Published online 11 November 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D81–D85***
doi:10.1093/nar/gkp982

ChimerDB 2.0—a knowledgebase for fusion genes updated

Pora Kim[1,2], Suhyeon Yoon[1,2], Namshin Kim[3], Sanghyun Lee[2], Minjeong Ko[1,2], Haeseung Lee[1,2], Hyunjung Kang[1,2], Jaesang Kim[1,2] and Sanghyuk Lee[1,2,*]

[1]Division of Life and Pharmaceutical Sciences, [2]Ewha Research Center for Systems Biology, Ewha Womans University, Seoul 120-750 and [3]Bioinformatics Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Korea

Received September 8, 2009; Revised and Accepted October 15, 2009

ABSTRACT

Chromosome translocations and gene fusions are frequent events in the human genome and have been found to cause diverse types of tumor. ChimerDB is a knowledgebase of fusion genes identified from bioinformatics analysis of transcript sequences in the GenBank and various other public resources such as the Sanger cancer genome project (CGP), OMIM, PubMed and the Mitelman's database. In this updated version, we significantly modified the algorithm of identifying fusion transcripts. Specifically, the new algorithm is more sensitive and has detected 2699 fusion transcripts with high confidence. Furthermore, it can identify interchromosomal translocations as well as the intrachromosomal deletions or inversions of large DNA segments. Importantly, results from the analysis of next-generation sequencing data in the short read archives are incorporated as well. We updated and integrated all contents (GenBank, Sanger CGP, OMIM, PubMed publications and the Mitelman's database), and the user-interface has been improved to support diverse types of searches and to enhance the user convenience especially in browsing PubMed articles. We also developed a new alignment viewer that should facilitate examining reliability of fusion transcripts and inferring functional significance. We expect ChimerDB 2.0, available at http://ercsb.ewha.ac.kr/fusiongene, to be a valuable tool in identifying biomarkers and drug targets.

INTRODUCTION

Fusion genes play important roles in tumorigenesis and cancer progression (1). Perhaps, the best-characterized case, BCR-ABL1 fusion is the cause of the chronic myelogenous leukemia and the target of the anticancer drug, Gleevec (imatinib) (2). Identification of fusion genes thus can lead to the discovery of diagnostic biomarkers and therapeutic targets as well as understanding the molecular basis of tumorigenesis.

Initial studies have concentrated on the hematological cancer in large part due to the sample availability (1,3). Over the last few years, however, there has been significant progress in fusion gene identification in solid tumors. Importantly, Chinnaiyan and colleagues (4–7) reported several cases of gene fusion in prostate cancer identified via integrative analysis of microarray data (TMPRSS2 and ETS transcription factors) and transcriptome resequencing. Soda *et al.* (8) identified the transforming EML4-ALK fusion gene in nonsmall cell lung cancer (NSCLC) using a function-based screening procedure. A proteomic study of phosphotyrosine kinases also revealed the ROS-ALK fusion in NSCLC cell lines (9). These cases clearly indicate that gene fusions play an important role in cancer development in solid tumors.

Recent progress in next-generation sequencing (NGS) techniques provides a tremendous opportunity for fusion gene discovery. Notably, paired-end sequencing, now a frequent if not standard procedure, compensates for the short read length of NGS techniques (10). Sequencing and analyzing whole genome or transcriptome lead to identification of many chromosomal aberrations including translocations, amplifications and deletions. Short read sequencing strategies were successfully applied to find

*To whom correspondence should be addressed. Tel: +82 2 3277 2888; Fax: +82 2 3277 3760; Email: sanghyuk@ewha.ac.kr

The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

fusion genes in prostate, lung and breast cancer cell lines (5,11–13).

There have been considerable efforts to make a catalog of fusion genes. The Mitelman's database and the Sanger cancer genome project (CGP) are the notable examples of collecting fusion genes from literature reports (1,3). The COSMIC and CancerGenes database include other types of chromosomal aberrations such as mutations, amplifications and deletions (14,15). Currently, the Mitelman's database and the Sanger CGP collection include 150 and 270 gene pairs, respectively, involved in gene fusion events.

Bioinformatics analysis of public transcriptome sequences in the GenBank also provides ample cases of fusion transcript candidates. Fusion genes may be classified into two groups, interchromosomal and intrachromosomal. The former results from fusion between two different chromosomes i.e. translocation and the latter originates from single chromosomes due to deletion, inversion or amplification of large DNA segments. Romani *et al.* (16) analyzed the mRNA sequences and Hahn *et al.* (17) analyzed the mRNA and EST sequences to identify fusion transcripts between different chromosomes. Similar data-mining approaches were adopted to construct databases of fusion genes such as ChimerDB (18), HybridDB (19) and TICdb (20) although computational details vary considerably.

ChimerDB is designed to be a knowledgebase of fusion genes that encompass the fusion transcripts identified from bioinformatics analysis of transcript sequences in the GenBank and various public resources such as the Sanger CGP (3), OMIM (21), Mitelman's database (1) and PubMed. The updated version, ChimerDB 2.0, features (i) algorithm modifications for increased sensitivity, (ii) extensive coverage of recent publications and relevant databases, (iii) analysis of NGS data in the NCBI's short read archives (SRA) and (iv) the enhanced user interface and the novel alignment viewer to support diverse types of search. ChimerDB 2.0 would be the most extensive catalog of fusion genes and transcripts publically available to date.

IMPLEMENTATIONS

Computational method for transcriptome analysis

The basic strategy is virtually identical to the procedure used in ChimerDB 1.0 where the genomic alignments of transcript sequences were analyzed to identify the fusion transcripts. We will describe the major differences and modifications here with more details provided in the Supplementary data and in the web site documentation.

The most important change is relieving the boundary conditions based on our observation that many reported cases did not satisfy the strict condition that the fusion boundary of the transcript should match the exon boundary. Therefore, we introduced the 'reliability class' as a measure of confidence level. We consider the alignment with multiple exons or single exon with matching boundaries as features of reliability. Entry to Class A requires that both head and tail transcripts consist of multiple exons or of single exons with matching boundaries, thus being the most reliable cases. Only one or neither of the head and tail genes satisfying this condition would put a given transcript in Class B or C, respectively.

Another important difference is the introduction of various refinement steps. For example, we removed the entries whose genomic alignments have many hits of comparable qualities in different genomic regions even though these genomic regions are not marked as repeat sequences. This step was necessary to avoid possible complications arising from gene duplication, pseudo-genes and retroposon sequences. In addition, the number of exons was estimated by using the Exonerate program rather than the BLAT alignments (22,23).

The computational pipeline for 454 sequences from the SRA is identical with the EST processing since the sequence length is comparable. Solexa reads are generally too short and we used them just as supporting evidence for the existence of fusion transcripts. Solexa transcriptome reads were aligned using the BWA program (24) against the fusion transcripts to determine if multiple reads cover the fusion point. The alignment of resulting candidates was manually examined.

In this updated version, we also include the fusion transcripts within the same chromosome. The head and tails genes, not being adjacent, should be separated by >1 Mb. We exclude the fusion cases between adjacent genes, which we named co-transcription and intergenic splicing, in order to limit our focus to genuine fusion genes originating from chromosomal aberrations.

Data sources

Transcript sequences were downloaded from the GenBank last updated on September, 2008. It included 323 914 mRNA and over 8 million EST sequences for human. We also downloaded NGS transcriptome sequences in the SRA that included $\sim$1.2 million 454 sequences and $\sim$762 million Solexa reads. The human genome map used for transcriptome analysis was the NCBI build 36.1 (hg18 in the UCSC genome browser database) (25).

Literature-related information was obtained as follows. PubMed articles related to fusion genes were retrieved by using the Entrez query of 'chromosomal translocation or fusion gene' and the MeSH terms on human cancers. Abstracts of 3618 articles were manually examined to obtain information on the fusion gene pairs. OMIM records retrieved by the query of 'translocation or fusion' (May 2009) were also manually examined to find fusion gene pairs. As for the Sanger CGP data, the 'cancer gene census' list released on December 2008 was downloaded from the web site. Mitelman's database was obtained from the web site for the recurrent chromosome aberrations in cancer (http://cgap.nci.nih.gov/Chromosomes/RecurrentAberrations) as of April 2009. Entries with specific gene symbols for both head and tail genes were retained as part of our literature-related data.

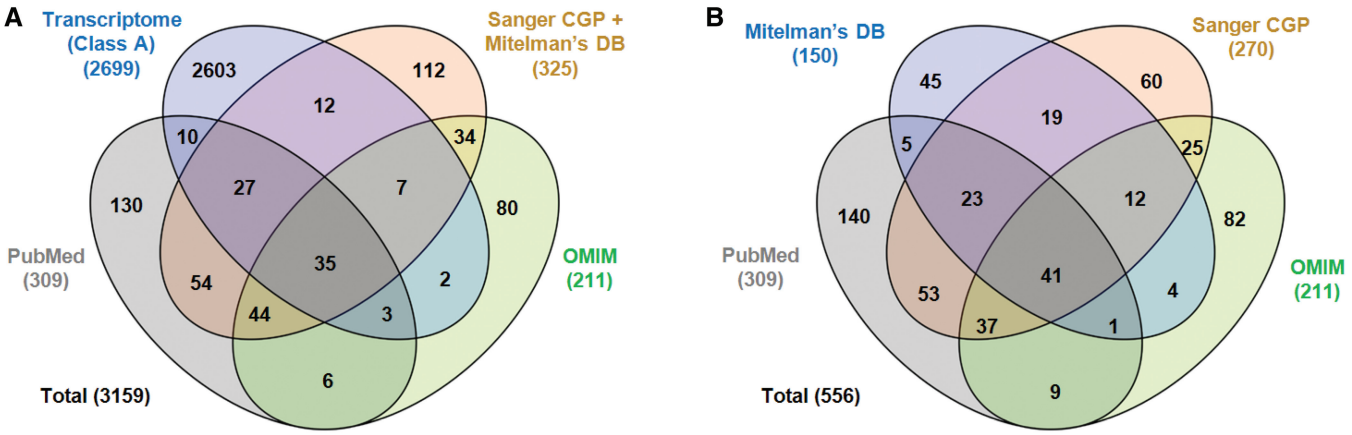

Figure 1. The number of gene pairs in the ChimerDB 2.0 according to the information source.

Table 1. Statistics of transcriptome analysis in ChimerDB 2.0

Class	Interchromosomal			Intrachromosomal		
	A	A + B	A + B + C	A	A + B	A + B + C
No. of transcripts	1900	6073	8833	515	887	1065
mRNA	479	855	900	110	143	146
EST	1247	3972	5397	396	677	781
NGS	174	1246	2536	9	67	138
No. of genes (9358 in total)	2855	6976	8710	703	1276	1543
No. of gene pairs	2209	7362	10639	490	909	1108
With multiple transcripts	278	807	1137	144	220	246
With Solexa evidence	14	14	15	65	67	67

RESULTS

ChimerDB 2.0 includes 9358 genes, 11747 fusion gene pairs and 9358 fusion transcripts. Figure 1A shows the number of fusion gene pairs according to the information source, counting just the Class A candidates for the transcriptome analysis. As expected, transcriptome analysis is the most ample source of fusion gene pairs and includes 2699 candidates, compared with ~300 candidates with the original version. Only 96 cases of those have the literature evidence from other resources, implying that the majority of candidates remain to be verified experimentally.

Comparison between databases for just the literature-based cases is also revealing (Figure 1B). The overlaps between Sanger CGP, OMIM, Mitelman's database and our own PubMed collections are much smaller than expected, and 327 cases out of 556 fusion gene pairs are found only in one of the literature databases. This reveals the incomplete coverage of manual efforts and the necessity for integration of various databases. ChimerDB 2.0 includes 537 genes and 556 fusion gene pairs from literature publications, which is a significant increase than other single database.

Detailed statistics of transcriptome analysis is shown in Table 1. We found 1046, 6178 and 2674 fusion transcripts from mRNA, EST and 454 sequences, respectively. In sum, 89% of the total cases are interchromosomal fusions.

A significant proportion of gene pairs (422) in the Class A features multiple fusion transcripts, thus indicating an even higher chance of representing genuine fusion. We also searched the short reads from the Solexa transcriptome sequences in the SRA that span the fusion boundary of our candidates. Notably, 82 fusion transcripts were found to have multiple short-read matches. One of these fusions, CRTC1-MAML2, has been reported to be a frequent feature of mucoepidermoid carcinoma (26). It is noteworthy that intrachromosomal events are overrepresented among our fusion transcript candidates. A close look reveals that a major portion consists of genes belonging to the same family or of pseudogenes. It remains to be seen whether they are from alignment ambiguity or genuine fusion events.

User interface

The user interface of ChimerDB is significantly improved in this updated version. Figure 2 shows the important features in the user interface. Most importantly, we support diverse types of search targeting transcripts, genes, gene pairs, cytobands and tissues. As for the fusion transcripts, users may choose the reliability class, number of exons, boundary types for the head and tail genes.

The result page includes the gene pairs, disease information, PubMed articles and linkouts to diverse resources. PubMed articles are displayed with the title and journal name for user convenience. Except for the

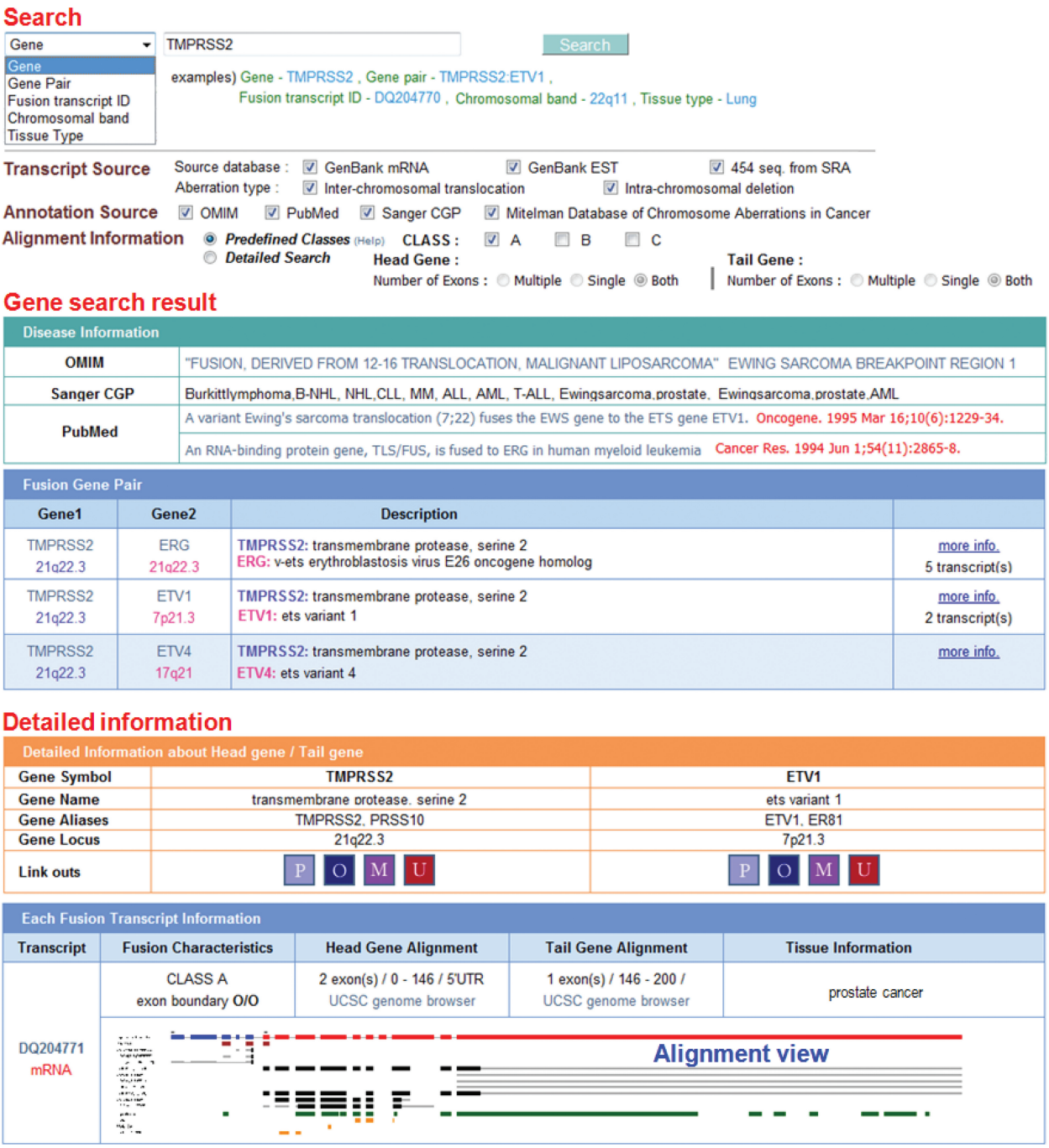

Figure 2. User interface of ChimerDB 2.0. The search page is designed to support diverse types of search. The 'search result' page shows the gene pairs and the disease-related information in OMIM, Sanger CGP and Mitelman's database with the title and journal name of PubMed articles. Clicking 'more info' link shows the detailed information on fusion genes and transcripts as seen in the bottom panel. The 'alignment view' shows the hypothetical fusion gene (head gene in blue, tail gene in red) and the candidate fusion transcript (in magenta) along with the UCSC-annotated genes (exons in black, UTRs in grey). The repeat regions and the Pfam domains are indicated in green and orange colors, respectively. Clicking on the alignment picture opens a magnified view. The information contents in this figure are trimmed for brevity.

few cases without the fusion sequence available, we show the alignment picture that includes the gene structure, domains and repeat sequences. We also provide information on tissue and pathology type from the GenBank records and CGAP (Cancer Genome Anatomy Project) library data. Links to the UCSC genome browser are provided to allow users to examine the detailed gene structure and alignment.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Korea Science and Engineering Foundation (KOSEF) funded by the Korea government (MEST) (R01-2008-000-20818-0 and 2007-03983); grant from BioGreen 21 Program of the Korean Rural Development Administration (20070401034010); 'Systems Biology Infrastructure Establishment Grant' provided by Gwangju Institute of Science and Technology in 2009 through Ewha Research Center for Systems Biology (ERCSB); grant from the National Core Research Center (NCRC) program (R15-2006-020) of the KOSEF funded by the MEST. Funding for open access charge: Korea Science and Engineering Foundation (R01-2008-000-20818-0).

Conflict of interest statement. None declared.

REFERENCES

1. Mitelman,F., Johansson,B. and Mertens,F. (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.
2. Hunter,T. (2007) Treatment for chronic myelogenous leukemia: the long road to imatinib. *J. Clin. Invest.*, **117**, 2036–2043.
3. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
4. Kumar-Sinha,C., Tomlins,S.A. and Chinnaiyan,A.M. (2008) Recurrent gene fusions in prostate cancer. *Nat. Rev. Cancer*, **8**, 497–511.
5. Maher,C.A., Kumar-Sinha,C., Cao,X., Kalyana-Sundaram,S., Han,B., Jing,X., Sam,L., Barrette,T., Palanisamy,N. and Chinnaiyan,A.M. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97–101.
6. Tomlins,S.A., Mehra,R., Rhodes,D.R., Cao,X., Wang,L., Dhanasekaran,S.M., Kalyana-Sundaram,S., Wei,J.T., Rubin,M.A. *et al*. (2007) Integrative molecular concept modeling of prostate cancer progression. *Nat. Genet.*, **39**, 41–51.
7. Tomlins,S.A., Rhodes,D.R., Perner,S., Dhanasekaran,S.M., Mehra,R., Sun,X.W., Varambally,S., Cao,X., Tchinda,J., Kuefer,R. *et al*. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.
8. Soda,M., Choi,Y.L., Enomoto,M., Takada,S., Yamashita,Y., Ishikawa,S., Fujiwara,S., Watanabe,H., Kurashina,K., Hatanaka,H. *et al*. (2007) Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*, **448**, 561–566.
9. Rikova,K., Guo,A., Zeng,Q., Possemato,A., Yu,J., Haack,H., Nardone,J., Lee,K., Reeves,C., Li,Y. *et al*. (2007) Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell*, **131**, 1190–1203.
10. Fullwood,M.J., Wei,C.L., Liu,E.T. and Ruan,Y. (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.*, **19**, 521–532.
11. Campbell,P.J., Stephens,P.J., Pleasance,E.D., O'Meara,S., Li,H., Santarius,T., Stebbings,L.A., Leroy,C., Edkins,S., Hardy,C. *et al*. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.
12. Ruan,Y., Ooi,H.S., Choo,S.W., Chiu,K.P., Zhao,X.D., Srinivasan,K.G., Yao,F., Choo,C.Y., Liu,J., Ariyaratne,P. *et al*. (2007) Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res.*, **17**, 828–838.
13. Zhao,Q., Caballero,O.L., Levy,S., Stevenson,B.J., Iseli,C., de Souza,S.J., Galante,P.A., Busam,D., Leversha,M.A., Chadalavada,K. *et al*. (2009) Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proc. Natl Acad. Sci. USA*, **106**, 1886–1891.
14. Bamford,S., Dawson,E., Forbes,S., Clements,J., Pettett,R., Dogan,A., Flanagan,A., Teague,J., Futreal,P.A., Stratton,M.R. *et al*. (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer*, **91**, 355–358.
15. 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.
16. Romani,A., Guerra,E., Trerotola,M. and Alberti,S. (2003) Detection and analysis of spliced chimeric mRNAs in sequence databanks. *Nucleic Acids Res.*, **31**, e17.
17. Hahn,Y., Bera,T.K., Gehlhaus,K., Kirsch,I.R., Pastan,I.H. and Lee,B. (2004) Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. *Proc. Natl Acad. Sci. USA*, **101**, 13257–13261.
18. Kim,N., Kim,P., Nam,S., Shin,S. and Lee,S. (2006) ChimerDB–a knowledgebase for fusion sequences. *Nucleic Acids Res.*, **34**, D21–D24.
19. Kim,D.S., Huh,J.W. and Kim,H.S. (2007) HYBRIDdb: a database of hybrid genes in the human genome. *BMC Genomics*, **8**, 128.
20. 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.
21. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
22. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
23. Slater,G.S. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.
24. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
25. Kuhn,R.M., Karolchik,D., Zweig,A.S., Wang,T., Smith,K.E., Rosenbloom,K.R., Rhead,B., Raney,B.J., Pohl,A., Pheasant,M. *et al*. (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.
26. Fehr,A., Roser,K., Heidorn,K., Hallas,C., Loning,T. and Bullerdiek,J. (2008) A new type of MAML2 fusion in mucoepidermoid carcinoma. *Genes Chromosomes Cancer*, **47**, 203–206.