D784–D789 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 28 November 2016*
doi: 10.1093/nar/gkw1083

ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining

Myunggyo Lee[1,†], Kyubum Lee[2,†], Namhee Yu[3,†], Insu Jang[4,†], Ikjung Choi[5], Pora Kim[6], Ye Eun Jang[1], Byounggun Kim[7], Sunkyu Kim[7], Byungwook Lee[4], Jaewoo Kang[2,7,*] and Sanghyuk Lee[1,3,5,*]

[1]Department of Bio-Information Science, Ewha Womans University, Seoul 03760, Republic of Korea, [2]Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea, [3]Department of Life Science, Ewha Womans University, Seoul 03760, Republic of Korea, [4]Korean Bioinformation Center, Korean Research Institute of Bioscience and Biotechnology, Daejeon 34141, Republic of Korea, [5]Ewha Research Center for Systems Biology, Ewha Womans University, Seoul 03760, Republic of Korea, [6]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA and [7]Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul 02841, Republic of Korea

Received September 15, 2016; Revised October 24, 2016; Editorial Decision October 24, 2016; Accepted October 27, 2016

ABSTRACT

Fusion gene is an important class of therapeutic targets and prognostic markers in cancer. ChimerDB is a comprehensive database of fusion genes encompassing analysis of deep sequencing data and manual curations. In this update, the database coverage was enhanced considerably by adding two new modules of The Cancer Genome Atlas (TCGA) RNA-Seq analysis and PubMed abstract mining. ChimerDB 3.0 is composed of three modules of ChimerKB, ChimerPub and ChimerSeq. ChimerKB represents a knowledgebase including 1066 fusion genes with manual curation that were compiled from public resources of fusion genes with experimental evidences. ChimerPub includes 2767 fusion genes obtained from text mining of PubMed abstracts. ChimerSeq module is designed to archive the fusion candidates from deep sequencing data. Importantly, we have analyzed RNA-Seq data of the TCGA project covering 4569 patients in 23 cancer types using two reliable programs of FusionScan and TopHat-Fusion. The new user interface supports diverse search options and graphic representation of fusion gene structure. ChimerDB 3.0 is available at http://ercsb.ewha.ac.kr/fusiongene/.

INTRODUCTION

Fusion genes have been firmly established as an important class of biomarkers and therapeutic targets in various types of cancer. A number of databases have been developed to catalogue fusion genes of clinical value. Initial efforts include the Mitelman database (1), COSMIC (2), ChimerDB 1.0 (3) and TICdb (4).

Since the advent of high-throughput sequencing technology, the deep sequencing data became the main source of identifying fusion genes. Numerous tools have been developed to predict fusion genes from genome or transcriptome sequencing data (5). However, the reliability of most programs falls short of the expectation from bench biologists or doctors who want the predictions to pass the validation experiments that typically require valuable resources such as patient tissues and time. Another major problem is the computer resources since typical programs require substantial amount of CPU time and memory. Thus, processing hundreds or thousands of RNA-Seq data for detection of fusion transcripts is almost impossible for most labs although massive amount of deep sequencing data are currently available in public. Nevertheless, several databases on fusion genes were released to include the results of analyzing transcriptome sequencing data. ChimerDB 2.0 was designed to be a knowledgebase of fusion genes with extensive manual curation and transcriptome sequencing data analysis (6). It has been used in many applications as a gold standard for developing prediction tools (7,8) and as a reference data set for fusion transcriptome simulation (9). ChiTaRS was developed with similar concepts and includes ∼29 000 fusion transcripts from eight species (10).

The amount of cancer genome sequencing data is exploding during the last several years. For example, The Cancer Genome Atlas (TCGA) provides RNA-Seq data for 10 539 tumor samples in 26 cancer types as of September 13,

*To whom correspondence should be addressed. Tel: +82 2 3277 2888; Fax: +82 2 3277 6809; Email: sanghyuk@ewha.ac.kr
Correspondence may also be addressed to Jaewoo Kang. Email: kangj@korea.ac.kr
†These authors contributed equally to the work as the first authors.

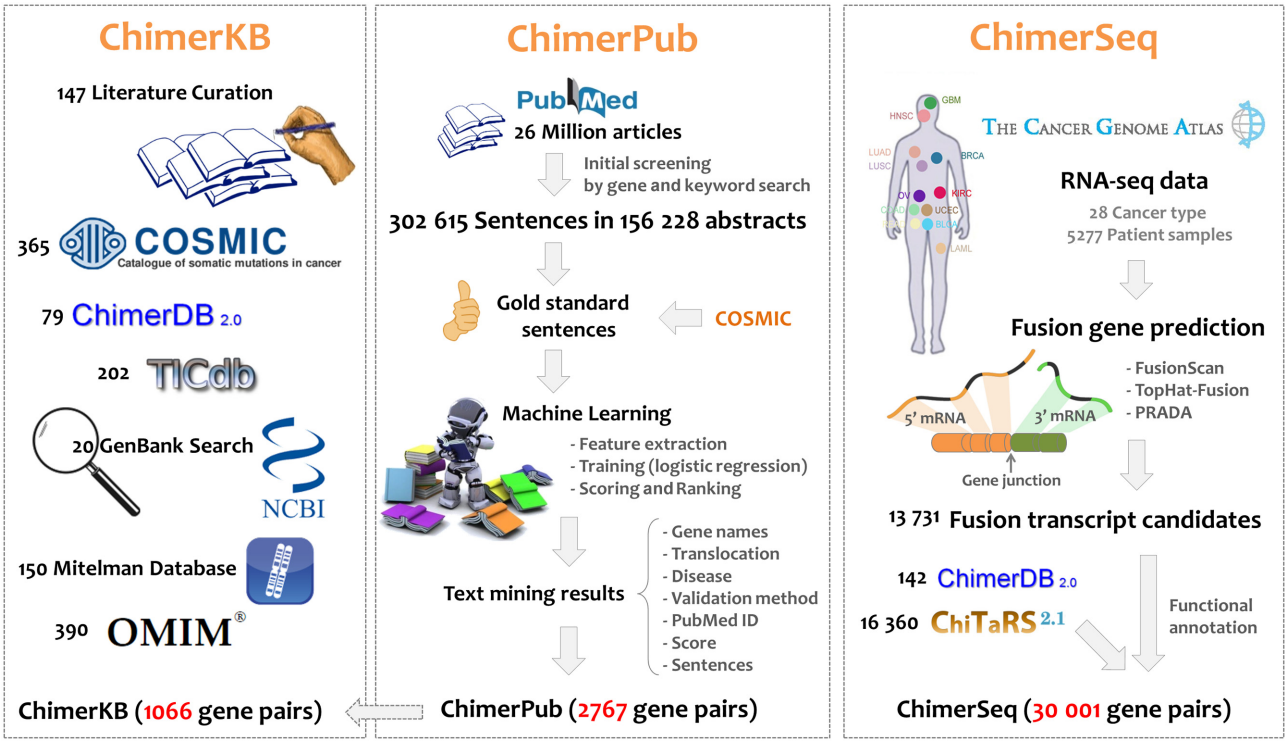

Figure 1. Overview of ChimerDB 3.0. Each number indicates the number of gene pairs from relevant resources.

2016. Stransky *et al*. focused on recurrent kinase fusion genes with oncogenic potentials (11). They applied an extensive filtering scheme of germline fusions observed in normal samples of the TCGA and Genotype-Tissue Expression projects (12). Approximately 3.0% of tumor samples were estimated to contain likely oncogenic, recurrent kinase fusion genes. Verhaak *et al*. developed a pipeline for predicting fusion genes named PRADA (13), and analyzed the TCGA RNA-Seq data to identify ~8000 fusion transcript candidates in 13 tumor types (14). However, these fusion gene databases contain computational results without experimental validation, and thus cannot be used for developing biomarkers or therapeutic targets of clinical value as they are.

Since the number of PubMed articles reporting fusion genes is rapidly increasing, cataloguing and curation itself has become a major challenge. Thus, automatic methods for text mining of PubMed articles to identify fusion genes would be of great help to build a comprehensive knowledgebase on fusion genes. In this updated version, we introduce a new module of text mining for fusion genes, provide an updated version of knowledgebase with manual curation and present a dramatically enhanced version of fusion transcripts obtained by analyzing TCGA RNA-Seq data with several advanced programs. ChimerDB 3.0 would be the most extensive catalog of fusion genes and transcripts publically available to date.

SYSTEM DESIGN AND METHODS

System overview

ChimerDB 3.0 is composed of three modules of ChimerKB, ChimerPub and ChimerSeq as shown in Figure 1. ChimerKB represents a knowledgebase of fusion genes that were compiled from well-known public resources such as GenBank, Mitelman (1), OMIM (15), COSMIC (2), TICdb (4), dbCRID (16) fusion databases and PubMed articles. All entries were manually curated for disease, sequences, breakpoints and experimental evidences. Specifically for PubMed articles reporting fusion genes, we examined the full text to find the relevant information.

ChimerPub is our effort to provide up-to-date information on *published* fusion genes. We developed an advanced text mining system to identify fusion genes from 26 million PubMed articles. Text mining techniques and an elaborate machine learning approach yielded a highly reliable pipeline for extracting genuine PubMed articles reporting fusion genes.

ChimerSeq collected fusion transcripts identified by computational analysis of transcriptome sequencing data including RNA-Seq and EST sequences. TCGA is the largest data set of deep sequencing for cancer patients currently available in public. We have re-analyzed RNA-Seq data from the TCGA with 4569 tumor samples in 23 cancer types using FusionScan (http://fusionscan.ewha.ac.kr/) and TopHat-Fusion (17) that we selected based on the benchmark test of precision and recall rates. Prediction results from PRADA pipeline were merged to build the TCGA fusion transcripts (PRADA ver. 1). We further compiled

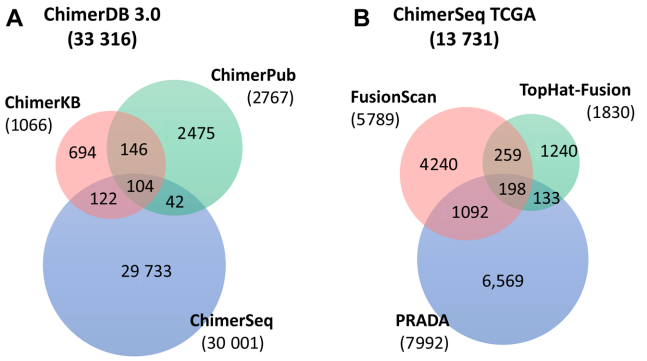

Figure 2. Statistics of ChimerDB 3.0. (**A**) Number of gene pairs from three modules. (**B**) Number of gene pairs from three prediction programs.

the fusion transcripts from ChiTaRS ver. 2.1 (18) and ChimerDB 2.0 (6) for better coverage.

ChimerPub development and implementation

ChimerPub is a new module that contains the fusion gene information obtained by text mining of PubMed abstracts. Computational procedure for identifying fusion-related PubMed abstracts is summarized in Figure 1 with more detailed information provided in Supplementary Figure S1. Out of ∼26 million PubMed abstracts as of June 2016, we searched all sentences containing multiple gene names that were recognized by the BEST entity extractor (19) after taxonomy filtering to remove articles on non-human species using PubTator (20). This initial screening yielded 302 615 sentences in 156 229 abstracts. We also extracted information on experimental methods, related diseases and translocation from the same abstract.

To build a classifier for fusion gene sentences based on machine learning technique, we prepared the positive and negative sentence sets using 283 known fusion cases from the COSMIC database (2) as detailed in Supplementary Figure S1. We identified 9277 sentences that contained both of gene names for the COSMIC fusion genes. Then, we randomly selected 1800 sentences and manually examined to obtain the gold standard positive set of 1549 sentences. Sentences not containing both of COSMIC fusion gene names were used as the negative data set in the training procedure even though they may include fusion genes not in the list of COSMIC fusion genes.

The procedure for constructing a classifier consists of the feature selection followed by logistic regression with extracted features. We compared the word distribution between positive and negative data sets to obtain 37 differential word features. We also added three features for fusion-specific information such as translocation description and experimental methods for validation. We applied logistic regression with 40 features to build a classification model. Finally, we scored all candidate sentences with the resulting regression model.

The reliability of high scoring sentences was evaluated in two ways. We checked 3000 top scoring sentences manually to see if they report genuine fusion genes and found that only one entry was not related to fusion gene. We further examined the cumulative probability of including two sets of positive sentences – gold standard sentences used in the training procedure and pseudo-positive sentences that included both of gene names of known fusion genes in the ChimerKB module. The number of known fusion genes is much larger in ChimerKB than in COSMIC database. Top 10 000 sentences recovered 38% and 49% of positive sets from gold positive and ChimerKB, respectively, with only 0.8% of the negative sentence set used in the training procedure (Supplementary Figure S2). For database construction, we selected 10 580 top scoring sentences, which were converted into 2767 fusion gene pairs. Thus, ChimerPub provides highly accurate list of fusion genes obtained from text mining of PubMed abstracts.

ChimerSeq development and implementation

ChimerSeq module includes fusion transcripts obtained from computational analysis of transcriptome sequencing data. RNA-Seq is the most popular source of data to predict fusion transcripts and a number of tools have been published so far. We have compared the performance of several tools including SOAPfuse (21), deFuse (22), FusionHunter (23), FusionMap (24), TopHat-Fusion (17) and FusionScan (our in-house developed program; http://fusionscan.ewha.ac.kr/) using RNA-Seq data of three cell lines (NCI-H660, K562 and MCF-7) whose genuine fusion genes were known. FusionScan and TopHat-Fusion achieved the best performance in the overall F_1-score, a combination of the precision and recall rates (Supplementary Table S1). Notably, FusionScan outperformed other programs in terms of the precision rate, which would be the most important factor for clinical utility (precision rate = 0.60). Thus, we chose FusionScan and TopHat-Fusion to analyze RNA-Seq data in the TCGA project.

The raw sequence data were downloaded from the CGHub of TCGA with the dbGap permission. Run time for FusionScan and TopHat-Fusion depends on the computer specification. Analyzing whole transcriptome data in the TCGA took several months with ∼600 CPU cores. To build a reliable list of fusion cases, we kept the fusion transcripts with the number of seed/junction reads ≥ 2 for FusionScan and PRADA, and the number of spanning pairs ≥ 100 for TopHat-Fusion. Number of fusion transcripts for each cancer type is shown in Supplementary Figure S3.

RESULTS

ChimerDB 3.0 includes 33 316 fusion gene pairs as summarized in the overall statistics of Table 1. Two representative modules for known fusion genes (i.e. ChimerKB and ChimerPub) takes ∼10% of the total and the remaining ∼90% are predicted ones from transcriptome sequencing (ChimerSeq), which require further experimental validation. Three modules are complementary since overlap among them is not large (Figure 2A).

Entries in the ChimerKB are most supported by other modules (45.1%, Table 1) as expected from the nature of knowledgebase. Of note, we collected fusion cases with known breakpoints carefully, where the genomic position and exon junction information were annotated for 106 and

Table 1. Statistics of ChimerDB 3.0

ChimerKB		ChimerPub		ChimerSeq	
Literature Curation	147	**Information available**		**TCGA**	13 731
COSMIC	365	Translocation	1248	FusionScan	5789
mRNA Sequence	273	Disease	1917	Tophat-Fusion	1830
Mitelman,OMIM,GenBank	495	Validation method	1147	PRADA	7992
		Others	741	**ChimerDB** 2.0	142
				ChiTaRS 2.1	16 360
Total	**1066**	**Total**	**2767**	**Total**	**30 001**
ChimerPub supported	250	ChimerKB supported	250	ChimerKB supported	226
ChimerSeq supported	226	ChimerSeq supported	146	ChimerPub supported	146
Known breakpoint cases				**Novel fusion***	29 733
Genomic position	106			TCGA	13 637
Exon junction	1450			ChiTaRS	16 149

All numbers represent the number of unique fusion genes.
*Transcripts not included in ChimerKB and ChimerPub were classified as novel fusion.

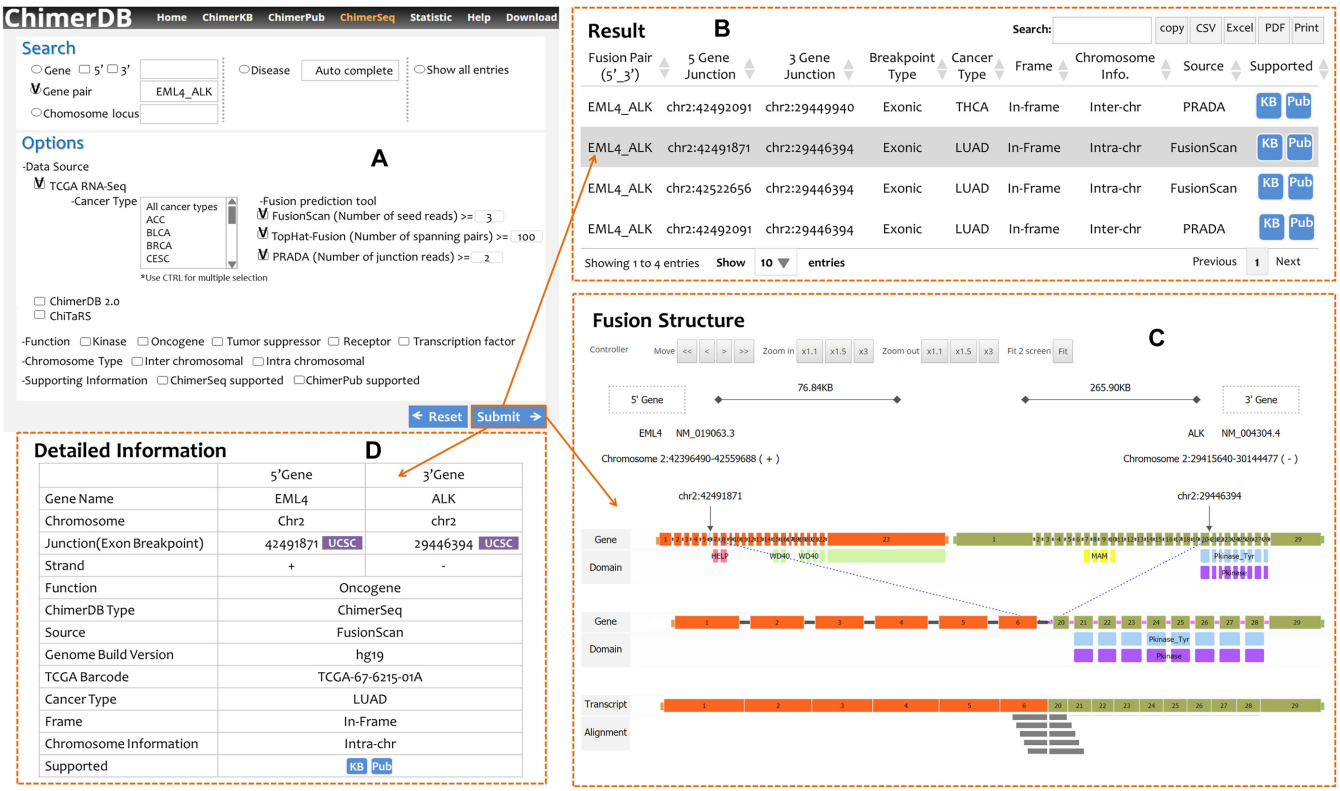

Figure 3. User interface of ChimerDB 3.0. (**A**) The search window for ChimerSeq is shown as an example of supporting user query for specific data source, cancer type and prediction tools. (**B**) Main output is the result table showing query hits with brief summary and links to further information. A click on each row activates the fusion structure window in (**C**) and the detailed information window (**D**). Note that we support the zooming and moving capability, exon structure with domain information and the alignment of seed/junction reads if available, in the fusion structure viewer.

1450 cases, respectively. This should be a useful resource for developing algorithms to predict fusion breakpoints *de novo* (25).

ChimerPub contains 2767 entries supported by literature publication. Only 250 of those are annotated in ChimerKB containing 1066 entries. It is evident that ChimerPub contributed a major portion of published fusion genes, emphasizing the importance of text mining in grasping current knowledge on fusion genes. ChimerPub entries were automatically annotated for information on disease, translocation and experimental methods, and the number of annotated fusion genes is shown in Table 1.

ChimerSeq takes ∼90% of fusion genes and most of its entries are not annotated or published, thus being the gold mine of novel fusion genes. However, it should be warned that many false positives are present even though we tried to use the most conservative programs to analyze the deep sequencing data. We expect that almost half of the prediction could be false, which should be acceptable considering the current status of prediction accuracy (Supplementary Table S1).

Since ChimerSeq represents a compilation of fusion transcripts from various resources based on computational analysis of transcriptome sequencing data, reliability estimation of each resource is very important for end users. We compared the overlap of predicted fusion transcripts with the ChimerKB and ChimerPub that represent the current knowledge of known fusion genes (Supplementary Table S2). The overlap ratio was in the order of TopHat-Fusion (1.75%), FusionScan (1.00%), PRADA (0.63%) and ChiTaRS (0.01%). Similarly, we also compared the overlap among three prediction programs (Figure 2B). The overlapping proportions were 32.2%, 26.8%, 17.8% for TopHat-Fusion, FusionScan and PRADA, respectively. The order was identical in two comparisons, thus users are recommended to take the reliability of the prediction tools in this order. TopHat-Fusion's prediction is most reliable but it misses many true positives. On the other hand, ChiTaRS has the most hits but it seems to contain many false positives. FusionScan seems to be a good compromise to this end.

USER INTERFACE

The user interface of ChimerDB was redesigned to accommodate new modules in this update. Figure 3 shows the important features in the user interface, taking *EML4-ALK* fusion as an example query to the ChimerSeq module. We support diverse types of search including gene, gene pair, chromosome locus and disease types. In ChimerSeq search, users may select the data source, cancer type and prediction tools with optional parameters. With ample annotations, we support diverse filtering options such as function filters for kinase, oncogene, tumor suppressor, receptor and transcription factor genes. Users may keep fusion transcripts supported from other modules as well for increased reliability or cross-checking.

Output GUI consists of a table of summary with search hits, a graphic illustration of fusion structure and detailed information on a specific fusion event. The output table supports many features of searching, sorting, exporting and linkouts to external resources. Click on each entry activates the graphic window of fusion gene structure and the detailed information table. The fusion gene graphic window shows readily the exons, domains and the breakpoint before and after the fusion event. This should be the most insightful picture for deducing functional significance of fusion event since the location of functional domains before and after gene fusion is illustrated. If available, we also show the alignment of short reads (seed/junction read only). We support zooming and panning for user convenience. The detailed information table provides all relevant information on the fusion transcript. Of note, the UCSC links guide users to the UCSC genome browser with short read alignment added as a custom track so that they can examine the detailed gene structure and alignment.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Research Foundation of Korea [NRF-2014M3C9A3065221 and NRF-2015K1A4A3047851 to S.L., NRF-2014R1A2A1A10051238 to J.K.]; KRIBB Research Initiative Program; Technology Innovation Program of the Ministry of Trade, Industry and Energy, Republic of Korea [10050154 to S.L.]. Funding for open access charge: National Research Foundation of Korea [NRF-2014M3C9A3065221].

Conflict of interest statement. None declared.

REFERENCES

1. Mertens,F., Johansson,B., Fioretos,T. and Mitelman,F. (2015) The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer*, **15**, 371–381.
2. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
3. Kim,N., Kim,P., Nam,S., Shin,S. and Lee,S. (2006) ChimerDB–a knowledgebase for fusion sequences. *Nucleic Acids Res.*, **34**, D21–D24.
4. Novo,F.J., de Mendibil,I.O. and Vizmanos,J.L. (2007) TICdb: a collection of gene-mapped translocation breakpoints in cancer. *BMC Genomics*, **8**, 33.
5. Wang,Q., Xia,J., Jia,P., Pao,W. and Zhao,Z. (2013) Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. *Brief. Bioinform.*, **14**, 506–519.
6. Kim,P., Yoon,S., Kim,N., Lee,S., Ko,M., Lee,H., Kang,H., Kim,J. and Lee,S. (2010) ChimerDB 2.0–a knowledgebase for fusion genes updated. *Nucleic Acids Res.*, **38**, D81–D85.
7. Abate,F., Zairis,S., Ficarra,E., Acquaviva,A., Wiggins,C.H., Frattini,V., Lasorella,A., Iavarone,A., Inghirami,G. and Rabadan,R. (2014) Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC Syst. Biol.*, **8**, 97.
8. Shugay,M., Ortiz de Mendibil,I., Vizmanos,J.L. and Novo,F.J. (2013) Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. *Bioinformatics*, **29**, 2539–2546.
9. Bruno,A.E., Miecznikowski,J.C., Qin,M., Wang,J. and Liu,S. (2013) FUSIM: a software tool for simulating fusion transcripts. *BMC Bioinformatics*, **14**, 13.
10. Frenkel-Morgenstern,M., Gorohovski,A., Lacroix,V., Rogers,M., Ibanez,K., Boullosa,C., Andres Leon,E., Ben-Hur,A. and Valencia,A. (2013) ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic Acids Res.*, **41**, D142–D151.
11. Stransky,N., Cerami,E., Schalm,S., Kim,J.L. and Lengauer,C. (2014) The landscape of kinase fusions in cancer. *Nat. Commun.*, **5**, 4846.
12. Mele,M., Ferreira,P.G., Reverter,F., DeLuca,D.S., Monlong,J., Sammeth,M., Young,T.R., Goldmann,J.M., Pervouchine,D.D., Sullivan,T.J. *et al.* (2015) Human genomics. The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.
13. 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.
14. Yoshihara,K., Wang,Q., Torres-Garcia,W., Zheng,S., Vegesna,R., Kim,H. and Verhaak,R.G. (2015) The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*, **34**, 4845–4854.
15. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
16. Kong,F., Zhu,J., Wu,J., Peng,J., Wang,Y., Wang,Q., Fu,S., Yuan,L.L. and Li,T. (2011) dbCRID: a database of chromosomal rearrangements in human diseases. *Nucleic Acids Res.*, **39**, D895–D900.

17. Kim,D. and Salzberg,S.L. (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.*, **12**, R72.
18. Frenkel-Morgenstern,M., Gorohovski,A., Vucenovic,D., Maestre,L. and Valencia,A. (2015) ChiTaRS 2.1–an improved database of the chimeric transcripts and RNA-seq data with novel sense-antisense chimeric RNA transcripts. *Nucleic Acids Res.*, **43**, D68–D75.
19. Lee,S., Kim,D., Lee,K., Choi,J., Kim,S., Jeon,M., Lim,S., Choi,D., Kim,S., Tan,A.C. *et al.* (2016) BEST: Next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PLoS One*, **11**, e0164680.
20. Wei,C.H., Kao,H.Y. and Lu,Z. (2013) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.*, **41**, W518–W522.
21. Jia,W., Qiu,K., He,M., Song,P., Zhou,Q., Zhou,F., Yu,Y., Zhu,D., Nickerson,M.L., Wan,S. *et al.* (2013) SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biol.*, **14**, R12.
22. McPherson,A., Hormozdiari,F., Zayed,A., Giuliany,R., Ha,G., Sun,M.G., Griffith,M., Heravi Moussavi,A., Senz,J., Melnyk,N. *et al.* (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.*, **7**, e1001138.
23. Li,Y., Chien,J., Smith,D.I. and Ma,J. (2011) FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics*, **27**, 1708–1710.
24. Ge,H., Liu,K., Juan,T., Fang,F., Newman,M. and Hoeck,W. (2011) FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics*, **27**, 1922–1928.
25. Wijaya,E., Shimizu,K., Asai,K. and Hamada,M. (2014) Reference-free prediction of rearrangement breakpoint reads. *Bioinformatics*, **30**, 2559–2567.