Published online 10 November 2021 *Nucleic Acids Research, 2022, Vol. 50, Database issue* ***D1221–D1230***
https://doi.org/10.1093/nar/gkab1056

FusionGDB 2.0: fusion gene annotation updates aided by deep learning

Pora Kim [1,*], Hua Tan [1], Jiajia Liu[1], Haeseung Lee[2], Hyesoo Jung[3], Himanshu Kumar[1] and Xiaobo Zhou [1,4,5]

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA, [2]Intellectual Information Team, Future Medicine Division, Korea Institute of Oriental Medicine, Daejeon, South Korea, [3]Department of Neurology, Asan Medical Center, Seoul, Korea, [4]McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA and [5]School of Dentistry, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Received September 14, 2021; Revised October 10, 2021; Editorial Decision October 16, 2021; Accepted November 03, 2021

ABSTRACT

A knowledgebase of the systematic functional annotation of fusion genes is critical for understanding genomic breakage context and developing therapeutic strategies. FusionGDB is a unique functional annotation database of human fusion genes and has been widely used for studies with diverse aims. In this study, we report fusion gene annotation updates aided by deep learning (FusionGDB 2.0) available at https://compbio.uth.edu/FusionGDB2/. FusionGDB 2.0 has substantial updates of contents such as up-to-date human fusion genes, fusion gene breakage tendency score with FusionAI deep learning model based on 20 kb DNA sequence around BP, investigation of overlapping between fusion breakpoints with 44 human genomic features across five cellular role's categories, transcribed chimeric sequence and following open reading frame analysis with coding potential based on deep learning approach with Ribo-seq read features, and rigorous investigation of the protein feature retention of individual fusion partner genes in the protein level. Among ∼102k fusion genes, about 15k kept their ORF as In-frames, which is two times compared to the previous version, FusionGDB. FusionGDB 2.0 will be used as the reference knowledgebase of fusion gene annotations. FusionGDB 2.0 provides eight categories of annotations and it will be helpful for diverse human genomic studies.

INTRODUCTION

Gene fusion is one of the hallmarks of the cancer genome via chromosomal rearrangement initiated by DNA double-strand breakage. Fusion genes have the breakpoints of structural variants on their gene body and provide a highlighted structural variant resource for studying the genomic breakages with expression and potential pathogenic impacts. A knowledgebase of the systematic functional annotation of fusion genes is critical for understanding genomic breakage context and developing therapeutic strategies. For this aim, previously, we built FusionGDB (1), which is a unique functional annotation database of human fusion genes. To date, FusionGDB has been widely used for diverse studies with different human genomic study aims. It has been visited ∼27k times by the users and this study was often cited from the peer-review scientific papers since its publication in 2019. To serve the research communities of diverse genomic mechanisms studies, we report FusionGDB 2.0, which has substantial updates of contents, including the analyses aided by deep learning approaches. This updated version of human fusion gene annotation will provide unique and essential analyses results to the research communities.

DATABASE UPDATE OVERVIEW

In this study, we report FusionGDB 2.0, the update version of fusion gene annotation database including the analyses aided by deep learning approaches such as (i) up-to-date human fusion genes with breakpoint location information with gene structure, gene assessment in pan-cancer fusions, and updated functional category assignment to individual fusions, (ii) fusion gene breakage tendency scores from FusionAI (2) deep learning model based on 20 kb DNA sequence around BP area, (iii) investigation of overlapping between fusion breakpoints (and high breakage feature importance scored regions) with 44 human genomic features across five cellular role's categories (i.e. integration sites of 6 viruses, 13 types of repeats, 5 types of structural variants, 15 different chromatin stated regions and 5 gene

*To whom correspondence should be addressed. Tel: +1 713 500 3636; Email: pora.kim@uth.tmc.edu

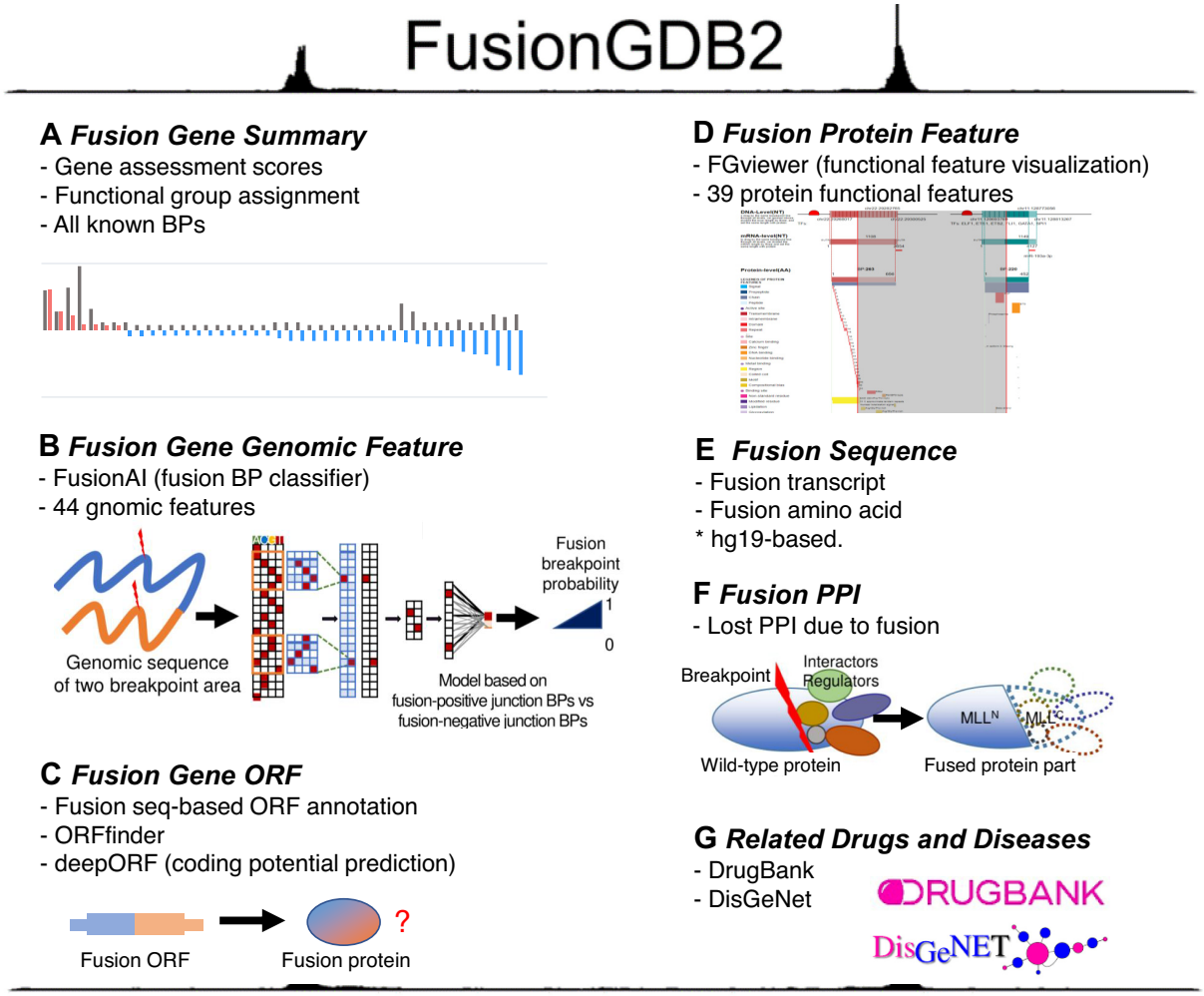

Figure 1. Overview of FusionGDB 2.0. Updated FusionGDB provides multiple annotations on fusion genes in eight categories including fusion gene summary, fusion gene genomic feature, fusion gene ORF, fusion protein features, fusion sequence, fusion PPI, related drugs and related diseases.

expression regulatory regions), (iv) transcribed and translated chimeric sequences, (v) open reading frame (ORF) analyses and enhanced coding potential investigation through deep learning approach using Ribo-seq alignment features (deepORF) rather than classical comparison between coding vs non-coding gene structure and (vi) investigation of the protein feature retention of individual fusion partner genes in the protein level using FGviewer (3), a visualization tool for functional features of human fusion genes. Among ∼102k fusion genes, about 15k fusions were predicted as keeping their open reading frames as the in-frame, which is two times compared to the previous version of FusionGDB.

FusionGDB 2.0 provides eight categories of annotations: Fusion Gene Summary, Fusion Gene ORF analysis, Fusion Gene Genomic Features, Fusion Protein Features, Fusion Gene Sequence, Fusion Gene PPI analysis, Related Drugs and Related Diseases. For 511 highly recurrent fusion genes that have been expressed in more than five samples, we performed manual curation of PubMed articles. All such information is included and downloadable in the database with a unique and non-redundant format. Figure 1 summarizes the overview of FusionGDB 2.0 annotations. All entries and annotation data are available for browsing and downloading on the website with unique and efficient visualization (https://compbio.uth.edu/FusionGDB2). The main features of FusionGDB 2.0 annotations are summarized below.

i. Fusion Gene Summary information category provides five unique and useful fusion gene information including basic gene information, gene assessment score in pan-cancer, study context, most frequent breakpoint information, and anticipated loss of major functional domain due to fusion event. In this update, we extended the functional gene groups to assign potential loss-of-functional effect with ORF annotation and provide all breakpoints based on the gene structures of individual partners using the UCSC genome browser.
ii. Fusion Gene ORF Analysis category provides the ORF annotation results. It provides the coding potential study results based on three approaches. First, we investigated the ORF whether in-frame or frame-shift if both breakpoints are located in the coding sequence

(CDS) area. Second, to have the potential amino acid sequence, we input the in-frame fusions' full-length transcript sequences to ORFfinder by NCBI (4). Third, we also input these in-frame fusions' transcript sequences into a deep learning model classifier to identify the coding potential of fusion transcripts. This model was built by training between the coding genes mapped by Ribo-seq reads with high reliability and non-coding genes not mapped by any Ribo-seq reads.

iii. Fusion Gene Genomic Features category provides the potential human genomic features related to the fusion gene breakpoints or near to that area. For all fusion transcripts whose breakpoints are located at the exon junction boundaries, we ran an in-house model, FusionAI, which is a deep learning-based classifier between fusion breakpoint sequence and no fusion breakpoint sequence. FusionAI predicts the fusion breakage potential of a fusion gene from a 20k bp length DNA sequence. Then, we investigated the overlap between the diverse human genomic features with the top 10% of the feature importance scored regions and all the 20kb sequences. To do this, we integrated 44 different types of human genomic feature information across five big categories.

iv. Fusion Protein Features category provides the retention information of 39 protein features of fusion proteins based on the canonical protein structures with considering multiple transcripts from gene isoforms and multiple breakpoints. By focusing on the types of protein features, the user can infer the overall functional loss or different regulation for a fusion gene. In this updated version, we also added the link to an in-house web tool, FGviewer. FGviewer provides functional feature annotations at four different levels: DNA-, RNA-, protein- and pathogenic levels. The same breakpoint lines across four tiers classify between fusion involving or non-involving zones with multiple types of functional features.

v. Fusion Gene Sequence information category provides the full-length fusion transcript sequence and predicted amino acid sequences considering the multiple breakpoints with matched gene isoforms.

vi. Fusion Gene PPI analysis category provides the potential protein–protein interactions between fusion protein and small molecules in the cells that are anticipated as losing or retaining their interactions due to gene fusion events.

vii. Related Drugs and Related Diseases categories provide the related information of 3635 approved drugs that target 1205 fusion genes as well as 5981 fusion genes that are reported to be associated with 11 264 different types of diseases.

DATA INTEGRATION AND ANNOTATIONS

Fusion gene information

50 360 and 52 737 fusion genes and their related information were downloaded from the comprehensive database of chimeric transcripts matched with druggable fusions and 3D chromatin maps (ChiTaRS 5.0, http:// http://chitars.md.biu.ac.il/, January 2020) (5) and an updated and expanded database of fusion genes (ChimerDB 4.0, http://www.kobic.re.kr/chimerdb/, January 2020) (6), respectively. Of these, 50 360 and 50 931 fusion genes were from Entrez Sanger sequences and TCGA samples, respectively. By the union of these, we obtained 102 647 unique fusion genes in total. For the genome coordinates information of some fusion breakpoints from ChimerDB 4.0, we lifted over from the human reference genome GRCh38 to GRCh37 using Batch Coordinate Conversion (liftOver) utility from UCSC Genome Browser (7). For all fusion gene information from these two resources, the following information was collected: sample ID or expressed sequence tag (EST) ID, fusion partner gene names, exon junction breakpoint information. We followed the definition of fusion gene's direction for the Hgene (Head gene or 5′-gene) and Tgene (Tail gene or 3′-gene) to these datasets.

Fusion genomic feature analyses

Recently, we developed a deep learning-based classifier between fusion gene and no fusion gene breakpoint sequences (FusionAI, https://compbio.uth.edu/FusionGDB2/FusionAI/) (2). It aims to learn sequence context around the fusion gene breakpoints. The input of FusionAI is the flanking DNA sequence of ±5 kb from the two breakpoints of a fusion gene. Since FusionAI is designed based on the exon junctional breakpoints, for all fusion transcripts whose breakpoints are located at the exon junction boundaries, we made input sequences and ran FusionAI. Then, we had FusionAI output scores, which is the probability of being used as the fusion gene breakpoint. We also investigated what human genomic sequence features are enriched in the fusion gene breakpoint area. To do this, we overlapped the top 10% of the feature importance scored regions among the 20kb sequence of fusion gene breakpoints with 44 different types of human genomic feature loci information across five big categories including virus integration sites, repeats, structural variants, chromatin states, and gene expression regulation. First, we downloaded the virus integration site information from the VISDB (8) and we lifted it over to the hg19 version using the liftover tool from the UCSC Genome Browser to set the same human genome version that was used in FusionAI model. We integrated 13 types of repeats (Alu repeats, A-Phased repeats, Directed repeats, DNA transposons, 'G-Quadruplex, forming repeats', Inverted repeats, L1 repeats, L2 repeats, 'Low_complexity, A/T rich regions', Microsatellites, MIR repeats, Mirror repeats, and Z-DNA motifs) from RepeatMasker (9) and MicroSatellite DataBase (MSDB) (10). For the diverse types of structural variants including the copy number variants, we downloaded the arranged breakpoint information of the structural variants from dbVar (11). The chromatin states category include the loci of 15 different types of chromatin states such as 1_TssA, 2_TssAFlnk, 3_TxFlnk, 4_Tx, 5_TxWk, 6_EnhG, 7_Enh, 8_ZNF_Rpts, 9_Het, 10_TssBiv, 11_BivFlnk, 12_EnhBiv, 13_ReprPC, 14_ReprPCWk and 15_Quies, from the previous study on the chromatin state calls using a 15-state model for 12 cell lines, were obtained from the Roadmap Epigenomics

Mapping Consortium (12,13). The gene expression regulatory category includes five types of features as CPGisland, Methylation, Promoters, ReplicationTiming, and TAD boundaries. The information of the first three feature categories was downloaded from the FANTOM5 collection (14). We downloaded the replication timing-specific peak regions from the ENCODE portal site by selecting the assay type of the replication timing (15). We used 2477 loci of common TAD boundaries from a previous study that made high-resolution chromosome conformation (Hi-C) datasets from five human cell lines based on the (16).

Open reading frame (ORF) annotations

Between the 5′-partner gene and the 3′-partner gene, we checked the open reading frame of the full-length fusion transcript sequence. When both breakpoints of 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 this fusion gene as 'in-frame'. If there is one or two nucleotide insertion, then we reported as the 'frame-shift'. Except for these two 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 'intron' is reported when the breakpoint is located 6 bp 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 breakpoints should be located inside of the exon. Therefore, if the breakpoint is located on the intron, then we report as an intron, and the ORFs including the intron in at least one of the partners, we set aside these categories to not available (NA) ORF cases in our ORF classification. For these analyses, we considered all matched Ensembl transcripts (ENSTs) (17). There were 73 784 and 79 196 breakpoints of 15 141 and 16 814 partner genes that were matched with 58 709 and 64 273 ENSTs for the 5′- and 3′-genes, respectively. Total 68 877 ENSTs were mapped to 18 407 genes involved in 150 496 fusion genes. Among these, we identified 16 146 in-frame fusion genes. With considering multiple breakpoints and gene isoforms, we made 83 291 full-length fusion transcript sequences. For these in-frame fusion transcripts, we ran the open reading frame finder (ORFfinder), which is mainly selecting the longest ORFs from the six-frames-based translation. Then, we found 42 110 fusion amino acid sequences from 14 569 fusion genes. From the result of ORFfinder, we adopted the translated ORF sequences and the position of start and end from ORFfinder outputs.

Prediction of the coding potential of fusion transcripts using a deep learning model

The tools to predict the coding potential of transcripts usually train their models by comparing the transcript sequences between coding genes versus non-coding genes of the human reference genome. However, we noticed that those are not enough data to find the best features of the coding transcripts that are bound by the ribosomes. Therefore, we made the positive data of the coding genes that were experimentally read by the ribosomes. We also made the negative data of non-coding genes that were not read by ribosomes. To have the ORFs read by Ribo-seq data in humans, we downloaded the ORFs that were read in 102 Ribo-seq data of 56 different human cells including cancers and normal cells from RPFdb v2.0 (18). We lifted over all genomic coordinates into hg38. To make the positive data, we selected the ORFs that have the median read count is greater or equal to 22 and the RibORF score (19) is bigger than 0.7 for the reliability. Then, there were 23,587 Ensembl transcripts with 2,367,842 read ORFs in coding genes. To make the negative data, first, we integrated the read lncRNA ORFs from the 102 Ribo-seq data as the minimum read count was bigger than one and the RiboORF score was bigger than 0.7. Then, there were 158 387 Ensembl transcripts with 25 218 872 read ORFs in the non-coding genes. Since we are investigating the lncRNA not read by ribosomes, we excluded these read lncRNA-based ORFs from all lncRNA genes of GENCODE v19. As a result, we were able to have 20 869 Ensembl transcripts of lncRNAs. In summary, we have a total of 23 587 coding and 23 587 lncRNA Ensembl transcripts as the positive and negative data sets. Since overall, the length distribution of the positive and negative data, we first decided to make deepORF model using the genes whose sequence length is less than 3k. Then, in total, there were 20 243 coding and 19 971 lncRNA transcripts as the positive and negative data sets, respectively. Using the divided data from the mixture of the positive and negative data into 70% and 30%, we trained and test deepORF, respectively.

We initially adopted the model design of RNAsamba (20), which is composed of two different sources of the whole sequence branch and the longest ORF branch. The usage of the IGLOO architecture to learn from sequence data and the integration of the whole transcript and ORF-derived information into a single coding score. By using IGLOO layers, RNAsamba could learn non-intuitive coding patterns. Based on the same model structure, we trained our model with a different training dataset, which is experimentally validated coding genes' transcript sequences as the positive data and the non-coding genes' transcript sequences never read by ribosomes. We believe this is a more reliable training data set. As a result, the initial version of deepORF (AUC: 0.94) showed better performance than RNAsamba (AUC: 0.79). In this study, we mainly adopted the existing model structure but enhanced the data quality of the training, which made better performance. This is to fill the gap of blank knowledge on the coding potential of known fusion genes as soon as possible, but with a better model. In the future, we will make a more unique and efficient algorithm by incorporating the technologies that are used in natural language process studies.

Retention analysis of 39 protein features from UniProt

We first downloaded the GFF (General Feature Format) format protein information of 10 651 UniProt accessions from UniProt for total of 10 619 genes involved in 15 030 fusion genes (21). UniProt provides the loci 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 loci information was based on amino acid sequence, the genomic breakpoint information was converted into the amino acid level while considering all UniProt protein accessions, ENST isoforms, and multiple breakpoints for each partner. To map each feature to the human genome sequence, we used the GENCODE gene model of human reference genome v19 (22). For the 5′-partner gene, we considered the protein feature to be retained in the fusion gene if the breakpoints occurred on the 3′-end of the protein feature. On the contrary, if a protein domain was not included completely in the fusion amino acid sequence, we reported that such fusion genes did not retain that protein feature. Similarly, for the 3′-partner gene, we considered the fusion gene to have retained the protein feature if the breakpoints occurred on the 5′-end of the protein feature region.

Creating fusion transcript and fusion amino acid sequences

Two different genes can form fusion genes with multiple breakpoints based on multiple gene isoforms. Therefore, we considered all gene isoforms at each breakpoint. To help for the identification and validation of fusion genes, we focused on the in-frame fusion genes. For more reliable fusion genes, we checked the distance between the two breakpoints in case of intra-chromosomal rearrangements and created fusion sequences when those genes are apart more than 100kb. We also selected fusion genes when both of their breakpoints are aligned at the exon junction. To call each exon sequence of given breakpoint, transcription start/end sites, and CDS start/end sites, we used the nibFrag utility from UCSC Genome Browser based on ENCODE hg19 genome structure. After filtering, we have created 83 290 fusion transcript sequences of 16 146 in-frame fusion genes. For these fusion transcript sequences, if it has the ORF annotation results from ORFfinder, then we adopted the start and end position for the in-frame ORF and translated amino acid sequence. As a result, we have 42 110 fusion amino-acid sequences in total.

Protein-protein interaction information

We downloaded interactor information from BIOGRID (v 3.4.260) to provide the PPI information to the wild-type protein of each fusion partner (23). There was a limitation of this dataset like providing the interactor name only. Since we need to know the loci information of each PPI to investigate the retention of the PPI at the fusion protein level, we recognized that the 'Region' feature, which is one of 39 protein features provided by UniProt, included the start and end information of interaction on each protein. Therefore, we followed the same method for the protein feature retention screening in this aim. During the process of protein feature retention search, we also checked whether these interaction loci are retained or not at fusion protein.

Table 1. Statistical comparison of FusionGDB and FusionGDB 2.0

	# fusion genes	# in-frame fusion genes	# all partner genes	# reviewed UniProt accessions of all partner genes
FusionGDB	43 895	9859	14 910	14 943
FusionGDB 2.0	102 645	16 146	26 688	17 300

Table 2. Number of fusion genes per gene group

Gene group	# in-frame fusion genes	# frame-shift fusion genes
Cancer gene census	2747	2862
Cell metabolism	4136	4342
Epigenetic factor	2076	2309
Essential gene	11 628	12 708
IUPHAR	4871	5325
Kinase	1430	1582
Trascription factor	2002	2347
Tumor suppressors	2513	2694

Functional or gene category assignment

To assign the functional or gene categories, we integrated cancer genes, tumor suppressors, epigenetic regulators, DNA damage repair genes, human essential genes, kinases and transcription factors. The first four genes were downloaded from a gene selection resource for cancer genome projects (CancerGenes) (24), an updated literature-based knowledgebase for tumor suppressor genes (TSGene) (25), a comprehensive database of human epigenetic factors and complexes (EpiFactors) (24,26), Knijnenburg et al. study about the genomic and molecular landscape of DNA damage repair deficiency across TCGA data (27), an update of the Database of Essential Genes that includes built-in analysis tools (DEG 15) (28). For these gene groups, we checked the retention and ORFs of the main protein functional features including 13 features belong to the region, such as 'calcium binding', 'coiled coil', 'compositional bias', 'DNA binding', 'domain', 'intramembrane', 'motif', 'nucleotide binding', 'region', 'repeat', 'topological domain', 'transmembrane' and 'zinc finger'.

Drug and disease information

Drug-target interactions (DTIs) were extracted from DrugBank (January 2021, version 5.1.8) with the duplicated DTI pairs excluded (29). All drugs were grouped using Anatomical Therapeutic Chemical (ATC) classification system codes. Disease-gene information was extracted from a database of gene–disease associations (DisGeNet, July 2020, version 7.0) (30).

Manual curation of PubMed articles

For the 551 highly recurrent fusion genes, which are expressed in more than five samples or cells, a literature query of PubMed was performed in June 2021 using the search expression that applied to each fusion gene. Using *BCR-ABL1* as an example, it is '((*BCR* [Title/Abstract]) AND

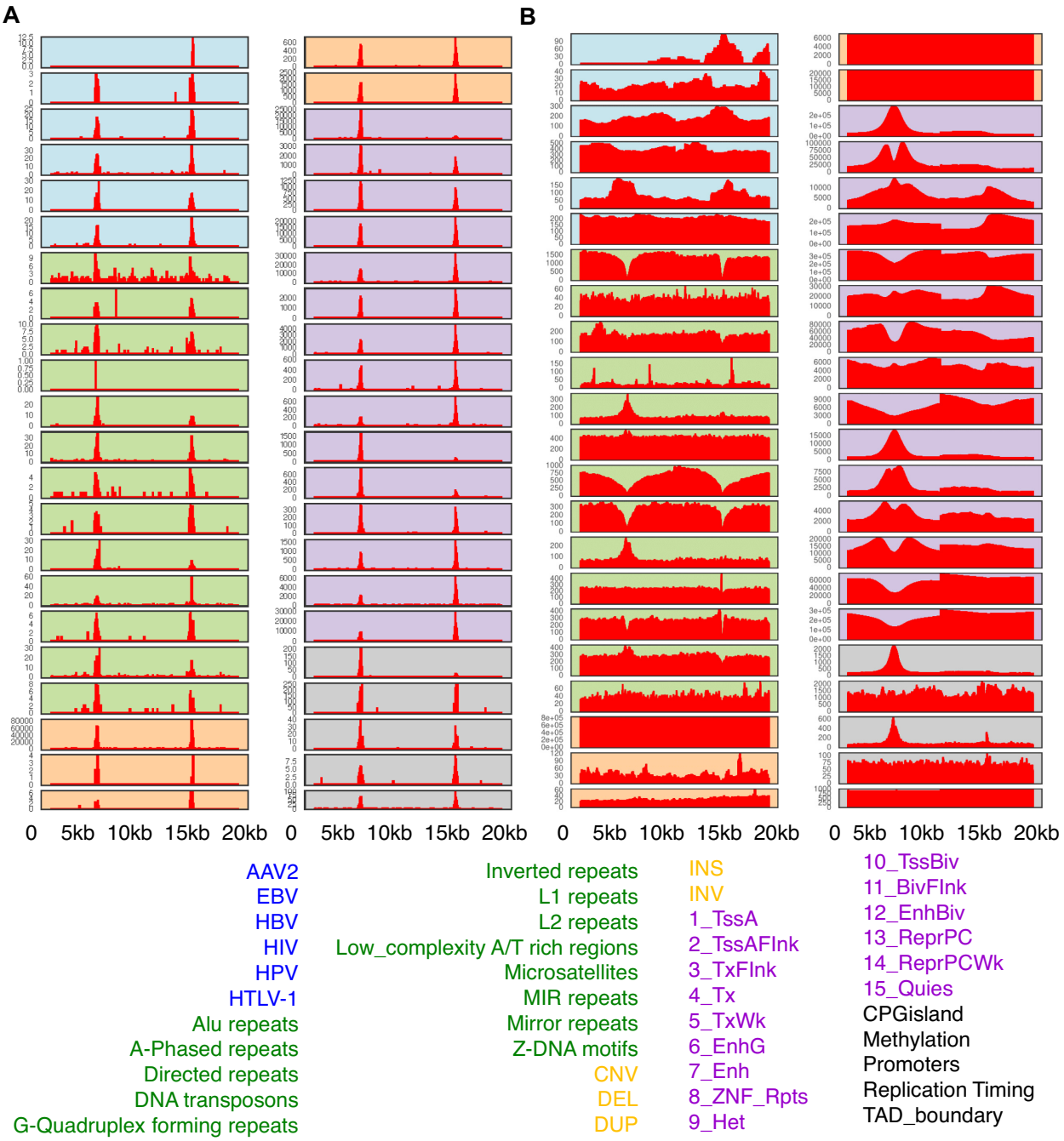

Figure 2. Feature importance (FI) score distributions across fusion breakpoint sequence of 20 Kbp length from FusionAI. (**A**) Distribution of overlaps between top 10% FI scored regions and 44 different types of human genomic features. (**B**) Distribution of overlaps between all regions and 44 different types of human genomic features. (Individual background corresponds to categories with the same colored font among 44 human genomic features).

ABL1 [Title/Abstract]) AND fusion [Title/Abstract])'. After a manual review of the abstracts, we found 108 fusion genes had literature evidence that support these fusion genes.

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 MySQL 3.23 with the MyISAM storage engine.

UPDATED WEB INTERFACE AND ANALYSIS RESULTS

Fusion gene information category (Fusion Gene Summary)

This category provides five types of information on a fusion gene including basic gene information, gene assessment

Table 3. Number of fusion genes per ORF types

ORF of fusion transcript	# fusion genes
3UTR-3UTR	3035
3UTR-5UTR	2311
3UTR-CDS	5499
3UTR-intron	8577
5UTR-3UTR	2204
5UTR-5UTR	5132
5UTR-CDS	9865
5UTR-intron	8781
CDS-3UTR	6890
CDS-5UTR	11 944
CDS-intron	28 940
Frame-shift	17 710
In-frame	16 146
intron-3UTR	13 411
intron-5UTR	12 883
intron-CDS	27 931
intron-intron	48 685

score in pan-cancer, study context, most frequent breakpoint information, and anticipated lost major functional domains due to fusion event. In this updated version, we have 102 647 fusion genes from two representative fusion gene resources of ChiTaRs 5.0 (5) and ChimerDB 4.0 (6) (Supplementary Table S1). Total, 26 688 genes with 17 300 UniProt accessions were involved in these human fusion genes (Table 1). The gene assessment scores that are showing in this section are the Degree of Frequency (DoF) score and Major Active Isofusion Index (MAII) score from previous studies (31,32). The study context of a fusion gene is showing PubMed article search results for the fusion genes present in at least five samples. To provide the most reliable and representative breakpoints, in this updated version, this category showed the most frequent breakpoint information per fusion gene. Last, we list the potentially affected major protein functional domains from specific gene groups including, kinase, transcription factor, cancer gene census, tumor suppressor, IUPHAR drug targets, cell metabolism genes, human essential genes, and epigenetic factors by investigation of retention and ORFs (between in-frame and frame-shift) of 13 protein features (See methods). As shown in Table 2, there were 1430 kinase fusion genes that retained their kinase domain but 12 708 essential gene fusion genes that lost their major functional domains due to the fusion event. Besides, this category shows the distribution of all breakpoints across two genes involved in the fusion gene with the UCSC genome browser. These are all unique and essential to understand the potential function/effect of fusion genes.

Fusion genomic feature information category (fusion gene genomic feature)

In this section, we sought to identify the genomic features of the fusion gene breakpoint area across the human genome sequences. We ran FusionAI by inputting all in-frame fusion genes that have both breakpoints located in the exon junction boundaries. The input sequence data from FusionGDB 2.0's fusion genes for FusionAI and the output scores can be accessed from the download page of the website (https://compbio.uth.edu/FusionGDB2/tables/TableS5.txt and https://compbio.uth.edu/FusionGDB2/tables/TableS6.txt). We investigated the feature importance scores across the 20 kb length fusion input sequence (Supplementary Table S7). Overall, the top 10% feature importance scored regions were enriched in near to the breakpoints among 20 kb sequence as shown in Supplementary Figure S1, which is the distribution of median values of the top 10% FI scores per nucleotide across 20 kb sequence. Next, we integrated 44 different human genomic features belong five important cellular mechanism categories such as integration site category of 6 viruses, 13 types of repeat category, 5 types of structural variant category, 15 different types of chromatin state category and 5 gene expression regulatory category to have the landscape of genomic features in the fusion breakpoint area (see Materials and Methods). For individual features of these five categories, we counted the unique number of the overlap between feature loci with the top 10% FI scored regions in every nucleotide across 20k sequence of all fusion genes (Figure 2A). The overall distribution of overlaps was enriched in the fusion gene breakpoint area.

Furthermore, we counted the unique number of the overlapped loci of the individual features with all regions of the 20 kb breakpoint sequence to see without potential confounding factors in the genomes (Figure 2B). From this distribution, we identified several genomic features that showed different distribution around the breakpoint area. In the repeat category (green background), two repeats like G-Quadruplex forming repeats and low complexity A/T rich regions were increased to the breakpoint area. On the other hand, Alu, L1 and L2 repeats were decreased to the breakpoint area. In the chromatin state category (purple background), 1_TssA and 10_TssBiv chromatin states showed increased distribution to the breakpoints. Those states represent active TSS and bivalent/poised TSS. In the expression regulation category (gray background), CpG island and promoter regions showed increased distribution around the breakpoints. Both plots in Figure 2 were drawn for all fusion genes' average values. However, on the website, we provide these two plots for individual fusion genes, separately. From these plots, the users can infer what genomic features are associated with individual fusion gene breakpoints. Overall, the distance between the top 10% feature importance scored regions and breakpoints was 70 nt of the median, 99.54 nt of mean with 211.28 nt of standard deviations (SD) as shown in Supplementary Figure S1.

Fusion gene ORF annotation category (fusion gene ORF)

This category provides three types of annotations of open reading frames of fusion genes of individual breakpoints. First, using in-house ORF annotation codes (see Methods), we annotate the ORF of fusion transcript at given breakpoints and transcript isoforms. From this result, we found 16 146 and 17 710 in-frame and frame-shift fusion genes (https://compbio.uth.edu/FusionGDB2/tables/TableS2.txt). The statistics of diverse ORFs of fusion genes are summarized in Table 3. For 16 146 in-frame fusion genes, we made full-length transcript sequences considering multiple gene isoforms and multiple breakpoints in the individual partner genes. There were total of 83 291 fusion tran-

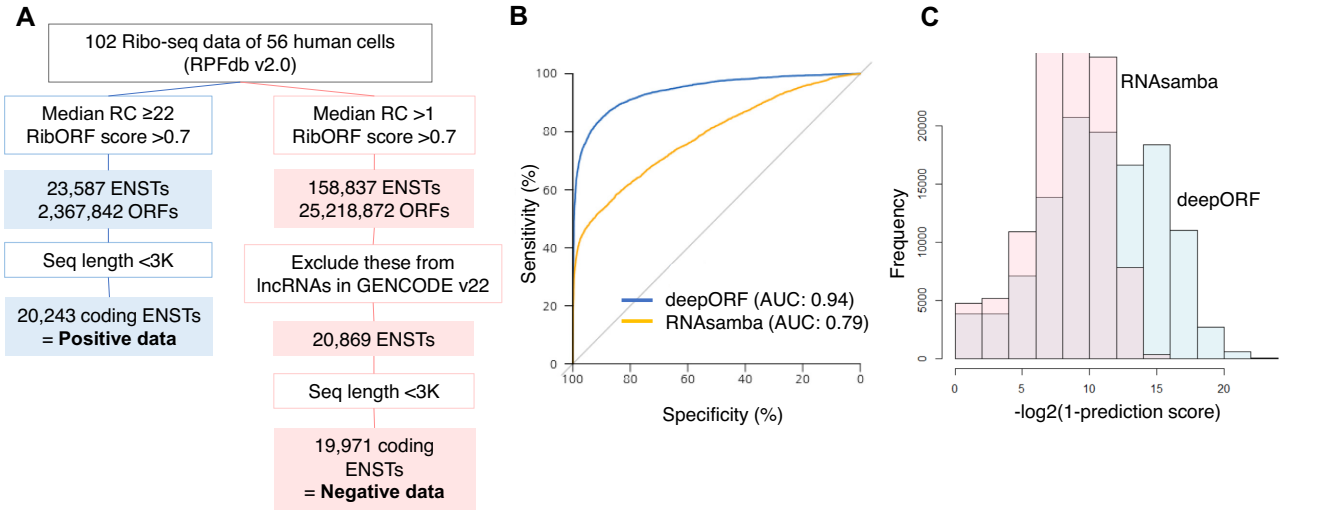

Figure 3. Prediction of the coding potential of fusion transcript with the deep learning approach. (**A**) Pipeline for creating the training data of deepORF. (**B**) Performance comparison between deepORF and RNAsamba. (**C**) Comparison of the distribution of the coding potential scores for in-frame fusion transcripts between deepORF data-based model (blue) and RAsamba (pink).

scripts. Then, we input these fusion transcript sequences to the ORFfinder. Then, ORFfinder resulted in 42 110 fusion amino acid sequences of 14 569 fusion genes. For these, we made the fusion amino acid sequences, which are available from the Fusion Gene Sequence category (https://compbio.uth.edu/FusionGDB2/tables/TableS3.txt). Last, we used a deep learning classifier to distinguish coding vs non-coding transcripts (see Methods), named as deepORF. For the deep learning model design, we adopted it from RNAsamba, a neural network-based assessment of the protein-coding potential of RNA sequences. By retraining the model using high-quality training data set as shown in the pipeline in Figure 3A, we could have dramatically increased the performance of the prediction of coding potentials (Figure 3B). We will enhance this model by creating a unique model design in the future. With the current model, we predicted the coding potential of individual in-frame fusion transcript sequences in FusionGDB 2.0 (Supplementary Table S4). The prediction scores' distributions using RNAsamba and deepORF are shown in Figure 3C.

Fusion protein feature information category (fusion protein Feature)

In this category, we provide the detailed annotation of fusion protein functions through the retention search of 39 protein features from UniProt based on the fusion protein sequence. For individual partners in fusion genes, we checked the retention of 39 protein features (https://compbio.uth.edu/FusionGDB2/tables/TableS8.txt and https://compbio.uth.edu/FusionGDB2/tables/TableS9.txt). Specifically, in this category, we are showing the results of the retention studies of 13 features belong to the region subsection including 'calcium binding', 'coiled coil', 'compositional bias', 'DNA binding', 'domain', 'intramembrane', 'motif', 'nucleotide binding', 'region', 'repeat', 'topological domain', 'transmembrane' and 'zinc finger', to focus on the major functional component due to the limited web page long. Most of all, in this updated version, we added a link to FGviewer (https://ccsmweb.uth.edu/FGviewer), a tool for visualizing functional features of the human fusion genes. FGviewer takes the genomic coordinate of the fusion gene breakpoints as the input data. The same breakpoint lines across four tiers across DNA, RNA, protein and pathogenic levels will classify between fusion involved or deleted zones with multiple types of functional features. Those features include fusion mRNA and amino acid sequences based on the user's breakpoint coordinates, swapped gene expression regulatory (i.e. transcription factor or miRNA binding sites), protein functional features (i.e., protein domains, protein-protein interactions, binding sites of all molecules, secondary structure level feature, etc.), clinically relevant variants, and etc. Through these annotations, users can easily infer the possible roles of fusion genes in tumorigenesis.

DISCUSSION AND FUTURE DIRECTION

FusionGDB 2.0 is the substantially updated version of FusionGDB that served as the unique and widely used systematic annotations of the functions of human fusion genes. FusionGDB 2.0 is the first database that systematically annotates the function of fusion genes across pan-cancer aided by deep learning approaches. From the genomic feature landscape of individual fusion genes, scientists can have novel insights into the breakage of the genome and forming fusion genes. Further studies may be able to identify novel genomic features in the specific fusion gene groups. From the advanced ORF annotations, the researchers can have the coding potential fusion genes in human cancer. From the protein feature study results, we identified 1430, 2002 and 2747 in-frame fusion genes retaining major functional domains of kinase, transcription factors, and cancer genes. We also identified 4342, 2309, 5325 and 2694 frame-shifted fusion genes that lost the major functional domains of cell metabolism genes, epigenetic

factors, IUPHAR drug target genes and tumor suppressors. These candidates can be novel sources of new drug development in cancer. From the fusion sequence category, users can download the full-length fusion transcript sequence and predicted fusion amino acid sequences.

To serve broad biomedical research communities, we will continuously update and curate human fusion genes routinely by checking new fusion gene or fusion protein data. We will investigate and extend the methods to find the clinically important fusion genes and affected target genes in the downstream part in the future. We believe this updated version will be more widely and actively visited/used in diverse human genetic diseases including cancers for a better understanding of pathogenesis, progression, cancer biology, and identification of drug-targetable features of specific fusion genes.

DATA AVAILABILITY

All annotation results are available from the FusionGDB 2.0 website (https://compbio.uth.edu/FusionGDB2) for academic purpose only. Further information and requests should be directed to Dr. Pora Kim (Pora.kim@uth.tmc.edu).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the users of FusionGDB for valuable questions, discussions and suggestions. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

FUNDING

National Institutes of Health [R35GM138184 to P.K. Funding for open access charge: Startup Fund to Dr Kim from the University of Texas Health Science Center at Houston.
Conflict of interest statement. None declared.

REFERENCES

1. Kim,P. and Zhou,X. (2019) FusionGDB: fusion gene annotation DataBase. *Nucleic Acids Res.*, **47**, D994–D1004.
2. Kim,P., Tan,H., Liu,J., Yang,M. and Zhou,X. (2021) FusionAI: predicting fusion breakpoint from DNA sequence with deep learning. *iScience*, **24**, 103164.
3. Kim,P., Yiya,K. and Zhou,X. (2020) FGviewer: an online visualization tool for functional features of human fusion genes. *Nucleic Acids Res.*, **48**, W313–W320.
4. Coordinators,N.R. (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **46**, D8–D13.
5. Balamurali,D., Gorohovski,A., Detroja,R., Palande,V., Raviv-Shay,D. and Frenkel-Morgenstern,M. (2020) ChiTaRS 5.0: the comprehensive database of chimeric transcripts matched with druggable fusions and 3D chromatin maps. *Nucleic Acids Res.*, **48**, D825–D834.
6. Jang,Y.E., Jang,I., Kim,S., Cho,S., Kim,D., Kim,K., Kim,J., Hwang,J., Kang,J., Lee,B. *et al.* (2020) ChimerDB 4.0: an updated and expanded database of fusion genes. *Nucleic Acids Res.*, **48**, D817–D824.
7. Navarro Gonzalez,J., Zweig,A.S., Speir,M.L., Schmelter,D., Rosenbloom,K.R., Raney,B.J., Powell,C.C., Nassar,L.R., Maulding,N.D., Lee,C.M. *et al.* (2021) The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.*, **49**, D1046–D1057.
8. Tang,D., Li,B., Xu,T., Hu,R., Tan,D., Song,X., Jia,P. and Zhao,Z. (2019) VISDB: a manually curated database of viral integration sites in the human genome. *Nucleic Acids Res.*, **48**, D633–D641.
9. Bao,W., Kojima,K.K. and Kohany,O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*, **6**, 11.
10. Avvaru,A.K., Sharma,D., Verma,A., Mishra,R.K. and Sowpati,D.T. (2020) MSDB: a comprehensive, annotated database of microsatellites. *Nucleic Acids Res.*, **48**, D155–D159.
11. Lappalainen,I., Lopez,J., Skipper,L., Hefferon,T., Spalding,J.D., Garner,J., Chen,C., Maguire,M., Corbett,M., Zhou,G. *et al.* (2013) DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.*, **41**, D936–D941.
12. Ernst,J. and Kellis,M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.*, **12**, 2478–2492.
13. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., Ziller,M.J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
14. Lizio,M., Abugessaisa,I., Noguchi,S., Kondo,A., Hasegawa,A., Hon,C.C., de Hoon,M., Severin,J., Oki,S., Hayashizaki,Y. *et al.* (2019) Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Res.*, **47**, D752–D758.
15. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
16. Akdemir,K.C., Le,V.T., Chandran,S., Li,Y., Verhaak,R.G., Beroukhim,R., Campbell,P.J., Chin,L., Dixon,J.R., Futreal,P.A. *et al.* (2020) Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.*, **52**, 294–305.
17. Howe,K.L., Achuthan,P., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J. *et al.* (2020) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.
18. Wang,H., Yang,L., Wang,Y., Chen,L., Li,H. and Xie,Z. (2019) RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res.*, **47**, D230–D234.
19. Ji,Z. (2018) RibORF: identifying genome-wide translated open reading frames using ribosome profiling. *Curr. Protoc. Mol. Biol.*, **124**, e67.
20. Camargo,A.P., Sourkov,V., Pereira,G.A.G. and Carazzolle,M.F. (2020) RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. *NAR Genom Bioinform*, **2**, lqz024.
21. The UniProt Consortium (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699.
22. 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.
23. 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.
24. 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.
25. 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.
26. Medvedeva,Y.A., Lennartsson,A., Ehsani,R., Kulakovskiy,I.V., Vorontsov,I.E., Panahandeh,P., Khimulya,G., Kasukawa,T., Drablos,F. and FANTOM Consortium (2015) EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database*, **2015**, bav067.
27. 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 Rep.*, **23**, 239–254.
28. Luo,H., Lin,Y., Liu,T., Lai,F.L., Zhang,C.T., Gao,F. and Zhang,R. (2021) DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. *Nucleic Acids Res.*, **49**, D677–D686.
29. 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.
30. 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.
31. 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.
32. 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.