Published online 4 November 2019 *Nucleic Acids Research, 2020, Vol. 48, Database issue* ***D817–D824***
doi: 10.1093/nar/gkz1013

ChimerDB 4.0: an updated and expanded database of fusion genes

Ye Eun Jang[1,†], Insu Jang[2,†], Sunkyu Kim[3,†], Subin Cho[1,†], Daehan Kim[3], Keonwoo Kim[3], Jaewon Kim[1], Jimin Hwang[1], Sangok Kim[4], Jaesang Kim[4], Jaewoo Kang[3], Byungwook Lee[2,*] and Sanghyuk Lee[1,4,*]

[1]Department of Bio-Information Science, Ewha Womans University, Seoul 03760, Republic of Korea, [2]Korean Bioinformation Center, Korean Research Institute of Bioscience and Biotechnology, Daejeon 34141, Republic of Korea, [3]Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea and [4]Department of Life Science, Ewha Womans University, Seoul 03760, Republic of Korea

Received September 15, 2019; Revised October 16, 2019; Editorial Decision October 16, 2019; Accepted October 17, 2019

ABSTRACT

Fusion genes represent an important class of biomarkers and therapeutic targets in cancer. ChimerDB is a comprehensive database of fusion genes encompassing analysis of deep sequencing data (ChimerSeq) and text mining of publications (ChimerPub) with extensive manual annotations (ChimerKB). In this update, we present all three modules substantially enhanced by incorporating the recent flood of deep sequencing data and related publications. ChimerSeq now covers all 10 565 patients in the TCGA project, with compilation of computational results from two reliable programs of STAR-Fusion and FusionScan with several public resources. In sum, ChimerSeq includes 65 945 fusion candidates, 21 106 of which were predicted by multiple programs (ChimerSeq-Plus). ChimerPub has been upgraded by applying a deep learning method for text mining followed by extensive manual curation, which yielded 1257 fusion genes including 777 cases with experimental supports (ChimerPub-Plus). ChimerKB includes 1597 fusion genes with publication support, experimental evidences and breakpoint information. Importantly, we implemented several new features to aid estimation of functional significance, including the fusion structure viewer with domain information, gene expression plot of fusion positive versus negative patients and a STRING network viewer. The user interface also was greatly enhanced by applying responsive web design. ChimerDB 4.0 is available at http://www.kobic.re.kr/chimerdb/.

INTRODUCTION

Fusion genes continue to serve as an important source of biomarkers and therapeutic targets in various types of cancer. Since the groundbreaking discovery of *BCR–ABL1* fusion gene in leukemia, numerous driver fusion alterations have been identified as druggable targets, including genes such as *TMPRSS2*, *ALK*, *RET*, *FGFR3*, *ROS1* and *ESR1*, leading to development of targeted therapies (1,2). Additionally, many fusion genes function as biomarkers for specific cancer types as can be seen in the examples of *DNAJB1–PRKACA* fusion in fibrolamella carcinoma (3) and *EWSR1–FLI1* in Ewing's sarcoma (4). Furthermore, a number of fusion genes have been identified as prognostic markers with biological roles. For example, fusion events in metastatic ER-positive breast cancer are more frequent than in primary cases, suggesting fusions as biomarkers of advanced and aggressive disease (5). Thus, fast and reliable identification of fusion genes is increasingly relevant for clinical and pharmaceutical applications.

Since the last update of ChimerDB 3.0 (6), an enormous amount of RNA-Seq data, the major source of mining fusion transcripts, has been released in public. The TCGA dataset represents the largest collection including 13 786 tumor samples in 33 cancer types, which were analyzed by the following two groups independently. Verhaak and colleagues built the TumorFusions database (7) that identified 20 731 gene fusions from 9966 tumor samples and 648 normal specimens in the TCGA database applying their own computational pipeline PRADA (8). The Fusion Analysis Working Group (FAWG) of the TCGA research network investigated 9624 tumors using multiple fusion calling tools and identified 25 664 'reliable' fusion events (9).

Reflecting the importance of fusion genes in cancer, numerous algorithms and databases have been developed to

*To whom correspondence should be addressed. Tel: +82 23 277 2888; Fax: +82 23 277 6809; Email: sanghyuk@ewha.ac.kr
Correspondence may also be addressed to Byungwook Lee. Tel: +82 42 879 8531; Fax: +82 42 879 8519; Email: bulee@kribb.re.kr
†The authors wish it to be known that, in their opinion, the first four authors should be regarded as Joint First Authors.

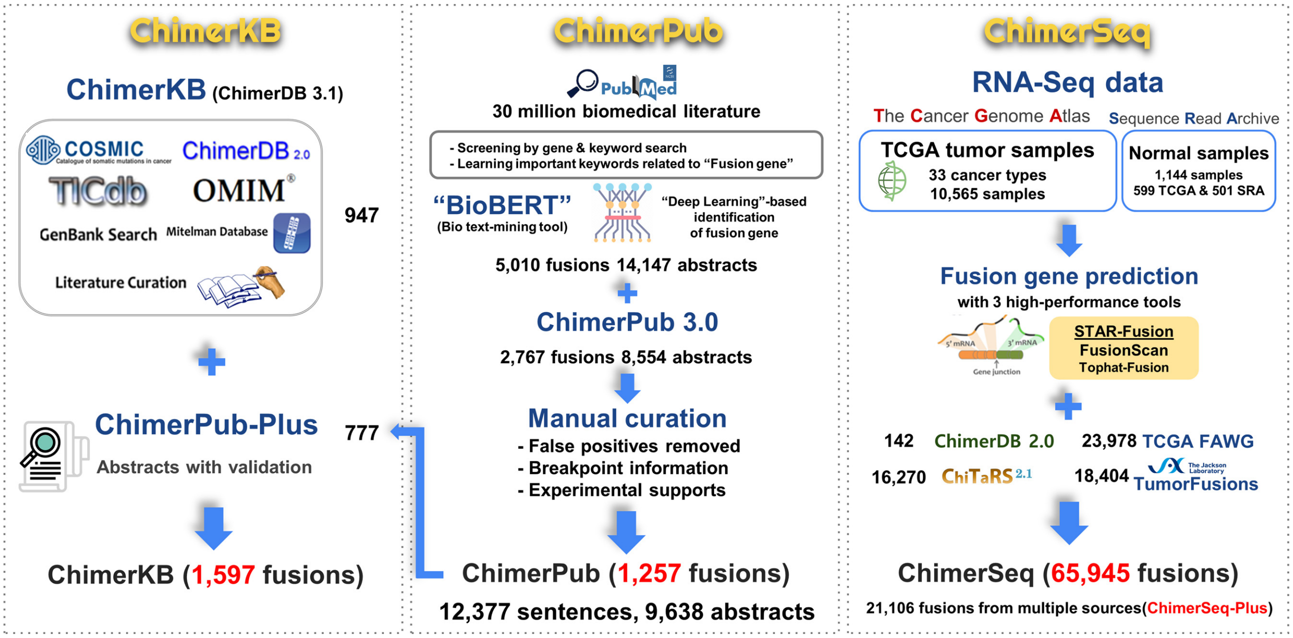


Figure 1. Overview of ChimerDB 4.0. Each number indicates the number of unique gene pairs from the relevant resources.

predict and catalog the fusion genes. Li and colleagues carried out comparative performance test for 12 public programs (10). Several more recent programs such as GFusion (11), FusionScan (12) and STAR-Fusion (13) claimed to have achieved higher sensitivity with less false positives. In addition, with so many fusion candidates predicted from transcriptome data, it is also critical to rapidly assess their reliability, functional significance and biological roles. Thus, data aggregation and functional annotation are necessary, ideally with powerful visualization support. INTEGRATE-Vis is a comprehensive visualization tool for gene fusion events (14). FusionGDB provides extensive functional annotations for fusion events aggregated from public resources such as TumorFusions (7), TCGA FAWG (9) and ChiTaRS 3.1 (15). More recently, FusionHub introduced an integrated web platform that supports both annotation and visualization for the largest collection of fusion gene datasets aggregated from 24 resources (16).

The previous version ChimerDB 3.0 was a unique effort to provide comprehensive pictures of fusion genes consisting of several modules with different purposes—ChimerKB as a knowledgebase with extensive manual annotation, ChimerPub as a text-mining utility for identifying fusion genes from PubMed abstracts and ChimerSeq to aggregate the prediction results by analyzing transcriptome sequencing data. To the best of our knowledge, ChimerPub was the first attempt of using text-mining technique to catalog the fusion gene events from the literature. In this update, we describe improvements for each of the three modules and new tools implemented to aid evaluating functional significance of fusion genes. Unlike other resources described earlier, ChimerDB provides a wide range of information encompassing known fusion genes and literature reports as well as candidates from deep sequencing data. ChimerDB would continue to be a comprehensive up-to-date database of fusion genes.

SYSTEM UPDATE AND METHODS

The overall procedure and contents of ChimerDB 4.0 are summarized in Figure 1. The tripartite module design remains unchanged, but each module was substantially improved. We have developed a new 'deep learning'-based text-mining method, which was applied to PubMed abstracts to identify 728 new fusion gene entries in ChimerPub update. Importantly, we manually scrutinized the full text of candidate-reporting papers to annotate the fusion breakpoints and experimental evidences as well as to remove false positives. ChimerSeq re-analyzed the whole TCGA transcriptome data using recently developed reliable programs and combined the results with those of TumorFusions and the TCGA FAWG. The number of patients became more than twice the previous version. ChimerKB was updated to reflect new authentic fusion genes identified from ChimerPub update after extensive manual curation. Methods for building ChimerSeq and ChimerPub are described later with further details provided in the Supplementary Information.

ChimerSeq module

ChimerSeq module analyzed RNA-Seq data available in public and aggregated the results from other databases or computation results. ChimerDB 3.0 covered 4569 tumor samples in 23 cancer types in the TCGA project. In this update, we re-analyzed 10 565 tumor samples across 33 cancer types downloaded from the GDC data portal of NCI. We have also analyzed 1144 normal samples from the TCGA and SRA (E-MTAB-2836, E-GEUV-1, GSE122401) archives to filter out germline fusions.

Our main objective was to reduce the number of false positives, which would lead to unnecessary efforts in the validation procedure. We used two high-performance fusion detection tools to analyze all TCGA RNA-Seq data—FusionScan (12) and STAR-Fusion (13). STAR-Fusion was added because of its high precision and fast computation. We kept the fusion candidates with ≥2 junction reads or with 1 junction read and ≥2 spanning reads. Fusions from the same gene family or from the paralogous genes were removed because of uncertainties in read alignments. We also filtered out fusion genes of germline origin that were observed in the pool of 1144 normal samples. Of note, however, several well-known fusion genes such as TMPRSS2–ERG fusion were identified in a few normal samples from the TCGA cohort. Thus, we rescued such fusion cases present in ChimerKB with literature evidence. The result was merged with two other public resources that analyzed the same dataset (TumorFusions and the TCGA FAWG). Additionally, we integrated the results from TopHat-Fusion prediction (17), EST and mRNA analysis from ChimerDB 2.0 and ChiTaRS 2.1 as included in ChimerDB 3.0. Most TCGA samples were analyzed by five independent programs or pipelines. We collected fusion cases called by multiple programs as the 'ChimerSeq-Plus' group representing a reliable list of fusion gene candidates in cancer.

Functional annotation for fusion genes is important in assessing their significance in cancer. We amassed the gene expression and copy number data of the TCGA samples from the UCSC Xena (https://xena.ucsc.edu). Gene lists for functional classes included kinases in the human kinome database (December 2007 update) (18), oncogenes in ONGene (19) and tumor suppressor genes in TSGene 2.0 (20).

ChimerPub module

ChimerPub was introduced in ChimerDB 3.0 to extract fusion-related sentences from PubMed abstracts semi-automatically. Initial screening of candidate sentences was based on identifying two gene names joined by a delimiter (e.g. BCR–ABL or BCR/ABL). We subsequently built an elaborate classifier model using feature selection and logistic regression methods, which resulted in 2767 fusion genes from 10 580 sentences.

In this update, we built a new 'deep learning'-based model to identify fusion relations that are not limited to two gene names joined by delimiters in sentences. Thus, the model can extract fusion relations from sentences in natural language forms without any hand-crafted features. Our model was built on BioBERT (https://github.com/dmis-lab/biobert), which is a 'deep learning'-based language model that showed state-of-the-art performance in representative biomedical text-mining tasks such as named entity recognition, relation extraction and question answering. We stacked one prediction layer on the last layer of the pre-trained BioBERT model.

The training data for the prediction layer were obtained from abstracts in ChimerKB 3.0 by extracting sentences with two gene names using BERN (21), a biomedical entity recognition tool. Resulting sentences were classified into positive and negative datasets based on whether the candidate fusion relation existed in ChimerKB or not. To compare the performance of this new tool with the old version, we divided cases into two classes according to the presence of fusion delimiters. We obtained 1295 positive and 5333 negative sentences as the training dataset for the fusion sentences with delimiters, and 379 positive and 96 993 negative sentences for the fusion sentences without delimiters. We also prepared the test datasets by random selection of candidate sentences followed by manual curation (251 positives and 256 negatives for the class with delimiters, and 230 positives and 266 negatives for the class without delimiters).

For sentences with delimiters, both methods achieved excellent performance with high precision of 0.98 and the recall rate of 0.95 in ChimerPub 3.0 and 0.98 in ChimerPub 4.0. For sentences without fusion delimiters, ChimerPub 4.0 achieved the precision of 0.901 and the recall rate of 0.909, which is still excellent although slightly worse than the results for sentences with delimiters. Thus, ChimerPub 4.0 now supports high-performance text mining of PubMed abstracts whether the fusion delimiter symbols are present or not.

The new 'deep learning'-based model was applied to analyze ∼30 million PubMed abstracts cumulated up to November 2018. Disease terms, validation methods and translocation information were also extracted from the abstracts. We obtained >14 000 abstracts from this new algorithm, and the total number of abstracts became >17 000 after summing abstracts from ChimerPub 3.0.

Text mining inevitably accompanies many false positives; thus, manual curation is essential to increase the credibility and quality of annotations. In building ChimerPub 4.0, we decided to examine the full text, not just the abstract to annotate fusion-related information as well as to remove false positives. Articles reporting cases in ChimerKB 3.0 were excluded from the manual curation and we manually inspected the full text of remaining 2816 articles. During the process of manual curation, we utilized the HGNC gene synonyms from the BioMart service and the cell–cell interaction database of G. Bader Lab (http://baderlab.org/CellCellInteractions).

Finally, we collected the authentic cases as the 'ChimerPub-Plus' group where fusion genes were supported with experimental evidences such as Sanger sequencing, reverse transcriptase-polymerase chain reaction (RT-PCR) or fluorescence *in situ* hybridization (FISH). These ChimerPub-Plus cases were added to ChimerKB since they would meet the stringent requirements to be genuine fusion genes.

RESULTS

ChimerDB 4.0 includes 67 610 fusion gene pairs as summarized in the overall statistics (Table 1). Compared with the previous version (6), the content of ChimerKB and ChimerSeq increased by ∼50% and ∼100%, respectively. ChimerPub's entries, however, decreased from 2767 to 1257 unique fusions due to extensive manual curation with the full-text proof. The overlap among three modules is limited (Figure 2A) implying that three modules play complemen-

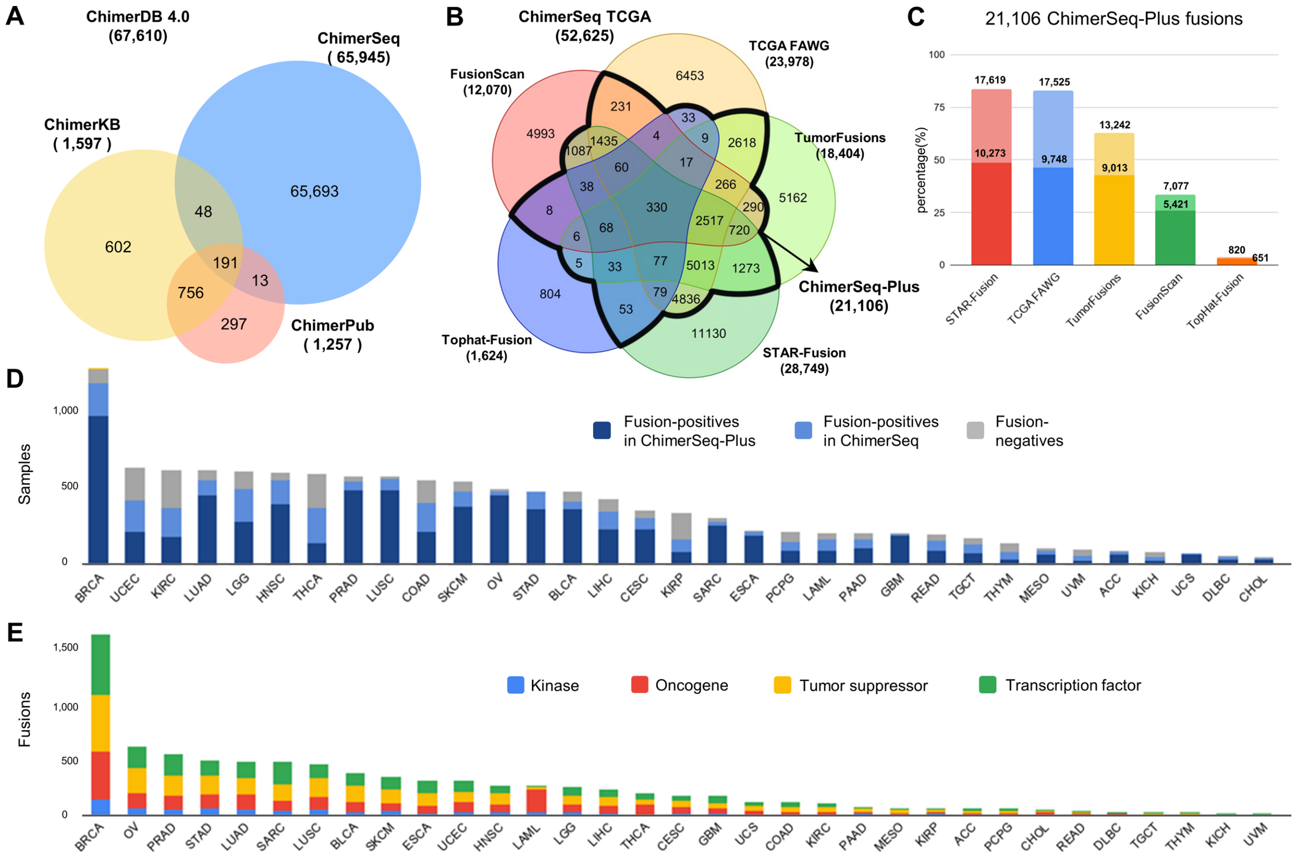


Figure 2. Statistics and contents of ChimerDB 4.0. (**A**) Venn diagram of unique fusions in three modules. (**B**) Venn diagram of unique fusions from five prediction pipelines that analyzed the TCGA dataset. (**C**) Contribution of each prediction pipeline to ChimerSeq-Plus. Dark colors indicate fusion genes that were identified by ≥3 prediction programs, whereas light colors indicate fusion genes predicted by the program of interest and one additional program. (**D**) Bar plot of TCGA samples for each cancer type. (**E**) Bar plot of fusion genes in different functional categories for each cancer type.

tary roles.

ChimerSeq was rebuilt by merging the results from five different pipelines that analyzed the whole TCGA transcriptome data (10 565 samples across 33 cancer types) (Figure 2B, Table 1). TCGA analysis yielded 49 648 fusion genes, representing ∼75% of all fusion genes in ChimerSeq. The proportion of singletons can be an indirect measure of credibility for each prediction method, which increased in the order of TCGA FAWG (27%), TumorFusions (28%), STAR-Fusion (39%), FusionScan (41%) and TopHat-Fusion (50%). Since common predictions from different programs are often regarded as more reliable fusions, we built a new 'ChimerSeq-Plus' group that contained 21 106 fusion genes supported by any two of the pipelines. We also examined the contribution of each prediction method to ChimerSeq-Plus, which decreased in the order of STAR-Fusion, TCGA FAWG, TumorFusions, FusionScan and TopHat-Fusion (Figure 2C). However, the portion of fusion genes predicted by ≥3 programs is higher in FusionScan and TopHat-Fusion, implying that their predictions were precise (i.e. low false discovery rate). The number of samples analyzed and the number of fusion genes identified for each cancer type are shown in Figure 2D.

ChimerPub was greatly enhanced by implementing a new 'deep learning'-based algorithm and by extensive manual curation of full-text articles. Our text-mining tool can identify sentences with fusion genes even when two gene are not necessarily joined by a delimiter symbol such as '–' or '/'. ChimerPub 4.0 now contains 1257 fusion genes with 728 fusions as new members (Figure 3A). We have 565 fusion genes from articles before June 2016, which should be the contribution from the new deep learning method. This number is comparable to that of the previous symbol-based method (529 fusion genes), illustrating the power of new algorithm. Moreover, a substantial portion of ChimerPub 3.0 (2238 fusion genes) was discarded after thorough curation process as described later.

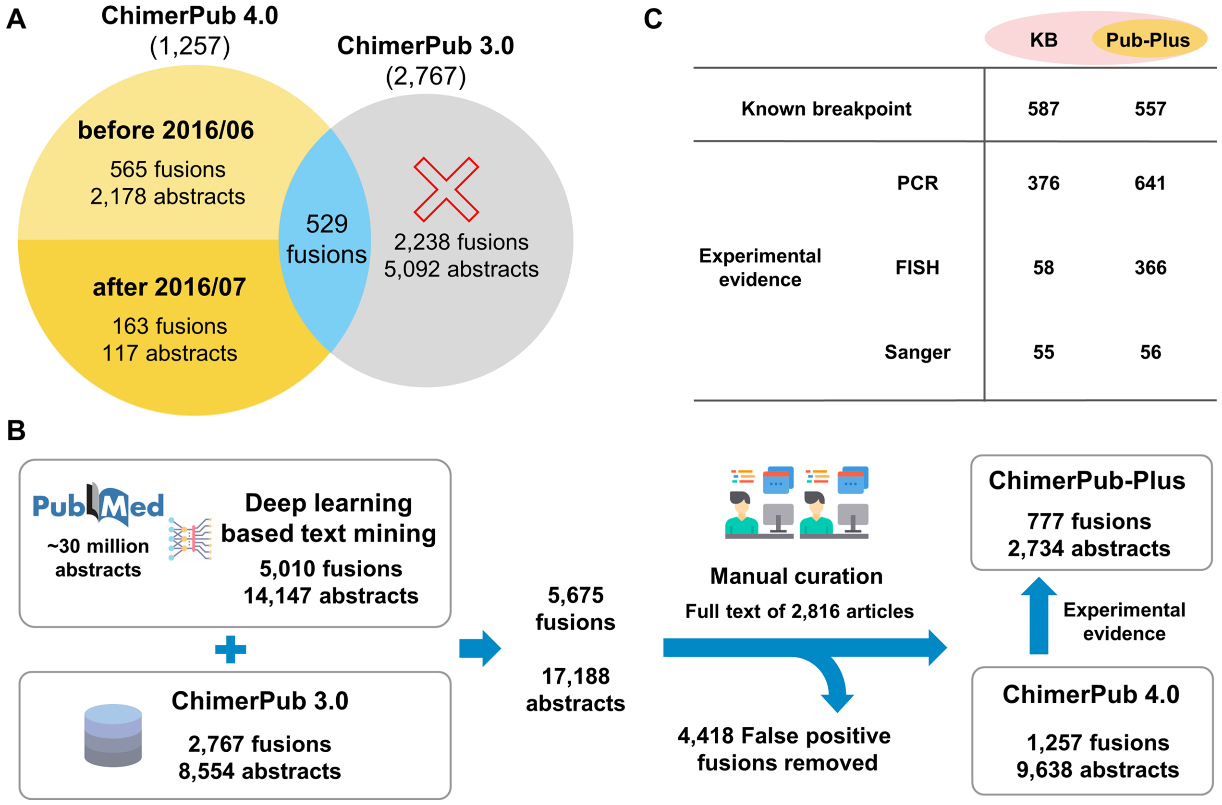
The manual curation process was extensively reinforced to enhance the quality of fusion records. We obtained 17 188 abstracts (5675 fusion genes) from text mining (Figure 3B). We found that many entries were false positives, where A/B (or A–B) meant the receptor–ligand interactions, gene–gene interactions, signaling or complex relations, or gene synonyms rather than the genuine gene fusion event. Even the gene order was reversed in some cases. The initial round of curation to remove such errors yielded 12 332 articles (2769 fusion genes). Then, articles reporting fusion genes in ChimerKB 3.0 were excluded from the manual curation, leaving 2816 articles (2182 fusion genes) for manual examination with the full-text proof. In the second round of curation, we removed further false positives reporting artificial, synthetic or nonhuman fusions. We have also annotated in-

Table 1. Statistics of ChimerDB 4.0

ChimerKB		ChimerPub		ChimerSeq	
Literature curation	147	**Information available**		**TCGA**	49 648
COSMIC	331	Translocation	925	STAR-Fusion	28 749
mRNA Sequence	272	Disease	1075	FusionScan	12 070
Mitelman, OMIM, GenBank	459	Validation method	1049	TumorFusions	18 404
ChimerPub-Plus	777			TCGA FAWG	23 978
				TopHat-Fusion	1624
				ChimerDB 2.0	142
				ChiTaRS 2.1	16 270
				Panel of **Normals**	2985
Total	**1597**	**Total**	**1257**	**Total**	**65 945**
ChimerPub supported	937	ChimerKB supported	937	ChimerKB supported	240
ChimerSeq supported	240	ChimerSeq supported	205	ChimerPub supported	205
				ChimerSeq-Plus	21 106
Known breakpoint cases				**Novel fusion**[a]	
Exon junction	1063			TCGA	52 534
				ChiTaRS	16 152

All numbers represent the number of unique fusion genes.

[a] Transcripts not included in ChimerKB and ChimerPub were classified as novel fusion.



		KB	Pub-Plus
Known breakpoint		587	557
Experimental evidence	PCR	376	641
	FISH	58	366
	Sanger	55	56

Figure 3. Statistics and contents of ChimerPub 4.0. (**A**) Comparison of ChimerPub 4.0 versus 3.0. (**B**) Curative procedure and resulting numbers at each step. (**C**) Number of ChimerKB entries with information on breakpoints and/or experimental evidences.

formation on the fusion breakpoints and set of experimental evidence from Sanger sequencing, RT-PCR and FISH.

As a result of manual curation and annotation, a substantial portion of ChimerPub entries became highly reliable with experimental evidences. Thus, we defined the 'ChimerPub-Plus' group whose fusion genes were supported by experimental evidences (Figure 3B). We identified 777 such cases and put them into ChimerKB, which greatly expanded the content by ∼50%. The information content of ChimerKB, ChimerPub-Plus is shown in Figure 3C. ChimerKB and ChimerPub now contain 1637 fusion genes with known breakpoints and 1150 fusion genes with experimental supports, which should be the largest collection of this kind.

Fusion genes in ChimerSeq were analyzed for functional roles such as kinases, oncogenes, tumor suppressors or tran-

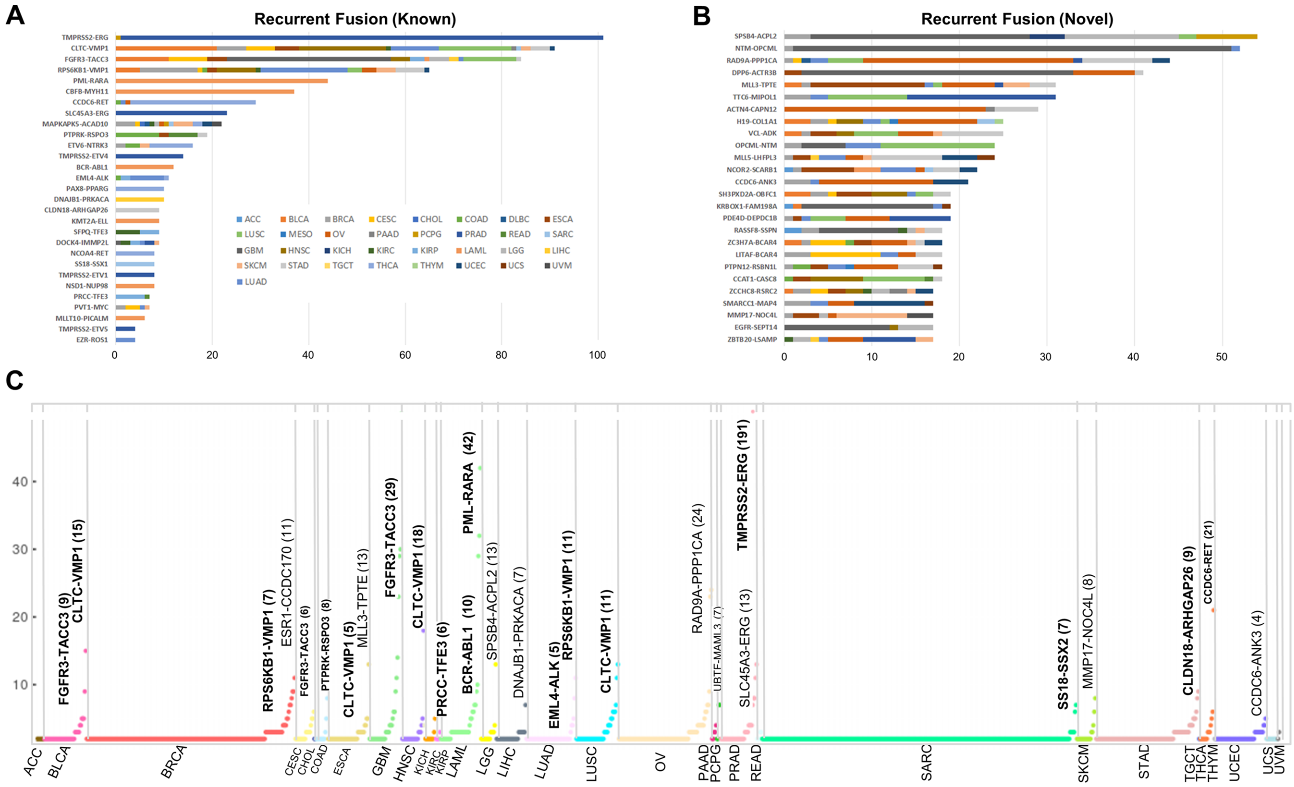


Figure 4. Recurrent fusion genes from the TCGA cohort. (**A**) Representative known fusion genes in ChimerKB and ChimerPub. (**B**) Representative novel fusion genes in ChimerSeq-Plus. (**C**) Recurrent fusion genes for each cancer type. Horizontal axes (**A–C**) indicate the number of patients with fusion genes identified in ChimerSeq-Plus.

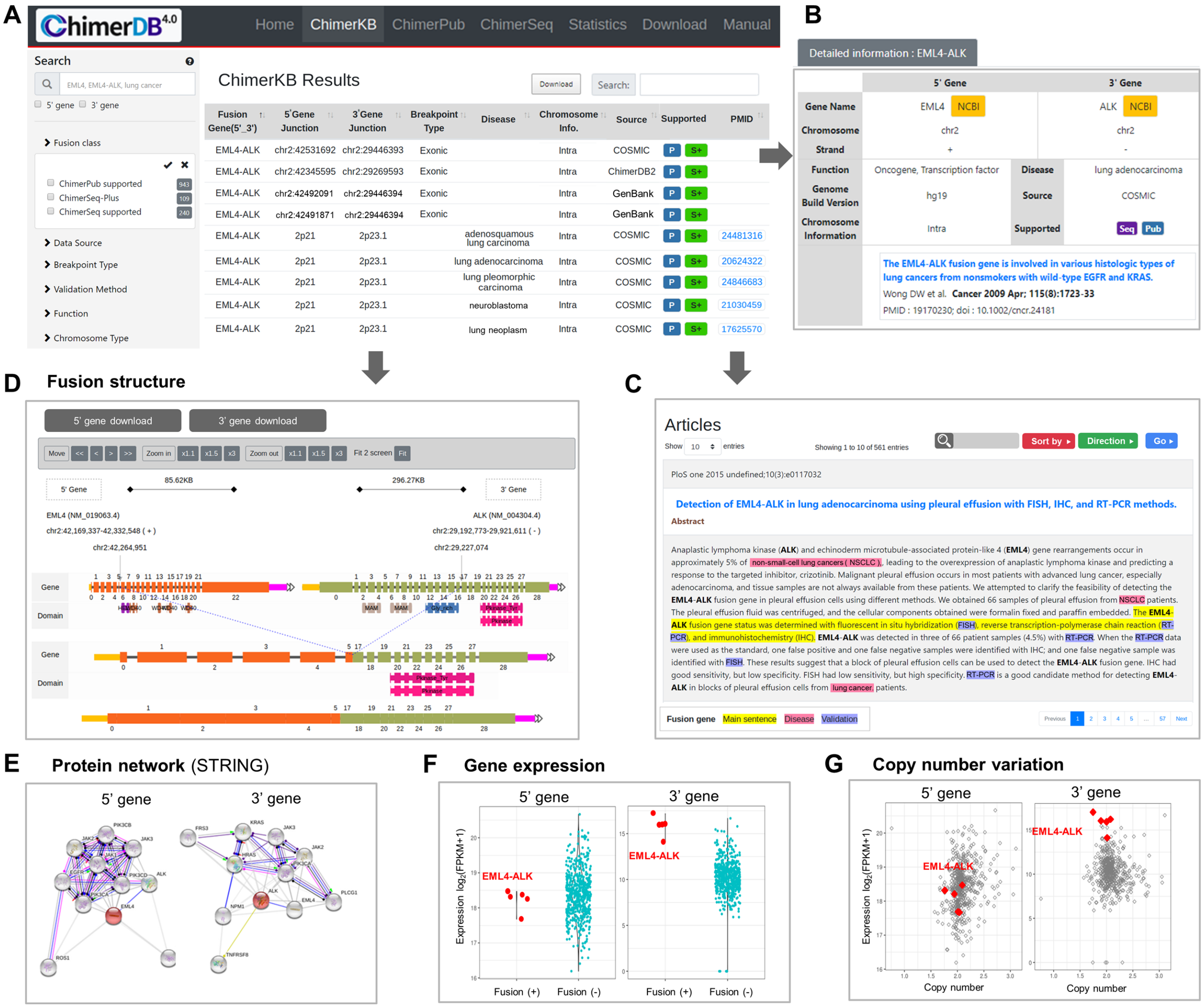
scription factors across cancer types (Figure 2E). Kinase fusion genes of in-frame are of particular interest because it can alter activity of signaling pathways. We have identified 2629 kinase fusion genes among which 1188 cases are 3′ kinases and 714 cases are in-frame. Likewise, we identified 6706 oncogene-associated, 10 538 tumor suppressor-associated and 9108 transcription factor-associated fusions. Majority of fusions are of CDS–CDS type, thus reading frame change being important to assess their functional significance. Most of known important fusion genes are in-frame fusions because frame shift necessarily leads to loss of function. We found 15 309 (23.2%) fusions as in-frame fusions, 8079 of those belonging to the ChimerSeq-Plus group.

Recurrence is the most critical property of clinically important fusion genes. Restricting our attention to fusion genes in ChimerSeq-Plus dataset, we found that 85 fusion genes from ChimerKB or ChimerPub were recurrent (Figure 4A). Additionally, we identified 1293 recurrent fusions where 61 fusions were observed in ≥10 patients (Figure 4B). Of note, novel recurrent fusion genes tend to come from diverse cancer types, highlighting the power of pan-cancer study. Sorting out recurrent fusions according to cancer types illustrates many famous fusion genes such as *TMPRSS2–ERG* in PRAD (191 patients), *PML–RARA* fusion in LAML (42 patients), *FGFR3–TACC3* in GBM (29 patients), *CCDC6–RET* in THCA (21 patients), *CLDN18–ARHGAP26* in STAD (9 patients) and *EML4–ALK* in LUAD (5 patients) (Figure 4C). Thus, mining recurrent genes from ChimerSeq can be an attractive strategy to identify novel cancer biomarkers. The full list of recurrent genes can be downloaded from the website.

USER INTERFACE

The user interface of ChimerDB adopted a responsive web design, in a similar fashion to the NCI's GDC data portal. Figure 5 shows the important features in the user interface, taking *EML4–ALK* fusion as an example query. Standard search can be done with gene names or disease terms with the autocomplete function in place (Figure 5A). Users may filter the output list with important features such as data sources, breakpoint information, validation methods, functional classifications, etc. The preloaded numbers and dynamic pie charts allow users to estimate the number of hits in advance. The result is shown in a tabular format where the contents can be searched, sorted and downloaded. The interface also includes links to more detailed information and linkouts to external resources such as NCBI Entrez genes or USCS Genome Browser (Figure 5B). For example, click on a ChimerPub entry opens a new window showing PubMed abstract with important information highlighted (Figure 5C).

In an effort to help users assess functional significance of fusion transcripts of interest, we implemented several novel graphic utilities. Fusion structure viewer shows the transcript structures before and after fusion event, where users can readily view exons, breakpoints and domains (Figure



ChimerKB Results

Fusion Gene(5'_3')	5'Gene Junction	3'Gene Junction	Breakpoint Type	Disease	Chromosome Info.	Source	Supported	PMID
EML4-ALK	chr2:42531692	chr2:29446393	Exonic		Intra	COSMIC	P S+	
EML4-ALK	chr2:42345595	chr2:29269593	Exonic		Intra	ChimerDB2	P S+	
EML4-ALK	chr2:42492091	chr2:29446394	Exonic		Intra	GenBank	P S+	
EML4-ALK	chr2:42491871	chr2:29446394	Exonic		Intra	GenBank	P S+	
EML4-ALK	2p21	2p23.1		adenosquamous lung carcinoma	Intra	COSMIC	P S+	24481316
EML4-ALK	2p21	2p23.1		lung adenocarcinoma	Intra	COSMIC	P S+	20624322
EML4-ALK	2p21	2p23.1		lung pleomorphic carcinoma	Intra	COSMIC	P S+	24846683
EML4-ALK	2p21	2p23.1		neuroblastoma	Intra	COSMIC	P S+	21030459
EML4-ALK	2p21	2p23.1		lung neoplasm	Intra	COSMIC	P S+	17625570

Detailed information : EML4-ALK

	5' Gene		3' Gene
Gene Name	EML4 NCBI		ALK NCBI
Chromosome	chr2		chr2
Strand	+		-
Function	Oncogene, Transcription factor	Disease	lung adenocarcinoma
Genome Build Version	hg19	Source	COSMIC
Chromosome Information	Intra	Supported	Seq Pub

The EML4-ALK fusion gene is involved in various histologic types of lung cancers from nonsmokers with wild-type EGFR and KRAS.
Wong DW et al. **Cancer 2009 Apr; 115(8):1723-33**
PMID : 19170230; doi : 10.1002/cncr.24181

PloS one 2015 undefined;10(3):e0117032

Detection of EML4-ALK in lung adenocarcinoma using pleural effusion with FISH, IHC, and RT-PCR methods.

Abstract

Anaplastic lymphoma kinase (**ALK**) and echinoderm microtubule-associated protein-like 4 (**EML4**) gene rearrangements occur in approximately 5% of non-small-cell lung cancers (NSCLC),, leading to the overexpression of anaplastic lymphoma kinase and predicting a response to the targeted inhibitor, crizotinib. Malignant pleural effusion occurs in most patients with advanced lung cancer, especially adenocarcinoma, and tissue samples are not always available from these patients. We attempted to clarify the feasibility of detecting the **EML4-ALK** fusion gene in pleural effusion cells using different methods. We obtained 66 samples of pleural effusion from NSCLC patients. The pleural effusion fluid was centrifuged, and the cellular components obtained were formalin fixed and paraffin embedded. The **EML4-ALK** fusion gene status was determined with fluorescent in situ hybridization (FISH), reverse transcription-polymerase chain reaction (RT-PCR), and immunohistochemistry (IHC). **EML4-ALK** was detected in three of 66 patient samples (4.5%) with RT-PCR. When the RT-PCR data were used as the standard, one false positive and one false negative samples were identified with IHC; and one false negative sample was identified with FISH. These results suggest that a block of pleural effusion cells can be used to detect the **EML4-ALK** fusion gene. IHC had good sensitivity, but low specificity. FISH had low sensitivity, but high specificity. RT-PCR is a good candidate method for detecting **EML4-ALK** in blocks of pleural effusion cells from lung cancer. patients.

Figure 5. User interface of ChimerDB 4.0. (**A**) The search and filter window and output table in ChimerKB. (**B**) Main output form for a ChimerKB entry. Colored blocks are links to detailed information. (**C**) Example of a PubMed abstract where key words are highlighted. (**D**) Example of fusion structure viewer. (**E**) STRING network view. (**F**) Gene expression plots of 5′ and 3′ genes in fusion-positive versus fusion-negative patients in the TCGA dataset. (**G**) Scatter plots of gene expression versus copy number for 5′ and 3′ genes in the TCGA dataset.

5D). The graphic supports zoom-in/out and panning for detailed examination. We also added a protein–protein interaction network using the STRING plugin (Figure 5E) (22). Gene expression of 5′ and 3′ genes is informative in assessing the activity of fusion genes. For gene fusion events from the TCGA cohorts, we provide the gene expression plots of 5′ and 3′ genes in the fusion-positive and fusion-negative patients (Figure 5F), such that users can see whether the gene fusion leads to any dysregulation of 5′ or 3′ genes. The scatter plot of gene expression versus copy number provides additional insight into the function of fusion genes (Figure 5G). In the case of *EML4–ALK* fusion, it is evident that the fusion event is associated with elevated expression of 3′ *ALK* gene and that this overexpression is independent of copy number variation. Such information would be of great help in sifting candidate fusion genes with functional significance.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We appreciate anonymous reviewers for insightful suggestions to enhance the quality of database contents. We also thank the Genome-InfraNet at Korean Bioinformation Center for sharing their data (ID: 10050154).

FUNDING

National Research Foundation of Korea [NRF-2014M3C9A3065221, NRF-2015K1A4A3047851]; KRIBB Research Initiative Program. Funding for open access charge: National Research Foundation of Korea [NRF-2014M3C9A3065221].

Conflict of interest statement. None declared.

REFERENCES

1. Brien,G.L., Stegmaier,K. and Armstrong,S.A. (2019) Targeting chromatin complexes in fusion protein-driven malignancies. *Nat. Rev. Cancer*, **19**, 255–269.
2. Mertens,F., Johansson,B., Fioretos,T. and Mitelman,F. (2015) The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer*, **15**, 371–381.
3. Dinh,T.A., Vitucci,E.C., Wauthier,E., Graham,R.P., Pitman,W.A., Oikawa,T., Chen,M., Silva,G.O., Greene,K.G., Torbenson,M.S. *et al.* (2017) Comprehensive analysis of The Cancer Genome Atlas reveals a unique gene and non-coding RNA signature of fibrolamellar carcinoma. *Sci. Rep.*, **7**, 44653.
4. Embree,L.J., Azuma,M. and Hickstein,D.D. (2009) Ewing sarcoma fusion protein EWSR1/FLI1 interacts with EWSR1 leading to mitotic defects in zebrafish embryos and human cell lines. *Cancer Res.*, **69**, 4363–4371.
5. Matissek,K.J., Onozato,M.L., Sun,S., Zheng,Z., Schultz,A., Lee,J., Patel,K., Jerevall,P.L., Saladi,S.V., Macleay,A. *et al.* (2018) Expressed gene fusions as frequent drivers of poor outcomes in hormone receptor-positive breast cancer. *Cancer Discov.*, **8**, 336–353.
6. 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.
7. 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.
8. Torres-Garcia,W., Zheng,S., Sivachenko,A., Vegesna,R., Wang,Q., Yao,R., Berger,M.F., Weinstein,J.N., Getz,G. and Verhaak,R.G. (2014) PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics*, **30**, 2224–2226.
9. 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.
10. Kumar,S., Vo,A.D., Qin,F. and Li,H. (2016) Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Sci. Rep.*, **6**, 21597.
11. Zhao,J., Chen,Q., Wu,J., Han,P. and Song,X. (2017) GFusion: an effective algorithm to identify fusion genes from cancer RNA-Seq data. *Sci. Rep.*, **7**, 6880.
12. Kim,P., Jang,Y.E. and Lee,S. (2019) FusionScan: accurate prediction of fusion genes from RNA-Seq data. *Genomics Inform.*, **17**, e22.
13. Hass,B.J., Dobin,A., Li,B., Stransky,N., Pochet,N. and Regev,A. (2019) Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.*, **20**, 213.
14. Zhang,J., Gao,T. and Maher,C.A. (2017) INTEGRATE-Vis: a tool for comprehensive gene fusion visualization. *Sci. Rep.*, **7**, 17808.
15. 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.
16. Panigrahi,P., Jere,A. and Anamika,K. (2018) FusionHub: a unified web platform for annotation and visualization of gene fusion events in human cancer. *PLoS One*, **13**, e0196588.
17. Kim,D. and Salzberg,S.L. (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.*, **12**, R72.
18. 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.
19. Liu,Y., Sun,J. and Zhao,M. (2017) ONGene: a literature-based database for human oncogenes. *J. Genet. Genomics*, **44**, 119–121.
20. 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.
21. Kim,D., Lee,J., So,C.H., Jeon,H., Jeong,M., Choi,Y., Yoon,W., Sung,M. and Kang,J. (2019) A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, **7**, 73729–73740.
22. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.