripts and RNA-seq data (ChiTaRS 3.1), an integrative resource for cancer-associated transcript fusions (TumorFusions), and The Cancer Genome Atlas (TCGA) fusions by Gao *et al*. For these ∼48K FGs, we performed functional annotations including gene assessment across pan-cancer fusion genes, open reading frame (ORF) assignment, and retention search of 39 protein features based on gene structures of multiple isoforms with different breakpoints. We also provided the fusion transcript and amino acid sequences according to multiple breakpoints and transcript isoforms. Our analyses identified 331, 303 and 667 in-frame FGs with retaining kinase, DNA-binding, and epigenetic factor domains, respectively, as well as 976 FGs lost protein-protein interaction. FusionGDB provides six categories of annotations: FusionGeneSummary, FusionProtFeature, FusionGeneSequence, FusionGenePPI, RelatedDrug and RelatedDisease.

INTRODUCTION

Gene fusion is one of the hallmark of cancer genome through chromosomal rearrangements triggered by DNA double-strand breakage. Accordingly, many fusion genes (FGs) have been identified as important biomarkers and therapeutic targets in multiple cancer types. Identification and analysis of fusion genes (FGs) will provide important insights into the mechanisms of cancer development and design novel therapeutic strategies (1). With the exponential growth of cancer genomic and other biomedical data, several studies have integrated and searched fusion genes across multiple cancer types (e.g. pan-cancer studies). The improved database of chimeric transcripts and RNA-seq data (ChiTaRS 3.1) provided 20 131 human breakpoints from expressed sequence tags (EST) (2). Since the TCGA database was open to the public, four research groups have been working on predicting FGs and tried to identify driver FGs from this dataset. Stransky *et al*. predicted fusion genes involving kinases across 20 cancer types from ∼7000 TCGA samples (3). In this study, the authors filtered out the FGs that were also detected in normal samples of TCGA and the Genotype-Tissue Expression (GTEx) project. Their studies suggested that 3.0% of tumor samples contained likely oncogenic, recurrent kinase fusion genes. ChimerDB 3.0 has an enhanced coverage of fusion gene data through literature mining using machine learning method (4) and provided 30K FG pairs by analyzing RNA-seq data of 5300 TCGA samples across 28 cancer types. TumorFusions, an integrative resource for cancer-associated transcript fusions, predicted FGs from 9966 TCGA samples across 33 cancer types (5). They applied stringent criteria and resulted in ∼21K FGs. Most recently, Gao *et al.* selected FGs called from at least two callers and predicted in ∼ 26K FGs from 9624 TCGA samples of 33 cancer types (6).

Although above studies have provided a huge amount of reliable FGs, such resources did not present detailed functional annotation of individual FGs. In addition, the identification of driver FGs was solely based on the kinase FGs. So far, a systematic annotation of FGs in cancer regarding the retention of diverse functional domains and protein features, which are important in understanding cellular process and tumorigenesis, has not been available. In this study, we investigated the retention of 39 protein features of 43 895 FGs with ORF annotation and identified 331, 303, 840 and 667 in-frame FGs with retaining kinase domain, DNA-binding domain, oncogene domains, and epigenetic factor (epifactor) domains, respectively. Furthermore, we

[*]To whom correspondence should be addressed. Tel: +1 713 500 3923; Email: Xiaobo.Zhou@uth.tmc.edu

identified 896 and 118 in-frame FGs that did not retain their functional domains of tumor suppressor genes and DNA damage repair genes, respectively. In contrast, there were 6863 FGs with domain retention, but they lost their function due to the frame-shift ORF by chromosomal rearrangement. We also analyzed the retention information for protein-protein interaction (PPI) in fusion protein. Such analysis can provide fusion gene candidates who lose important interactions with cellular regulators. Through this analysis, we identified 718 and 976 FGs with PPI retention and without PPI retention, respectively. 761 FGs have no PPI functionalities due to ORF frameshifts. Since the identification and browsing of FGs for analysis and validation are based on the exact genomic breakpoints, obtaining accurate fusion transcript and fusion amino acid sequences are very important for further studies in both of the dry- and wet-lab, and are urgent needs for many cancer researchers. However, the exact fusion transcript or fusion amino acid sequences of all existing fusion genes considering multiple isoforms and breakpoints are not available. We have created these fusion sequences based on the genomic breakpoints at the Ensembl gene isoform structures (7).

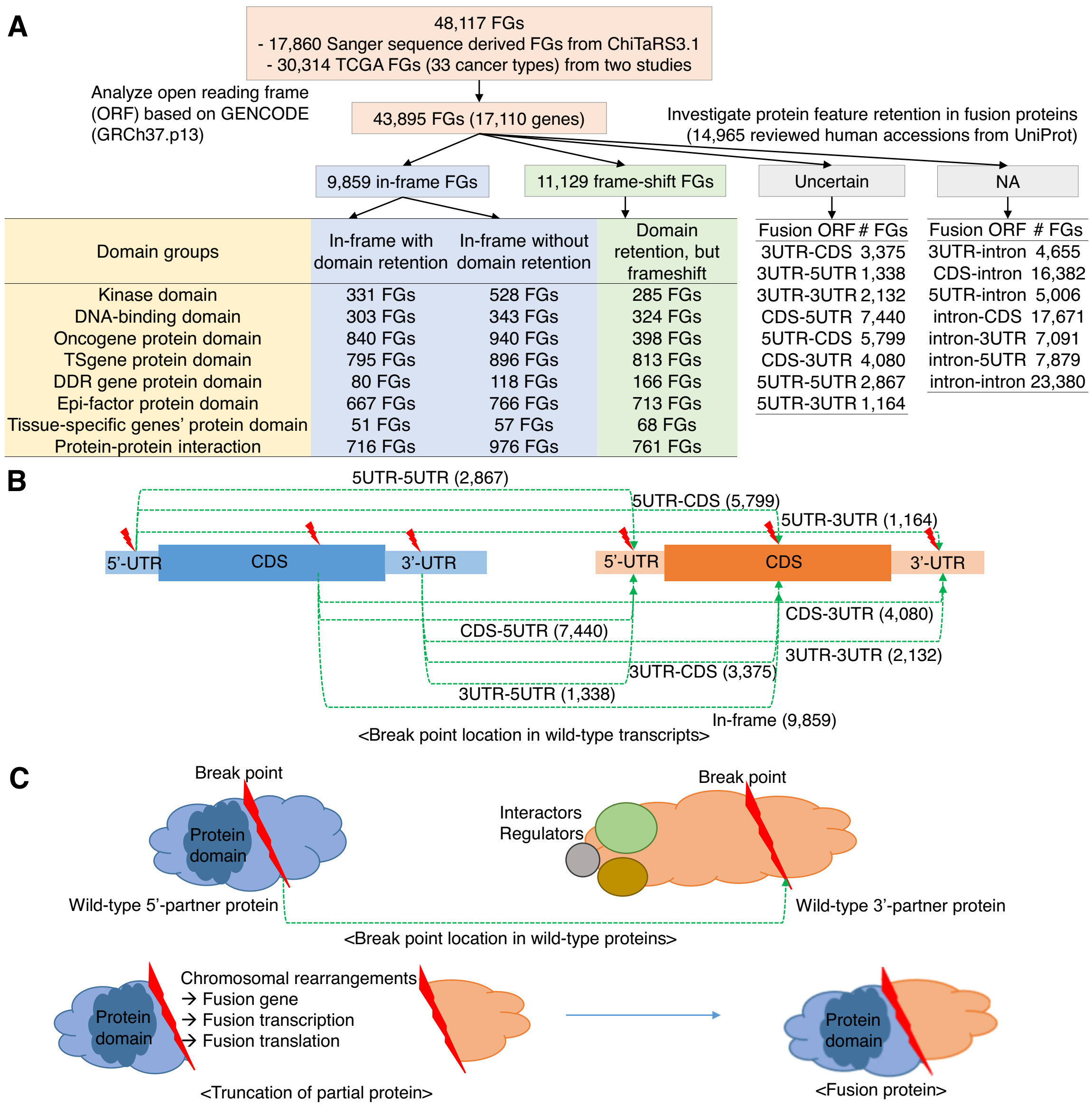
Here, we give a detailed introduction of the FusionGDB (Fusion Gene annotation DataBase), including the web interface and its applications. Our database includes features of all human FGs based on large cancer dataset analysis using systematic bioinformatics approaches, providing resources or references for functional annotation of fusion genes. FusionGDB will be a unique resource of cancer research for understanding the mechanisms of cancer development and identifying potential therapeutic targets against cancer.

DATABASE OVERVIEW

We first collected 48 117 FGs across pan-cancer from three representative fusion gene resources: the improved database of chimeric transcripts and RNA-seq data (ChiTaRS 3.1) (2), an integrative resource for cancer-associated transcript fusions (TumorFusions) (5) and The Cancer Genome Atlas (TCGA) fusions by Gao *et al.* (6) (Supplementary Table S1). For these ~48K FGs, we performed functional annotations including gene assessment across pan-cancer fusion genes, open reading frame (ORF) assignment, and protein domain retention searches based on multiple isoform gene structures and breakpoints, and finally provided the fusion transcripts and amino acid sequences for each breakpoint and gene isoforms (Figure 1 and Supplementary Table S2 and S3 (https://ccsm.uth.edu/FusionGDB/tables/TableS2.zip, https://ccsm.uth.edu/FusionGDB/tables/TableS3A.zip, and https://ccsm.uth.edu/FusionGDB/tables/TableS3B.zip)). For each fusion partner gene, the user can access multiple annotations such as gene summary, assessment scores of each gene in pan-cancer, biological process gene ontologies, functional description, and retention information of 39 protein features from UniProt (8) and protein–protein interaction (PPI), related drugs and diseases through six categories. Among ~44K FGs, which were checked ORFs, there were ~ 10K in-frame FGs and ~11K frameshift FGs. Of these, we identified 331, 303, 840 and 667 in-frame FGs with retained kinase, DNA-binding, oncogene and epigenetic effector domains. In addition, we identified 896 and 118 in-frame FGs that do not retain their functional domains of tumor suppressor genes and DNA damage repair genes, respectively. 6863 FGs with reserved functional domains have no functionalities due to frameshifts (Figure 1 and Supplementary Table S3A and S3B). We also investigated the retention of PPI in FGs and found that 976 FGs have no retained PPI regions and 761 FGs lose PPIs due to ORF frameshifts. For the 175 highly recurrent FGs that have expressed in more than five samples, we performed manual curation of PubMed articles (Supplementary Table S4). All of these information is included in the database and downloadable with unique and efficient data formats.

The main features of the FusionGDB annotations are summarized as follows. (i) The FusionGeneSummary category displays an overview of multiple annotations of fusion genes. Specifically, in this category, we added two genetic assessment scores, such as the Degree of Frequency (DoF) and the Major Active Isofusion Index (MAII), to provide the impact of each gene on pan-cancer gene fusions, which were created from previous studies (Supplementary Table S5A and B (9,10). Figure 2 shows the top-ranked impact genes for the 5′-genes (Head genes or Hgenes) and 3′-genes (Tail genes or Tgenes) of pan-cancer FGs through our scoring system. Furthermore, this category also shows the assignment of functional classes of each fusion gene to help understand the tumorigenic mechanisms. (ii) FusionProtFeature category provides the retention information of 39 protein features of fusion proteins based on their multiple isoforms of gene structures and multiple breakpoints. Through focusing on the protein domains or regions of interest among the 39 features, users can understand more about the overall function of specific fusion genes and make a hypothesis/plan for further research on tumorigenesis. (iii) In the FusionGeneSequence category, we present full-length fusion transcript and coding region (CDS) sequences, and amino acid sequences based on the multiple breakpoints and matched gene isoforms. (iv) FusionGenePPI category provides protein–protein interaction (PPI) information of fusion proteins. This category also present a link to the chimeric protein-protein interaction (ChiPPI) (11) for users to get fusion domain networks. It also shows the list of interactors with each partner protein in wild-type to infer the original interactions. Specifically, this category provides retention annotations of PPIs for better understand the possible loss or gain of interactions with important cellular regulatory factors due to structural disruptions of gene fusion. (v) RelatedDrug and RelatedDisease categories are designed for providing FG related drugs and diseases. By using these two categories, we identified 1341 approved drugs for 1302 genes involved in 8423 FGs and 6169 genes associated with 4679 different types of diseases from DrugBank (12) and DisGeNet (13), respectively. The details of the data and analysis processes are described in the following section.

Figure 1 and Table 1 summarize the overview of FG annotation and the overall statistics of protein feature retention status of ~48K FGs. All entries and annotation data can be viewed and downloaded on the FusionGDB


A
48,117 FGs
- 17,860 Sanger sequence derived FGs from ChiTaRS3.1
- 30,314 TCGA FGs (33 cancer types) from two studies
Analyze open reading frame (ORF) based on GENCODE (GRCh37.p13)
43,895 FGs (17,110 genes)
Investigate protein feature retention in fusion proteins (14,965 reviewed human accessions from UniProt)
9,859 in-frame FGs
11,129 frame-shift FGs
Uncertain
NA
B
5UTR-5UTR (2,867)
5UTR-CDS (5,799)
5UTR-3UTR (1,164)
5'-UTR
CDS
3'-UTR
5'-UTR
CDS
3'-UTR
CDS-3UTR (4,080)
CDS-5UTR (7,440)
3UTR-3UTR (2,132)
3UTR-CDS (3,375)
3UTR-5UTR (1,338)
In-frame (9,859)
<Break point location in wild-type transcripts>
C
Break point
Protein domain
Wild-type 5'-partner protein
Break point
Interactors
Regulators
Wild-type 3'-partner protein
<Break point location in wild-type proteins>
Protein domain
Chromosomal rearrangements
→ Fusion gene
→ Fusion transcription
→ Fusion translation
<Truncation of partial protein>
Protein domain
<Fusion protein>

Domain groups	In-frame with domain retention	In-frame without domain retention	Domain retention, but frameshift
Kinase domain	331 FGs	528 FGs	285 FGs
DNA-binding domain	303 FGs	343 FGs	324 FGs
Oncogene protein domain	840 FGs	940 FGs	398 FGs
TSgene protein domain	795 FGs	896 FGs	813 FGs
DDR gene protein domain	80 FGs	118 FGs	166 FGs
Epi-factor protein domain	667 FGs	766 FGs	713 FGs
Tissue-specific genes' protein domain	51 FGs	57 FGs	68 FGs
Protein-protein interaction	716 FGs	976 FGs	761 FGs

Fusion ORF	# FGs
3UTR-CDS	3,375
3UTR-5UTR	1,338
3UTR-3UTR	2,132
CDS-5UTR	7,440
5UTR-CDS	5,799
CDS-3UTR	4,080
5UTR-5UTR	2,867
5UTR-3UTR	1,164

Fusion ORF	# FGs
3UTR-intron	4,655
CDS-intron	16,382
5UTR-intron	5,006
intron-CDS	17,671
intron-3UTR	7,091
intron-5UTR	7,879
intron-intron	23,380

Figure 1. Overview of FusionGDB. (**A**) Work flow of functional annotation of FGs. (**B**) ORF types based on fusion transcript. (**C**) Schematic overview of retention of protein domain or interactors in fusion protein. Fusion protein in this figure retained protein domain of 5′-partner and lost protein-protein interaction (PPIs) with cellular regulators/interactors of 3′-partner.

website with a unique and efficient visualization interface (https://ccsm.uth.edu/FusionGDB).

DATA INTEGRATION AND ANNOTATIONS

Fusion gene information

We downloaded breakpoint information of 17 860, 15 854 and 23 944 FGs and their related information from ChiTaRS 3.1(http://chitars.md.biu.ac.il/, January 2017), TumorFusions (http://www.tumorfusions.org, November 2017), and TCGA fusions (Gao *et al.*, April 2018), respectively. Of these, 17 860 and 30 270 FGs were from Entrez Sanger sequences and TCGA samples. By integrating the above data, we obtained 48 117 unique FGs. For the genome coordinates information for fusion breakpoints from Gao et al. study, we lifted it over from the human reference genome GRCh38 to GRCh37 using Batch Coordinate Conversion (liftOver) utility from UCSC Genome Browser (14) to fit with the reference genome of the other two resources. The following FG information from these resources have been collected: sample ID or expressed sequence tag (EST) ID, the name of fusion partner gene and exon junction breakpoint. We followed the definition of FGs direction for the Hgene (Head gene or 5′-gene) and Tgene (Tail gene or 3′-gene) to these datasets.

Manual curation of PubMed articles

For the 175 highly recurrent FGs, which expressed in more than five samples or cells, a literature query of PubMed for individual FGs was conducted in July 2018. Using *BCR-ABL1* as an example, it is ‘((*BCR* [Title/Abstract]) AND

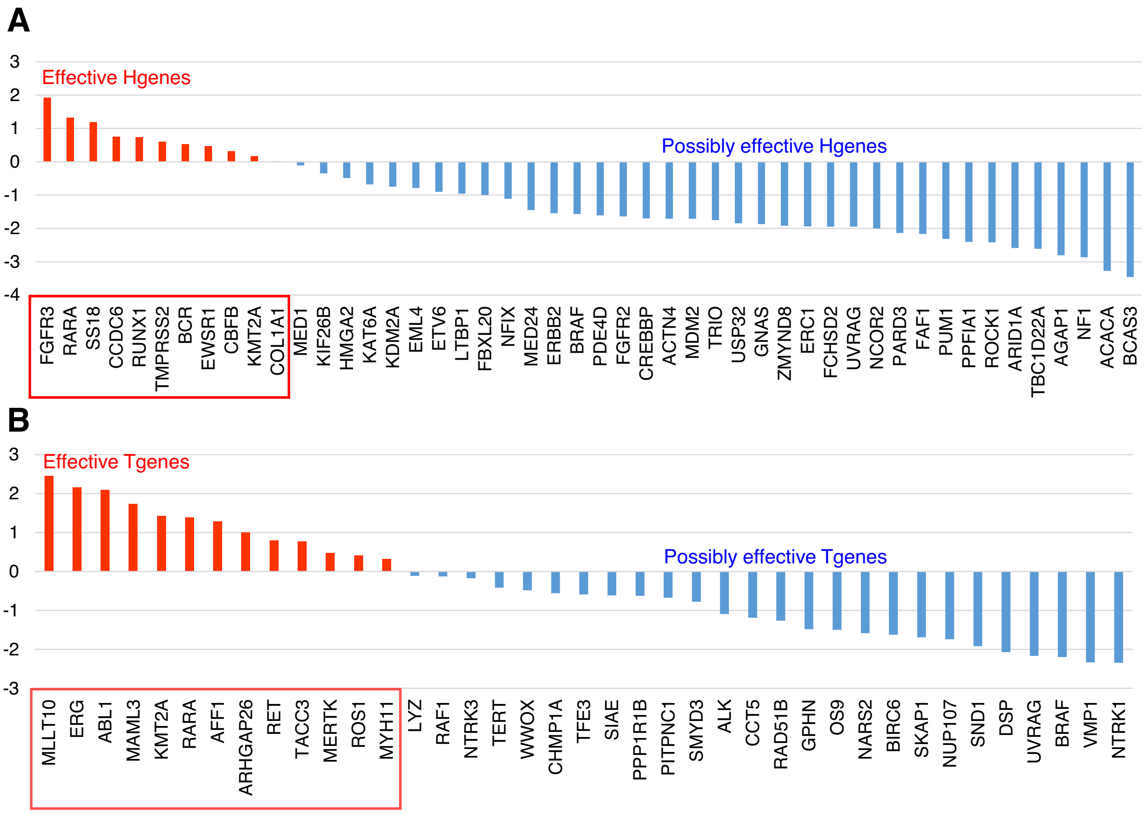


Figure 2. Gene assessment across pan-cancer in-frame FGs. (**A**) 5′-partner genes with high-MAII scores. (**B**) 3′-Partner genes with high-MAII scores. Y-axis presents Major Active Iso-fusion Index (MAII) score. MAII score can be calculated by $\log_2$(observed frequency/DoF score × 10). Degree of Frequency (DoF) score can be calculated by (# cancer types) × (# partners) × (# breakpoints). The genes that have the positive and bigger values of MAIIs are 'effective genes in pan-cancer fusion genes (eGinPCFG)'. The genes that have the negative and less values of MAIIs are 'possible effective genes in pan-cancer fusion genes (peGinPCFG)'.

ABL1 [Title/Abstract]) AND fusion [Title/Abstract])'. After a manual review of the abstracts, we found 85 FGs had literature evidence to support these FGs.

Open reading frame (ORF) annotation

We examined the open reading frames of individual fusion transcript sequences between the 5′- and 3′-partner genes. When both of the breakpoints in 5′- and 3′-genes are located inside of coding region (CDS) and the number of fusion transcript sequences from the transcription start site of 5′-gene to transcription end site of 3′-gene is a multiple of three, then we reported such fusion genes as 'in-frame'. If there is one or two nucleotide insertions, then we reported such FGs as 'frame-shift'. Except these two types of ORFs, there are 15 more ORFs such as '3UTR-CDS', '3UTR-3UTR', '3UTR-5UTR', '3UTR-intron', 'CDS-3UTR', 'CDS-5UTR', 'CDS-intron', '5UTR-CDS', '5UTR-3UTR', '5UTR-5UTR', '5UTR-intron', 'intron-CDS', 'intron-3UTR', 'intron-5UTR' and 'intron-intron'. Here, the FGs are named 'intron', when the breakpoint is located 6bp apart from the exon junction site to the intron direction. Since our fusion breakpoints were derived from the ESTs and RNA-seq data, all the breakpoint should be located inside of the exon. Therefore, if the breakpoint is located on the intron in at least one of the partners, then we report it as 'intron'. These categories are marked as not available (NA) ORF cases in our ORF classification. To do so, we took all matched Ensembl transcripts (ENSTs) into consideration (7). 37 900 and 40 109 breakpoints of 11 873 and 13 771 partner genes were matched with 48 781 and 54 296 ENSTs for the 5′- and 3′-genes, respectively. Total 60 466 ENSTs were mapped to 15 555 genes involved in 43 895 FGs.

Retention analysis of 39 protein features from UniProt

We first downloaded the protein information in general feature format (GFF) of 15,025 accessions of UniProt for a total of 17 110 genes involved in 43 895 FGs (8). UniProt provides the locus information of 39 protein features, including six molecule processing features, 13 region features, four site features, six amino acid modification features, two natural variation features, five experimental info features, and three secondary structure features. Since such feature locus information are based on amino acid sequence, the genomic sequence of breakpoints was converted to the amino acid information for each kinase, for all protein accessions of UniProt, ENST isoforms and multiple breakpoints of each partner. To map each feature to the human genome sequence, we used the GENCODE gene model of the human reference genome (hg19v19) available from the Encyclope-

Table 1. Statistics of retention status of 39 protein features from UniProt sequence annotation in the in-frame and frame-shift FGs

Subsection	**Gene location**	**Retention search in Hgenes**			**Retention search in Tgenes**		
	Retention categories	**# in-frame FGs with retention**	**# in-frame FGs without retention**	**# frame-shift FGs with retention**	**# in-frame FGs with retention**	**# in-frame FGs without retention**	**# frame-shift FGs with retention**
Molecular processing	Initiator methionine	1013	136	1172	141	808	135
	Signal peptide	829	145	812	190	1066	255
	Transit peptide	145	31	181	24	193	31
	Propeptide	92	139	93	146	139	205
	Chain	179	6836	260	1640	6343	1868
	Peptide	6	18	22	37	15	48
Regions	Topological domain	554	1094	543	941	1048	1152
	Transmembrane	744	1171	746	1226	867	1513
	Intramembrane	18	49	20	45	24	63
	Domain	1458	3264	1635	2449	2235	2664
	Repeat	461	786	563	657	525	713
	Calcium binding	44	99	58	93	51	94
	Zinc finger	281	705	293	428	317	550
	DNA binding	97	174	87	143	104	166
	Nucleotide binding	532	633	615	599	523	596
	Region	929	2029	1023	1649	1183	1688
	Coiled coil	373	957	401	773	569	754
	Motif	403	881	420	629	334	761
	Compositional bias	1016	1723	1097	1322	967	1474
Sites	Active site	310	679	324	682	257	706
	Metal binding	302	557	331	488	282	531
	Binding site	360	718	439	671	350	699
	Site	348	478	316	451	249	389
Amino acid modifications	Non-standard residue	0	2	2	2	0	4
	Modified residue	3315	4222	3683	3535	2819	3678
	Lipidation	82	116	94	120	75	172
	Glycosylation	731	967	746	1080	846	1289
	Disulfide bond	413	655	433	739	534	906
	Cross-link	486	795	588	621	374	651
Natural variations	Alternative sequence	2745	4267	3103	3637	3405	4020
	Natural variant	2764	4224	2994	4155	2917	4593
Experimental info	Mutagenesis	1087	1945	1184	1569	917	1634
	Sequence uncertainty	0	0	0	0	0	0
	Sequence conflict	2726	4108	3050	3883	2655	4116
	Non-adjacent residues	0	0	0	0	0	0
	Non-terminal residue	0	0	0	0	0	0
Secondary structure	Helix	1836	2675	1965	2324	1652	2473
	Beta strand	1728	2430	1824	2144	1511	2260
	Turn	1281	2072	1387	1826	1163	1892

dia of genes and gene variants (GENCODE) (15). For the 5′-partner genes, if the breakpoints occur after the 3′ end of the protein features, then the protein features are considered to be successfully retained in the fusion genes. On the contrary, if the kinase domain is not completely contained in the resultant FG, such fusion gene is thought to not retain its protein feature. Similarly, for 3′-partner gene, we considered the fusion genes to have retained protein features if the breakpoints appear on the 5′-end of the protein feature region. Table 1 shows the overall statistics of protein feature retention status for individual of 5′- and 3′-partner genes of FGs.

Creating fusion transcript and fusion amino acid sequences

Two different genes can form different FGs with multiple breakpoints based on multiple gene isoforms. Therefore, we considered all gene isoforms at each breakpoint. This study is designed to help users identify and validate FGs. Thus, we focused on the in-frame FGs. To obtain reliable FGs, we checked the distance between the two breakpoints in the case of intra-chromosomal rearrangements and created fusion sequences when these genes are apart more than 100 kb. We also selected FGs with both of their breakpoints aligned at the exon junction. To call each exon sequence of a given breakpoint, transcription start/end sites, and CDS start/end sites, we used the nibFrag utility of

UCSC Genome Browser based on ENCODE hg19 genome structure (14). After filtering, we have created 20 935, 20 944, and 45 634 fusion sequences corresponding to 8714, 8714 and 8788 in-frame FGs for amino acids, CDS transcripts and full-length transcripts, respectively. CDS transcript and amino acid fusion sequences were generated from 54 619 combinations of Ensembl transcripts between 13 491 and 13 207 ENSTs for the 5′- and 3′-genes, respectively.

Protein-protein interaction information

We downloaded the interactor information from BIOGRID (v 3.4.260) to provide PPI information for the wild-type proteins of individual fusion partners (16). There are limitations of this dataset, such as providing only the names of the interactors. Since we need to know the locus information for each PPI to search the retention of the PPI at the fusion protein level, we recognized that the 'Region' feature is one of the 39 protein features provided by UniProt, which includes the start and end locus information on the structure of each wild-type protein for each PPI. Therefore, we followed the same approach for the protein feature retention screening here. During the protein feature retention search, we also checked whether the fusion proteins retain these interaction loci.

Functional or gene category assignment

To assign the functional or gene categories, we integrated cancer genes in our study, such as oncogenes, tumor suppressors, epigenetic regulators, DNA damage repair genes, kinases, and transcription factors. The first four types of genes were downloaded from CancerGenes, a gene selection resource for cancer genome projects (17); TSGene, an updated literature-based knowledgebase for tumor suppressor genes (18); EpiFactors, a comprehensive database of human epigenetic factors and complexes (17,19); and the data from the studies by Knijnenburg *et al*. (20). For the gene groups of kinases and transcription factors, we examined the genes with kinase domains or DNA-binding domains and protein features.

Drug and disease information

The drug-target interactions (DTIs) were extracted from the DrugBank (April 2018, version 5.1.0). The duplicated DTI pairs were excluded (12). All drugs were grouped using the Anatomical Therapeutic Chemical (ATC) classification system codes. Disease-gene information was extracted from DisGeNet (June 2018, version 4.0) (13), a database of gene-disease associations.

Database architecture

The FusionGDB system is based on a three-tier architecture: client, server and database. It includes a user-friendly web interface, Perl's DBI module, and MySQL database. This database was developed on the MySQL 3.23 with the MyISAM storage engine.

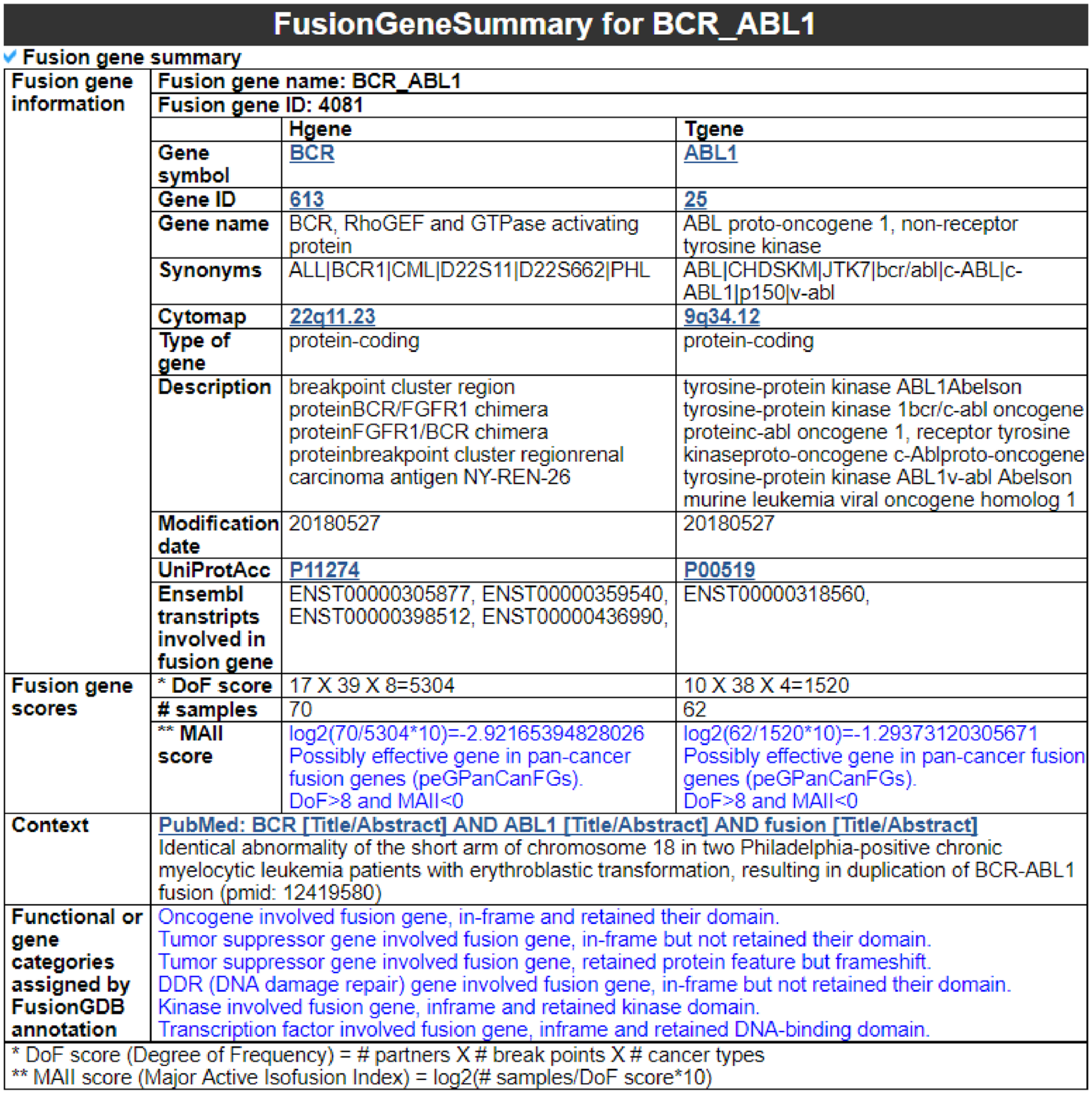
WEB INTERFACE AND ANALYSIS RESULTS

Fusion gene information category (FusionGeneSummary)

This category presents detailed information for both of partner genes and fusion genes (Figure 3). For each partner gene, it shows the basic gene information with the Ensembl transcript accessions, including the breakpoints of fusion genes in their gene structures. This category also provides the gene impact assessment scores in pan-cancer fusion genes, including the Degree of Frequency (DoF) score and Major Active Isofusion Index (MAII) score from previous studies (9,10). It is hypothesized if a gene is involved in a fusion gene in multiple cancer types with multiple breakpoints and multiple partner genes, this gene will play a critical role in tumorigenesis. Based on this hypothesis, we defined the DoF score by multiplying three factors. Through dividing the number of fusion positive samples by the number of all possible combination of gene fusion (DoF score), we defined the impact of each isofusion (Supplementary Table S5). Here, isofusion refers to a specific gene fusion combination with one specific partner gene and one specific breakpoint in a particular cancer type (MAII score). When the MAII score is greater than zero and DoF score exceeds 8, we refer to this gene as an 'effective gene in pan-cancer FGs (eGinPCFGs)'. For a gene with a MAII score less than zero and more than 8 DoF score, we refer to it as a 'possible effective Gene in Pan-Cancer FGs (peGinPCFGs)'., The user can also obtain the overall function of the fusion gene from this category based on the functional or gene categories assigned from the annotation of protein feature retention and overlapping with particular gene groups. The tables of this category list the Gene Ontology of each partner gene with evidence of Inferred from Direct Assay (IDA), the original fusion gene breakpoint information, and open reading frame (ORF) annotation results for each isofusion (21).

Fusion protein feature information category (FusionProtFeature)

In this category, we provide the detailed annotation of fusion protein function through the retention search of 39 protein features of UniProt based on the broken protein sequence in fusion protein. The first table shows the description of function of each partner, which is adopted from UniProt's explanation. The following tables show the information of protein feature retention status due to the breakage in the middle of coding region in fusion protein. To achieve this, we downloaded the gff format file of the sequence annotation (features) describing regions or sites of interest in the protein sequence, such as post-translational modifications, binding sites, enzyme active sites, local secondary structure or other characteristics from UniProt. For each fusion breakpoint with multiple gene isoform structures, we screened and investigated whether the protein features of each fusion gene are retained or not. Due to the limited space of the webpage, we only shows 13 regional features in the Table 1, such as 'calcium binding', 'coiled coil', 'compositional bias', 'DNA binding', 'domain', 'intramembrane', 'motif', 'nucleotide binding', 'region', 'repeat', 'topological domain', 'transmembrane',] and 'zinc



FusionGeneSummary for BCR_ABL1

✔ Fusion gene summary

Fusion gene information	Fusion gene name: BCR_ABL1		
	Fusion gene ID: 4081		
		Hgene	Tgene
	Gene symbol	BCR	ABL1
	Gene ID	613	25
	Gene name	BCR, RhoGEF and GTPase activating protein	ABL proto-oncogene 1, non-receptor tyrosine kinase
	Synonyms	ALL|BCR1|CML|D22S11|D22S662|PHL	ABL|CHDSKM|JTK7|bcr/abl|c-ABL|c-ABL1|p150|v-abl
	Cytomap	22q11.23	9q34.12
	Type of gene	protein-coding	protein-coding
	Description	breakpoint cluster region proteinBCR/FGFR1 chimera proteinFGFR1/BCR chimera proteinbreakpoint cluster regionrenal carcinoma antigen NY-REN-26	tyrosine-protein kinase ABL1Abelson tyrosine-protein kinase 1bcr/c-abl oncogene proteinc-abl oncogene 1, receptor tyrosine kinaseproto-oncogene c-Ablproto-oncogene tyrosine-protein kinase ABL1v-abl Abelson murine leukemia viral oncogene homolog 1
	Modification date	20180527	20180527
	UniProtAcc	P11274	P00519
	Ensembl transtripts involved in fusion gene	ENST00000305877, ENST00000359540, ENST00000398512, ENST00000436990,	ENST00000318560,
Fusion gene scores	* DoF score	17 X 39 X 8=5304	10 X 38 X 4=1520
	# samples	70	62
	** MAII score	log2(70/5304*10)=-2.92165394828026 Possibly effective gene in pan-cancer fusion genes (peGPanCanFGs). DoF>8 and MAII<0	log2(62/1520*10)=-1.29373120305671 Possibly effective gene in pan-cancer fusion genes (peGPanCanFGs). DoF>8 and MAII<0
Context	PubMed: BCR [Title/Abstract] AND ABL1 [Title/Abstract] AND fusion [Title/Abstract] Identical abnormality of the short arm of chromosome 18 in two Philadelphia-positive chronic myelocytic leukemia patients with erythroblastic transformation, resulting in duplication of BCR-ABL1 fusion (pmid: 12419580)		
Functional or gene categories assigned by FusionGDB annotation	Oncogene involved fusion gene, in-frame and retained their domain. Tumor suppressor gene involved fusion gene, in-frame but not retained their domain. Tumor suppressor gene involved fusion gene, retained protein feature but frameshift. DDR (DNA damage repair) gene involved fusion gene, in-frame but not retained their domain. Kinase involved fusion gene, inframe and retained kinase domain. Transcription factor involved fusion gene, inframe and retained DNA-binding domain.		
* DoF score (Degree of Frequency) = # partners X # break points X # cancer types ** MAII score (Major Active Isofusion Index) = log2(# samples/DoF score*10)			

Figure 3. FusionGeneSummary category. This category shows the overall function of fusion gene and each partner gene. It also provides information of the impact of each gene in pan-cancer fusion genes and functional category assigned by multiple functional annotations.

finger'. All of the retention information for 39 features are available from the download page.

ALK fusions are good examples of kinase domain retention in the fusion proteins. ALK is a transmembrane tyrosine kinase receptor. Wild-type ALK undergoes dimerization and subsequent autophosphorylation of the intracellular kinase domain upon ligand binding to its extracellular domain (22). Up to date, ALK has been found to be rearranged, mutated, or amplified in a series of tumors, including anaplastic large cell lymphomas (ALCL), neuroblastoma, and non-small cell lung cancer (NSCLC). Chromosomal rearrangements are the most common alterations in ALK and result in genetic fusions, such as EML4-ALK, TFG-ALK, NPM1-ALK,] and SQSTM1-ALK (23). Dimerization of fusion proteins leads to the constitutive activation of the kinase and transforming activity (24). Table 2 shows the retention search results for the ALK fusion protein features provided in FusionProtFeature category. In all of these fusion genes, the 'Protein kinase' domain of ALK was retained, but the 'Extracellular' topological domain was lost. This explains the mechanism of constitutive activation of ALK kinase domain in fusion proteins. Through the systematic annotations of the functional characteristics of a protein in fusion proteins, the user can easily understand the possible roles of fusion genes in relation to their specific tumorigenesis.

Fusion transcript and amino acid sequence category (FusionGeneSequence)

This category provides the fusion sequences for transcripts and amino acids. The DNA level sequences are not included in this category since our fusion breakpoints are exon junction sites or inside of exons identified from the Sanger transcript sequences or RNA-seq reads. After filtering steps (see

Table 2. Domain retention annotation of ALK fusion proteins from FusionGDB

Fusion gene	Gene	ENST	BP	Strand	BPexon	TotalExon	Protein feature loci	BPloci	TotalLen	Protein feature	Protein feature note
A. Retained protein features of in-frame ALK fusion proteins.											
EML4-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1116_1392	1057	1621	Domain	Protein kinase
EML4-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1197_1199	1057	1621	Region	Note = Inhibitor binding
EML4-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1060_1620	1057	1621	Topological domain	Cytoplasmic
NPM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1116_1392	1057	1621	Domain	Protein kinase
NPM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1197_1199	1057	1621	Region	Note = Inhibitor binding
NPM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1060_1620	1057	1621	Topological domain	Cytoplasmic
SQSTM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1116_1392	1057	1621	Domain	Protein kinase
SQSTM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1197_1199	1057	1621	Region	Note = Inhibitor binding
SQSTM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1060_1620	1057	1621	Topological domain	Cytoplasmic
TFG-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1116_1392	1057	1621	Domain	Protein kinase
TFG-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1197_1199	1057	1621	Region	Note = Inhibitor binding
TFG-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1060_1620	1057	1621	Topological domain	Cytoplasmic
B. Lost protein features of in-frame ALK fusion proteins.											
EML4-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	816_940	1057	1621	Compositional bias	Note = Gly-rich
EML4-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	264_427	1057	1621	Domain	MAM 1
EML4-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	478_636	1057	1621	Domain	MAM 2
EML4-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	437_473	1057	1621	Domain	Note = LDL-receptor class A
EML4-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	19_1038	1057	1621	Topological domain	Extracellular
EML4-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1039_1059	1057	1621	Transmembrane	Helical
NPM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	816_940	1057	1621	Compositional bias	Note = Gly-rich
NPM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	264_427	1057	1621	Domain	MAM 1
NPM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	437_473	1057	1621	Domain	Note = LDL-receptor class A
NPM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	478_636	1057	1621	Domain	MAM 2
NPM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	19_1038	1057	1621	Topological domain	Extracellular
NPM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1039_1059	1057	1621	Transmembrane	Helical
SQSTM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	816_940	1057	1621	Compositional bias	Note = Gly-rich
SQSTM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	264_427	1057	1621	Domain	MAM 1
SQSTM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	437_473	1057	1621	Domain	Note = LDL-receptor class A
SQSTM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	478_636	1057	1621	Domain	MAM 2
SQSTM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	19_1038	1057	1621	Topological domain	Extracellular
SQSTM1-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1039_1059	1057	1621	Transmembrane	Helical
TFG-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	816_940	1057	1621	Compositional bias	Note = Gly-rich
TFG-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	264_427	1057	1621	Domain	MAM 1
TFG-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	437_473	1057	1621	Domain	Note = LDL-receptor class A
TFG-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	478_636	1057	1621	Domain	MAM 2
TFG-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	19_1038	1057	1621	Topological domain	Extracellular
TFG-ALK	ALK	ENST00000389048	chr2:29446394	-	18	29	1039_1059	1057	1621	Transmembrane	Helical

Table 3. Protein-protein interaction (PPI) retention annotation of KMT2A-MLLT10 fusion protein from FusionGDB

A. Protein–protein interactors with each partner protein of wild type from BIOGRID-3.4.160

Hgene	Hgene's interactors	Tgene	Tgene's interactors
KMT2A	PPIE, PPP1R15A, KMT2A, ASH2L, HCFC1, HCFC2, MEN1, RBBP5, WDR5, AVP, INS, OXT, MAP3K5, HDAC1, CTBP1, CBX4, BMI1, CREBBP, SMARCB1, CXXC1, MYB, CTNNB1, SNW1, E2F2, E2F4, E2F6, PSIP1, MLLT4, POLR2A, KAT8, RNF2, TP53, SBF1, MTM1, SET, HIST1H3A, HIST1H4A, KAT6A, ELL, AFF1, AFF4, CDK9, CCNT1, CTR9, LEO1, PAF1, CDC73, WDR61, MLLT3, DOT1L, SKP2, HIST3H3, SVIL, HIST2H3C, SIN3A, MLLT1, RUNX1, CBFB, H3F3A, SIRT7, ASB2, TCEB1, TCEB2, CBX8, TOP1, TAF6, NCL, HECW2, LGR4, CSNK2A2, SENP3, SYMPK, PKN1, PIH1D1, KRAS, TAF1, CHD3, SMARCA2, SMARCC2, SMARCC1, HDAC2, RBBP4, RBBP7, TBP, MBD3, SAP30, RAN, TAF9, TASP1, HIST1H2BG, EWSR1, DYNLT1, KIF11, ING4, ZNF131, ASB7	MLLT10	YEATS4, SMARCB1, SS18, MLLT10, DOT1L, MLLT1, MLLT3, AFF1, DISC1, NDEL1, ELAVL1, CENPJ, TMPO, ZNF526, MLLT6, TCP10L, KIF6, PMF1

B. Still kept PPIs in the in-frame KMT2A-MLLT10 fusion protein with retention.

Gene	Hbp	Tbp	ENST	Strand	BPexon	TotalExon	Protein feature loci	BPloci	TotalLen	Still interaction with
MLLT10	chr11:118352807	chr10:22002700	ENST00000377091	+	0	5	141_233	−59	127	FSTL3
MLLT10	chr11:118352807	chr10:22002700	ENST00000377100	+	0	4	141_233	−119	180	FSTL3
MLLT10	chr11:118355029	chr10:21959377	ENST00000377091	+	0	5	141_233	−59	127	FSTL3
MLLT10	chr11:118355029	chr10:21959377	ENST00000377100	+	0	4	141_233	−119	180	FSTL3
MLLT10	chr11:118355690	chr10:21959377	ENST00000377091	+	0	5	141_233	−59	127	FSTL3
MLLT10	chr11:118355690	chr10:21959377	ENST00000377100	+	0	4	141_233	−119	180	FSTL3

C. Lost PPIs in the in-frame KMT2A-MLLT10 fusion protein without retention.

Gene	Hbp	Tbp	ENST	Strand	BPexon	TotalExon	Protein feature loci	BPloci	TotalLen	Interaction lost with
KMT2A	chr11:118352807	chr10:22002700	ENST00000354520	+	7	35	1584_1600	1337	3932	histone H3K4me3
KMT2A	chr11:118352807	chr10:22002700	ENST00000389506	+	7	36	1584_1600	1337	3970	histone H3K4me3
KMT2A	chr11:118352807	chr10:22002700	ENST00000534358	+	7	36	1584_1600	1337	3973	histone H3K4me3
KMT2A	chr11:118355029	chr10:21959377	ENST00000354520	+	9	35	1584_1600	1406	3932	histone H3K4me3
KMT2A	chr11:118355029	chr10:21959377	ENST00000389506	+	9	36	1584_1600	1406	3970	histone H3K4me3
KMT2A	chr11:118355029	chr10:21959377	ENST00000534358	+	9	36	1584_1600	1406	3973	histone H3K4me3
KMT2A	chr11:118355690	chr10:21959377	ENST00000354520	+	1	35	1584_1600	0	3932	histone H3K4me3
KMT2A	chr11:118355690	chr10:21959377	ENST00000389506	+	10	36	1584_1600	1444	3970	histone H3K4me3
KMT2A	chr11:118355690	chr10:21959377	ENST00000534358	+	10	36	1584_1600	1444	3973	histone H3K4me3
KMT2A	chr11:118352807	chr10:22002700	ENST00000354520	+	7	35	3764_3771	1337	3932	WDR5
KMT2A	chr11:118352807	chr10:22002700	ENST00000389506	+	7	36	3764_3771	1337	3970	WDR5
KMT2A	chr11:118352807	chr10:22002700	ENST00000534358	+	7	36	3764_3771	1337	3973	WDR5
KMT2A	chr11:118355029	chr10:21959377	ENST00000354520	+	9	35	3764_3771	1406	3932	WDR5
KMT2A	chr11:118355029	chr10:21959377	ENST00000389506	+	9	36	3764_3771	1406	3970	WDR5
KMT2A	chr11:118355029	chr10:21959377	ENST00000534358	+	9	36	3764_3771	1406	3973	WDR5
KMT2A	chr11:118355690	chr10:21959377	ENST00000354520	+	1	35	3764_3771	0	3932	WDR5
KMT2A	chr11:118355690	chr10:21959377	ENST00000389506	+	10	36	3764_3771	1444	3970	WDR5
KMT2A	chr11:118355690	chr10:21959377	ENST00000534358	+	10	36	3764_3771	1444	3973	WDR5
MLLT10	chr11:118352807	chr10:22002700	ENST00000307729	+	12	23	141_233	566	1069	FSTL3
MLLT10	chr11:118352807	chr10:22002700	ENST00000377072	+	13	24	141_233	582	1476	FSTL3
MLLT10	chr11:118355029	chr10:21959377	ENST00000307729	+	8	23	141_233	265	1069	FSTL3
MLLT10	chr11:118355029	chr10:21959377	ENST00000377072	+	8	24	141_233	265	1476	FSTL3
MLLT10	chr11:118355690	chr10:21959377	ENST00000307729	+	8	23	141_233	265	1069	FSTL3
MLLT10	chr11:118355690	chr10:21959377	ENST00000377072	+	8	24	141_233	265	1476	FSTL3

Methods), we have created 20 935, 20 944 and 45 634 fusion sequences corresponding to 8714, 8714 and 8788 in-frame FGs for amino acids, CDS transcripts, and full-length transcripts, respectively. Taking the in-frame TMPRSS2-ERG fusion gene as example, this category provides 18 full-length fusion transcript sequences, 18 fusion transcript sequences of CDS and 18 fusion amino acid sequences based on three TMPRSS2 isoforms with four breakpoints and three ERG isoforms with two breakpoints. In contrast, when we searched the fusion transcripts and amino acid sequences of the TMPRSS2-ERG fusion from the National Center for Biotechnology Information (NCBI), we were only able to obtain eight fusion transcript sequences and one fusion amino acid sequence (25). Therefore, FusionGeneSequence category in FusionGDB can be used as an important reference or resource for the cancer and drug research communities.

Fusion protein-protein interaction information category (FusionGenePPI)

Protein–protein interaction (PPI) plays a crucial role in the cellular biological processes (26). Specifically, studies have shown that the loss of PPI in the fusion genes leads to abnormal regulation of downstream genes. The well-known case is lysine methyltransferase 2A (MLL) fusion proteins. MLL translocations are associated with a wide array of hematologic malignancy and mutations in several family members are associated with cancer and developmental disorders (27). Due to the truncation of the region of PPIs in the C-terminal MLL protein, MLL fusion proteins fail to retain the PHD finger region, the HCF1 interaction region and the SET domain region. As a result, MLL fusion proteins lose their ability to catalyze H3K4 methylation (28). MLL1 and MLL2 function as large macromolecular complexes composed of >30 subunits, including several core components. One of the components, WD repeat domain 5 (WDR5) specifically recognizes histone H3 methylated at lysine (29). Consequently, MLL regulated genes in down-stream showed different expressional regulation (30). Through the FusionGenePPI category, *in-silico* evidence of these facts can be viewed. Firstly, this category shows the interactors at the wild type proteins for both of 5′- and 3′-partners as shown in Table 3A. The Table 3B and C provides the information of retained or lost PPIs in fusion proteins, respectively. As shown in Table 3C, MLL fusion proteins lose their interactions with histone H3K4me3 and WDR5, consistent with the previous studies. Therefore, our systematic investigation of PPI retention in fusion proteins will be very useful for understanding the genetic and epigenetic effect of fusion proteins in cancer.

Pharmacological information and disease information categories (FusionGeneDrug and FusionGeneDisease)

FusionGeneDrug category provides the pharmacological information associated with the genes involved in FGs from DrugBank (12). Overall, FusionGDB includes 1397 drugs targeting 1314 proteins involved in 8478 FGs. Interestingly, 1341 (95.99%) are FDA approved drugs targeting 1302 genes. 1619 out of 8478 FGs were overlapped with 8714 in-frame FGs that have fusion amino acid sequences. Among the 1302 genes with FDA approved targeting drugs, 529, 472, 138, 128, 85 and 69 were overlapped with IUPHAR drug target genes, essential genes, oncogenes, tumor suppressors, kinases, and transcription factors (31–33), respectively. According to the generic name stems from Drug Information Portal of National Library of Medicine, among 1312 drugs, there were 19 tyrosine-kinase inhibitors (TKIs) with a stem of '-tinib' at the end of their generic names: Afatinib, Axitinib, Bosutinib, Cabozantinib, Ceritinib, Crizotinib, Dasatinib, Erlotinib, Gefitinib, Ibrutinib, Imatinib, Lapatinib, Lenvatinib, Nilotinib, Ponatinib, Ruxolitinib, Sunitinib, Tofacitinib and Trametinib. These TKIs targeted 51 protein kinases involved in 303 different kinase fusion genes. Furthermore, there were 37 monoclonal antibodies (mAbs) with a suffix '-mab' at the end of their generic names. Monoclonal antibodies-based therapy for cutaneous T-cell lymphoma has demonstrated high response rates and a favorable toxicity profile in clinical trials (34). Here, 197 FGs are involved in the 37 mAbs-targeted 37 different genes.

FusionGeneDisease category shows the related disease information for each gene generated from DisGeNet (13). 6169 out of 15 555 genes involved in 43 895 FGs were reported to be associated with 4679 different types of diseases reported previously. Overall, 2164 (35.08%) out 6169 FGs are overlapped with the cancer genes from the Catalogue of Cancer Genes.

DISCUSSION AND FUTURE DIRECTION

FusionGDB is the first database that systematically annotates the function of fusion genes across pan-cancer. To serve broad biomedical research communities, we will continuously update and curate FGs routinely by checking new fusion gene or fusion protein data. We have identified 331, 303, 840 and 667 in-frame FGs with retaining kinase domains, DNA-binding domains, oncogene domains, and epifactor domains, respectively, and 976 FGs with no retained PPIs. Their genomic relationships, interactions, association with other oncogenes, and downstream effects are important for studying their potential roles in tumorigenesis, as well as developing possible molecular targets. We will extend our current approaches to further investigate the clinically important FGs and address their acting mechanisms as described above in near future. The easy-to-use website provides multiple annotation results to researchers and facilitates comprehensive functional studies of FGs. Thus, FusionGDB will be a useful resource for many investigators in pathology, cancer genomics and precision medicine, drug and therapeutic research, among others.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the members of the Center for Computational Systems Medicine (CCSM) for valuable discussion. We thank Dr. Weiling Zhao for improving the English of the manuscript.

FUNDING

National Institutes of Health [NIH R01GM123037, U01AR069395-01A1, U01CA166886 to X.Z.]; The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Funding for open access charge: Dr Carl V. Vartian Chair Professorship Funds to Dr Zhou from the University of Texas Health Science Center at Houston.

Conflict of interest statement. None declared.

REFERENCES

1. Liu,C., Ma,J., Chang,C.J. and Zhou,X. (2013) FusionQ: a novel approach for gene fusion detection and quantification from paired-end RNA-Seq. *BMC Bioinformatics*, **14**, 193.
2. Gorohovski,A., Tagore,S., Palande,V., Malka,A., Raviv-Shay,D. and Frenkel-Morgenstern,M. (2017) ChiTaRS-3.1-the enhanced chimeric transcripts and RNA-seq database matched with protein-protein interactions. *Nucleic Acids Res.*, **45**, D790–D795.
3. Stransky,N., Cerami,E., Schalm,S., Kim,J.L. and Lengauer,C. (2014) The landscape of kinase fusions in cancer. *Nat. Commun.*, **5**, 4846.
4. Lee,M., Lee,K., Yu,N., Jang,I., Choi,I., Kim,P., Jang,Y.E., Kim,B., Kim,S., Lee,B. *et al.* (2017) ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. *Nucleic Acids Res.*, **45**, D784–D789.
5. Hu,X., Wang,Q., Tang,M., Barthel,F., Amin,S., Yoshihara,K., Lang,F.M., Martinez-Ledesma,E., Lee,S.H., Zheng,S. *et al.* (2018) TumorFusions: an integrative resource for cancer-associated transcript fusions. *Nucleic Acids Res.*, **46**, D1144–D1149.
6. Gao,Q., Liang,W.W., Foltz,S.M., Mutharasu,G., Jayasinghe,R.G., Cao,S., Liao,W.W., Reynolds,S.M., Wyczalkowski,M.A., Yao,L. *et al.* (2018) Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.*, **23**, 227–238.
7. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Giron,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
8. UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699.
9. Kim,P., Jia,P. and Zhao,Z. (2018) Kinase impact assessment in the landscape of fusion genes that retain kinase domains: a pan-cancer study. *Brief. Bioinform.*, **19**, 450–460.
10. Kim,P., Ballester,L.Y. and Zhao,Z. (2017) Domain retention in transcription factor fusion genes and its biological and clinical implications: a pan-cancer study. *Oncotarget*, **8**, 110103–110117.
11. Frenkel-Morgenstern,M., Gorohovski,A., Tagore,S., Sekar,V., Vazquez,M. and Valencia,A. (2017) ChiPPI: a novel method for mapping chimeric protein-protein interactions uncovers selection principles of protein fusion events in cancer. *Nucleic Acids Res.*, **45**, 7094–7105.
12. Wishart,D.S., Feunang,Y.D., Guo,A.C., Lo,E.J., Marcu,A., Grant,J.R., Sajed,T., Johnson,D., Li,C., Sayeeda,Z. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
13. Pinero,J., Bravo,A., Queralt-Rosinach,N., Gutierrez-Sacristan,A., Deu-Pons,J., Centeno,E., Garcia-Garcia,J., Sanz,F. and Furlong,L.I. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.
14. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
15. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
16. Chatr-Aryamontri,A., Oughtred,R., Boucher,L., Rust,J., Chang,C., Kolas,N.K., O'Donnell,L., Oster,S., Theesfeld,C., Sellam,A. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, **45**, D369–D379.
17. Higgins,M.E., Claremont,M., Major,J.E., Sander,C. and Lash,A.E. (2007) CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res.*, **35**, D721–D726.
18. Zhao,M., Kim,P., Mitra,R., Zhao,J. and Zhao,Z. (2016) TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res.*, **44**, D1023–D1031.
19. Medvedeva,Y.A., Lennartsson,A., Ehsani,R., Kulakovskiy,I.V., Vorontsov,I.E., Panahandeh,P., Khimulya,G., Kasukawa,T., Consortium,F. and Drablos,F. (2015) EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database*, **2015**, bav067.
20. Knijnenburg,T.A., Wang,L., Zimmermann,M.T., Chambwe,N., Gao,G.F., Cherniack,A.D., Fan,H., Shen,H., Way,G.P., Greene,C.S. *et al.* (2018) Genomic and molecular landscape of DNA damage repair deficiency across the cancer genome atlas. *Cell Reports*, **23**, 239–254.
21. Mi,H., Huang,X., Muruganujan,A., Tang,H., Mills,C., Kang,D. and Thomas,P.D. (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, **45**, D183–D189.
22. Della Corte,C.M., Viscardi,G., Di Liello,R., Fasano,M., Martinelli,E., Troiani,T., Ciardiello,F. and Morgillo,F. (2018) Role and targeting of anaplastic lymphoma kinase in cancer. *Mol. Cancer*, **17**, 30.
23. Wu,W., Haderk,F. and Bivona,T.G. (2017) Non-canonical thinking for targeting ALK-fusion onco-proteins in lung cancer. *Cancers*, **9**, 164–181.
24. Souttou,B., Carvalho,N.B., Raulais,D. and Vigny,M. (2001) Activation of anaplastic lymphoma kinase receptor tyrosine kinase induces neuronal differentiation through the mitogen-activated protein kinase pathway. *J. Biol. Chem.*, **276**, 9526–9531.
25. Sharma,S., Ciufo,S., Starchenko,E., Darji,D., Chlumsky,L., Karsch-Mizrachi,I. and Schoch,C.L. (2018) The NCBI BioCollections database. *Database*, **2018**, doi:10.1093/database/bay006.
26. Jiao,X. and Ranganathan,S. (2017) Prediction of interface residue based on the features of residue interaction network. *J. Theor. Biol.*, **432**, 49–54.
27. Alicea-Velazquez,N.L., Shinsky,S.A., Loh,D.M., Lee,J.H., Skalnik,D.G. and Cosgrove,M.S. (2016) Targeted disruption of the interaction between WD-40 repeat protein 5 (WDR5) and Mixed Lineage Leukemia (MLL)/SET1 family proteins specifically inhibits MLL1 and SETd1A methyltransferase complexes. *J. Biol. Chem.*, **291**, 22357–22372.
28. Zhang,Y., Chen,A., Yan,X.M. and Huang,G. (2012) Disordered epigenetic regulation in MLL-related leukemia. *Int. J. Hematol.*, **96**, 428–437.
29. Cierpicki,T. and Grembecka,J. (2014) Challenges and opportunities in targeting the menin-MLL interaction. *Future Med. Chem.*, **6**, 447–462.
30. Slany,R.K. (2005) Chromatin control of gene expression: mixed-lineage leukemia methyltransferase SETs the stage for transcription. *PNAS*, **102**, 14481–14482.
31. Harding,S.D., Sharman,J.L., Faccenda,E., Southan,C., Pawson,A.J., Ireland,S., Gray,A.J.G., Bruce,L., Alexander,S.P.H., Anderton,S. *et al.* (2018) The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Res.*, **46**, D1091–D1106.
32. Wang,T., Birsoy,K., Hughes,N.W., Krupczak,K.M., Post,Y., Wei,J.J., Lander,E.S. and Sabatini,D.M. (2015) Identification and characterization of essential genes in the human genome. *Science*, **350**, 1096–1101.
33. Manning,G., Whyte,D.B., Martinez,R., Hunter,T. and Sudarsanam,S. (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.
34. Geskin,L.J. (2015) Monoclonal antibodies. *Dermatol. Clin.*, **33**, 777–786.