D142–D151 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 9 November 2012*
doi:10.1093/nar/gks1041

ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data

Milana Frenkel-Morgenstern[1], Alessandro Gorohovski[1], Vincent Lacroix[2], Mark Rogers[3], Kristina Ibanez[1], Cesar Boullosa[1], Eduardo Andres Leon[4], Asa Ben-Hur[3] and Alfonso Valencia[1,*]

[1]Structural Biology and BioComputing Program, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain, [2]UMR CNRS 5558, Laboratoire de Biométrie et Biologie Evolutive, INRIA Bamboo, Université Claude Bernard, Villeurbanne 69100, France, [3]Department of Computer Science, Colorado State University, Fort Colins, CO 80523-1873, USA and [4]Bioinformatics Unit, Structural Biology and BioComputing Program, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain

Received August 15, 2012; Revised October 2, 2012; Accepted October 5, 2012

ABSTRACT

Chimeric RNAs that comprise two or more different transcripts have been identified in many cancers and among the Expressed Sequence Tags (ESTs) isolated from different organisms; they might represent functional proteins and produce different disease phenotypes. The ChiTaRS database of Chimeric Transcripts and RNA-Sequencing data (http://chitars.bioinfo.cnio.es/) collects more than 16 000 chimeric RNAs from humans, mice and fruit flies, 233 chimeras confirmed by RNA-seq reads and ~2000 cancer breakpoints. The database indicates the expression and tissue specificity of these chimeras, as confirmed by RNA-seq data, and it includes mass spectrometry results for some human entries at their junctions. Moreover, the database has advanced features to analyze junction consistency and to rank chimeras based on the evidence of repeated junction sites. Finally, 'Junction Search' screens through the RNA-seq reads found at the chimeras' junction sites to identify putative junctions in novel sequences entered by users. Thus, ChiTaRS is an extensive catalog of human, mouse and fruit fly chimeras that will extend our understanding of the evolution of chimeric transcripts in eukaryotes and can be advantageous in the analysis of human cancer breakpoints.

INTRODUCTION

The eukaryote transcriptome is composed of RNAs transcribed from almost any location in the genome (1–6). Although most RNAs can be assigned to a single locus, some of them, called chimeras, are composed of exons from distinct genes and are therefore assigned to several loci (1,7–24). In some cases, the loci are close to each other in the genome, suggesting that the chimera is generated by read-through transcription (1,12). In other instances, the loci are megabases apart or on different chromosomes, suggesting that the chimera is generated through genome rearrangements or trans-splicing (9,22). Although the possibility that some chimeras are the *in vitro* artifact of template switching by the reverse transcriptase cannot be totally ruled out (reverse transcriptase–free assays are much harder to perform) (25), the recent evidence that some chimeras are translated corroborates their authenticity and motivated us to establish a systematic catalog of all chimeras (19). Another reason to categorize chimeras is their association with cancer, when the transcriptome is notoriously more complex owing to a large number of genome rearrangements, mutations and alterations of the splicing machinery (26,27).

The best-characterized chimeric transcript example is the BCR–ABL1 fusion that is expressed strongly in chronic myelogenous leukemia (23,24). Indeed, this fusion is the target of the anticancer drug imatinib (28,29). Thus, the therapeutic relationship highlights the benefits that can be obtained from identifying chimeric transcripts in cancers and other diseases, both as potential drug targets and as diagnostic tools (30–32).

Next-generation sequencing technology provides a great opportunity to identify chromosomal aberrations and novel fusion genes (14,23,24,31,33,34). Indeed, the TMPRSS2 and ETS fusion was identified in prostate cancer by RNA-sequencing and microarray data analysis (23,24,34). Similarly, the EML4–ALK fusion gene was

*To whom correspondence should be addressed. Tel: +34917328000; Email: avalencia@cnio.es
Present address:
Alessandro Gorohovski, National Technical University of Ukraine (KPI), Kiev 03056, Ukraine.

identified in non–small-cell lung cancer using a functional screening procedure (35,36). Short-read sequencing strategies were successfully applied to find fusion genes in prostate, lung and breast cancer cell lines (18,23,24,34,37,38). These are just a few examples where it has been possible to show how gene fusions are associated with solid tumor development, and more examples are likely to come.

We recently screened thousands of candidate chimeric transcripts using functional annotation, high-throughput RNA sequencing and mass spectrometry, and identified 175 chimeric RNAs and 12 novel chimeric proteins expressed in humans (19,20). Generally, chimeric transcripts are expressed weakly in normal tissues, although these chimeras tend to incorporate highly expressed parental genes (19). Moreover, we presented evidence of chimeras that had lost certain functional domains and that might therefore actively compete with the functional wild-type proteins, producing dominant negative effects in cancers and other diseases (20). Hence, the screening of the Expressed Sequence Tag (EST) databases and RNA-sequence mapping may have certain advantages when attempting to identify novel chimeric transcripts in cancers, or in normal cells (39).

An enormous effort has been made to catalog chimeric transcripts from the literature: the Mitelman database (40) and the Sanger cancer genome project (41) , including the COSMIC database (27,42,43). GenBank (44) also provides a resource to identify candidate inter-chromosomal or intra-chromosomal chimeric transcripts from EST and mRNA data sets (13,45,46). Several fundamental databases have been constructed to incorporate chimeric transcripts from different resources and using a variety of computational procedures: ChimerDB 2.0 (47), ChimerDB (48), HybridDB (49), TICdb (50) and dbCrid (51). Although these databases have been very useful and supported the research in the area, none of them integrates EST or mRNA sequences and literature resources together with RNA-sequencing data, expression level and tissue specificity of chimeric transcripts in different tissues and organisms.

Our ChiTaRS database is designed to incorporate chimeric transcripts from three organisms (human, mouse and fruit fly (drosophila)), which helps to provide evidence of chimeras conserved in these organisms. The database was generated by performing a bioinformatics analysis of transcript sequences for the three organisms in GenBank (44). The special features of ChiTaRS include the use of an algorithm optimized for the quick retrieval and search of 16 262 chimeric transcripts in the three organisms, using various search parameters. It includes an extensive coverage of recent publications and relevant databases that collate 1892 cancer breakpoints and read-through fusions, as well as manual verification of the entries. Moreover, the database incorporates evidence from RNA-seq reads that map 233 chimeric junction sites from multiple next-generation sequencing data sets for the three organisms, providing information regarding the level of expression and the tissue specificity of the entries. The download page includes all the entries and tables in the database, together with the RNA-seq and mass spectrometry data supporting the existence of these chimeras. ChiTaRS also enables the transcripts and their junction site to be visualized by SpliceGrapher (52), using the genome annotation for humans, mice and the fruit flies. Finally, the database has a unique feature to analyze the junction consistency, ranking the chimeras according to the evidence of the same junction site. This feature is advantageous to researchers seeking empirical confirmation of highly ranked chimeric transcripts. As a result, ChiTaRS represents the most extensive catalog of chimeric RNA transcripts in the human, mouse and fruit fly, which makes it particularly important that the data are presented in an easily understandable and user-friendly format.

RESULTS

Data sets of candidate chimeric transcripts

We have created a data set of chimeric transcripts using ESTs and mRNAs sequences for human [the UCSC reference genome (51–53): GRCh37/hg19], mouse (NCBI37/mm9) and *Drosophila* (BDGP R5/dm3) from GenBank (44). All sequences were aligned to the corresponding reference genomic sequences using the UCSC BLAT program (53,54). The sequence was considered a chimera whenever the first part aligned to one gene and the second to another gene located at least 750 kb away [the default maximum intron size in BLAT (54)]. For the alignments, identity was set at a minimum of 95%, and the minimum length was set at 50 nucleotides (nt). We allowed an overlap of up to 10nt between the two subparts of a chimera; therefore ChiTaRS also includes chimeric transcripts with short homologous sequences (46). Furthermore, we did not put any constraint on the splices sites, hence ChiTaRS contains chimeras with either canonical or non canonical splice sites. In this way, 14 512 human, 10 550 mouse and 4084 fruit fly candidate chimeras were identified. The chimeric transcripts incorporating opposite strands of the same gene were removed to avoid fusion by cotranscription and intergenic splicing (CoTIS) (47,48). Moreover, the chimeric junction sites were characterized to distinguish between genuine chimeras and artifacts, as the junction in a chimera is typically around the exon–exon splice sites (45,47). Applying this filter, whereby candidate chimeric sequences were removed if the junction was situated >50 nt away from a known splice site, reduced the number of chimeric candidates to 9379, 4828 and 2055 in the human, mouse and fruit fly data sets, respectively. These candidate chimeras involved 7808, 5141 and 1784 unique genes from human, mouse and fruit fly, respectively. It is worth noting that this pipeline did not capture read-through fusions (1), as they involve genes located <750 kb away. The read-through chimeras that we included in the database were added separately, based on adequate published supporting evidence (see 'Full Collection & Search').

RNA-sequencing analysis of candidate chimeric transcripts

To assess the expression and validate the authenticity of the candidate chimeric transcripts, we screened RNA-seq data sets from the corresponding organism (19).

For human candidate chimeras, we used the Human Body Map 2.0 data generated on the HiSeq 2000 by Illumina in 2010. This data set comprises 1097 million (M) reads of 75 nt derived from the sequencing of RNA from 16 different tissues. For the candidate drosophila chimeras, we used a data set of 22 M reads of 75 nt resulting from the sequencing of the ovarian cell line Kc167 (the RGASP competition, the modENCODE group). For the candidate drosophila chimeras, we used a data set of 22M reads of 75 nt resulting from the sequencing of the ovarian cell line Kc167 (the RGASP competition, the modENCODE group) (55). Clearly, the depth (number of reads sequenced) and breadth (number of tissues sampled) of the sequencing differed between the data sets, which explains why the proportion of chimeras we confirmed was different for each organism: 192 for humans, 12 for mice, 29 for fruit flies (see 'Full Collection').

To ensure that a RNA-seq read could be unambiguously assigned to a chimera and not to another location in the genome, we followed a specific mapping protocol (19). First, we mapped all RNA-seq reads to the reference genome and annotated exon junctions using the Grape RNAseq Analysis Pipeline Environment (GRAPE) (http://big.crg.cat/services/grape), and thereby removed any reads that could be linearly assigned to genomic regions. The remaining reads served as the set of putative chimeric reads and were mapped to our candidate chimeras. Selection of candidate chimeras required that an RNA-seq read map precisely to the chimeric junction, with at least six nucleotides on each side of the junction, with no more than three mismatches (32). Finally, 192 human chimeric transcripts were confirmed by at least two RNA-seq reads covering the gene–gene junction site. Based on this RNA-sequencing analysis, the ChiTaRS database contains information regarding the number of reads across the chimera junction, its tissue specificity and the abundance of a given chimera in human tissues (see 'Full Collection & Search').

Cancer-associated chimeric transcripts

The human data set of chimeric transcripts includes chromosomal fusions found in cancers that we extracted from the TICdb (50), dbCrid (51), ChimerDB 2.0 (47) and Mitelman (56) databases. The chimeric transcripts collected in our database are the result of chromosomal translocations, insertions, deletions, inversions, ring chromosomes, derivatives and many others (see 'Breakpoints') (50,51,56). The manual inspection of >7000 (3343 unique) articles was applied to confirm the correspondence between the fusion event, disease and the two genes incorporated into the chimeras. Thus, the ChiTaRS database is composed of 1892 fusions involving >1000 unique genes (see 'Breakpoints' and Figure 1) with cross-links to the chimeric ESTs. The database incorporates also the published read-through and trans-splicing fusions (1), which can be found explicitly under the 'Full Collection & Search' page (use a check-box for 'Published Fusions'). To the best of our knowledge, ChiTaRS is the first catalog that enables cross-referencing between chimeric transcripts found in GenBank (44), relevant Pubmed articles regarding putative breakpoints, the two genes involved and the 'chimeric' RNA-seq reads covering the chimeric junctions in a specific tissue or a cell type. For example, there is a chromosomal translocation t(10;11)(p13;q14), which creates a fusion between the PICALM and MLLT10 genes, characteristic of hematological malignancies. The translocation described corresponds to the chimeric transcript found in our database, ESTid = 'EF051633'. Interestingly, searching ChiTaRS with this chimeric RNA transcript revealed that this chimera was also expressed in a female patient with chronic obstructive lung disease, according to the Human Body Map 2.0 data, with the expression level of 0.33 reads per kilobase per million reads (RPKM) (1–2 transcript per cell in average) (see 'Breakpoints' and Figure 1).

Features of the ChiTaRS database

The search options

The ChiTaRS database is accessible at its home page: http://chitars.bioinfo.cnio.es. The search (see the 'Full Collection & Search' page and Figure 2) can be performed using ESTid ('ChimeraID'), the names of the genes participating in the chimeras ('Gene Name', e.g. LMNA, DDX5), a sequence identity score ('Identity', e.g. 100, 95), a tissue type ('Tissue Name', e.g. lung), gene synonyms ('Gene Synonym') or a keyword ('Keyword', e.g. RARA). The 'Full Collection' can be obtained using the 'Full Collection' option and clicking on 'Search'. All 16 262 entries in the databases for human, mouse and fruit fly are listed together (Figure 2).

The search results page shows all the relevant instances associated with the chimeric transcripts available, the RNA-seq data of the mapping to the chimeric junction site (19), the level of transcript expression and the cancer breakpoints (see 'pop-ups' windows clicking on the 'RNA-seq' column in the 'Full Collection' and Figure 2). It contains detailed information about the identifier and the link to the corresponding GenBank entry, the junction site, the gene names and the identity of the two genes incorporated into the chimera. The chimeras can be visualized through as splice graphs (see description given later in the text). Together with the two genes and the disease information, the table of fusion transcripts includes general links to relevant resources, such as the Entrez Gene, GenBank (44), the Mitelman database (56), TICdb (50), dbCrid (51) and PubMed references. The search results can be saved as a tab-delimited text file using the 'Get Results as Text' button (up to 100 sequences).

In addition, 'Junction Search' provides the option to screen through the list of RNA-seq reads found at the chimeras' junction sites (19) to identify putative junction sites in novel sequences provided by a user. The 'Junction Search' is available for all three organisms in the database, and both the transcript sequence and the GenBank accession number can be used as inputs. The search is an automatic procedure that identifies a junction site in the transcript entered by a user and that aligns the previously

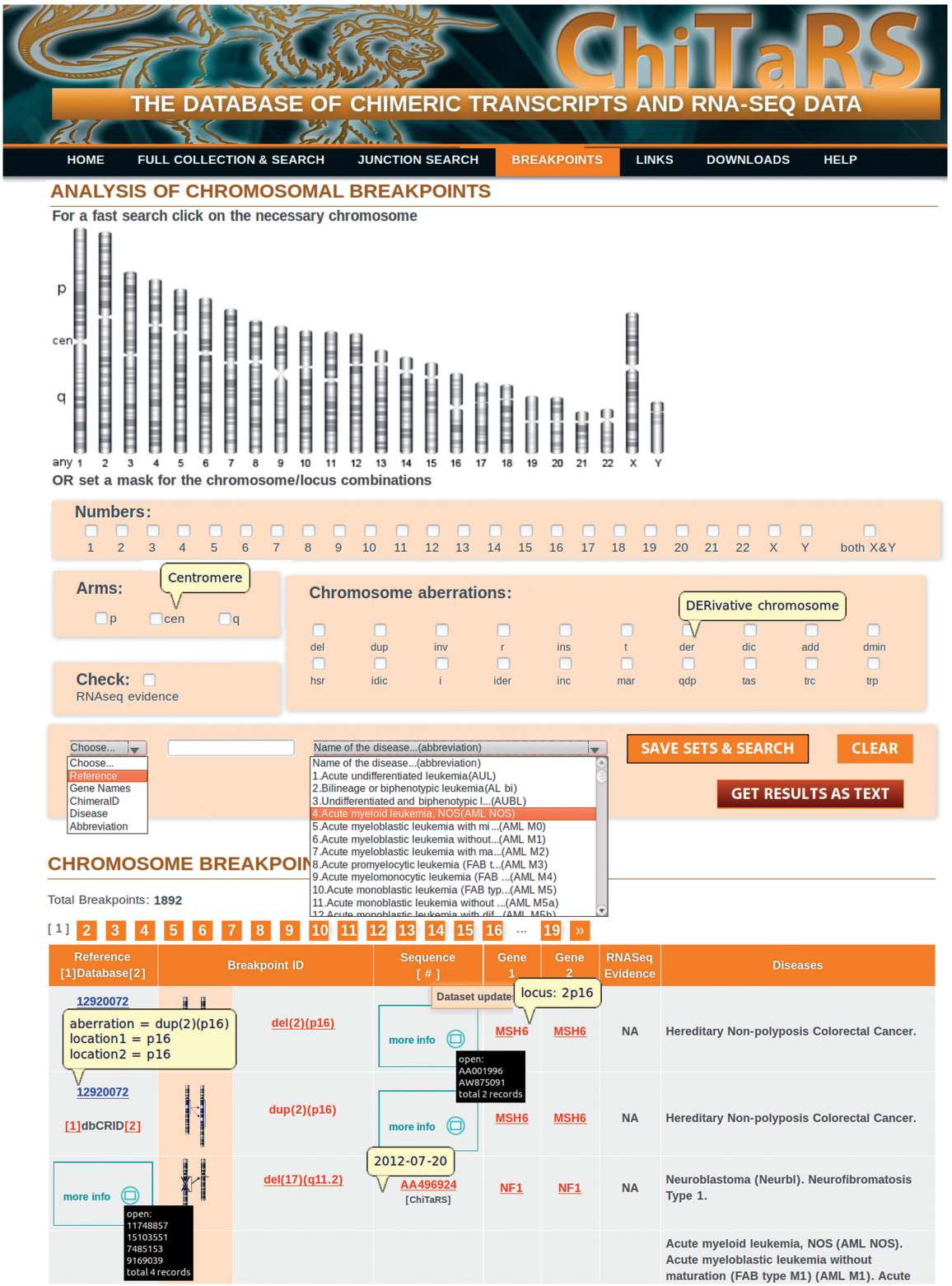

Figure 1. The ChiTaRS breakpoints collection page. The breakpoints collection ('Breakpoints') includes ~2000 human cancer breakpoints with the links to TICdb (50), dbCrid (51), ChimerDB 2.0 (47) and the Mitelman database (38,56). The search can be performed on the 'Breakpoints' page by a PubMed 'Reference', a 'Gene' name, ESTid ('ChimeraID'), a 'Disease' and a type of 'Chromosomal Aberrations'. The information for the search is recognized automatically between 'Reference', 'Gene', 'ChimeraID' or 'Disease'. A specific combination of chromosomes, arms and the locus can be used as a 'Search' option as well. Finally, the RNA-sequencing results are presented by clicking on 'RNA-seq' and 'Save sets and Search'.

Figure 2. The ChiTaRS full collection page. The full collection ('Full Collection & Search') consists of 16 262 transcripts in human (*H. Sapiens*), mouse (*M. Musculus*) and fruit fly (*D. Melanogaster*). The search can be performed using ESTid ('ChimeraID'), the names of the genes participating in the chimeras ('Gene Name', e.g. LMNA, DDX5), a sequence identity score ('Identity', e.g. 100, 95), a tissue type ('Tissue Name', e.g. lung), gene synonyms ('Gene Synonym') or a keyword ('Keyword', e.g. RARA). The RNA-sequencing results and the breakpoints can be extracted by clicking on the corresponding check-boxes and then 'Search'.

found 'chimeric' RNA-seq reads to this junction site. This special feature of ChiTaRS allows users to identify to what extent their chimeric transcripts are similar to those for which there is RNA-seq data in the database. It is essential for scientists to be able to analyze their chimeras in the complex setting of a large high-throughput data set and with multiple sequences. In the 'Downloads' section, we provide all the unmapped reads for the different RNA-seq data sets for three organisms. These data sets enable users to search for the junction coverage among other available chimeric transcripts in the different databases.

Unique gene names

The distinct aliases used for unique gene names represent one of the main problems when dealing with different

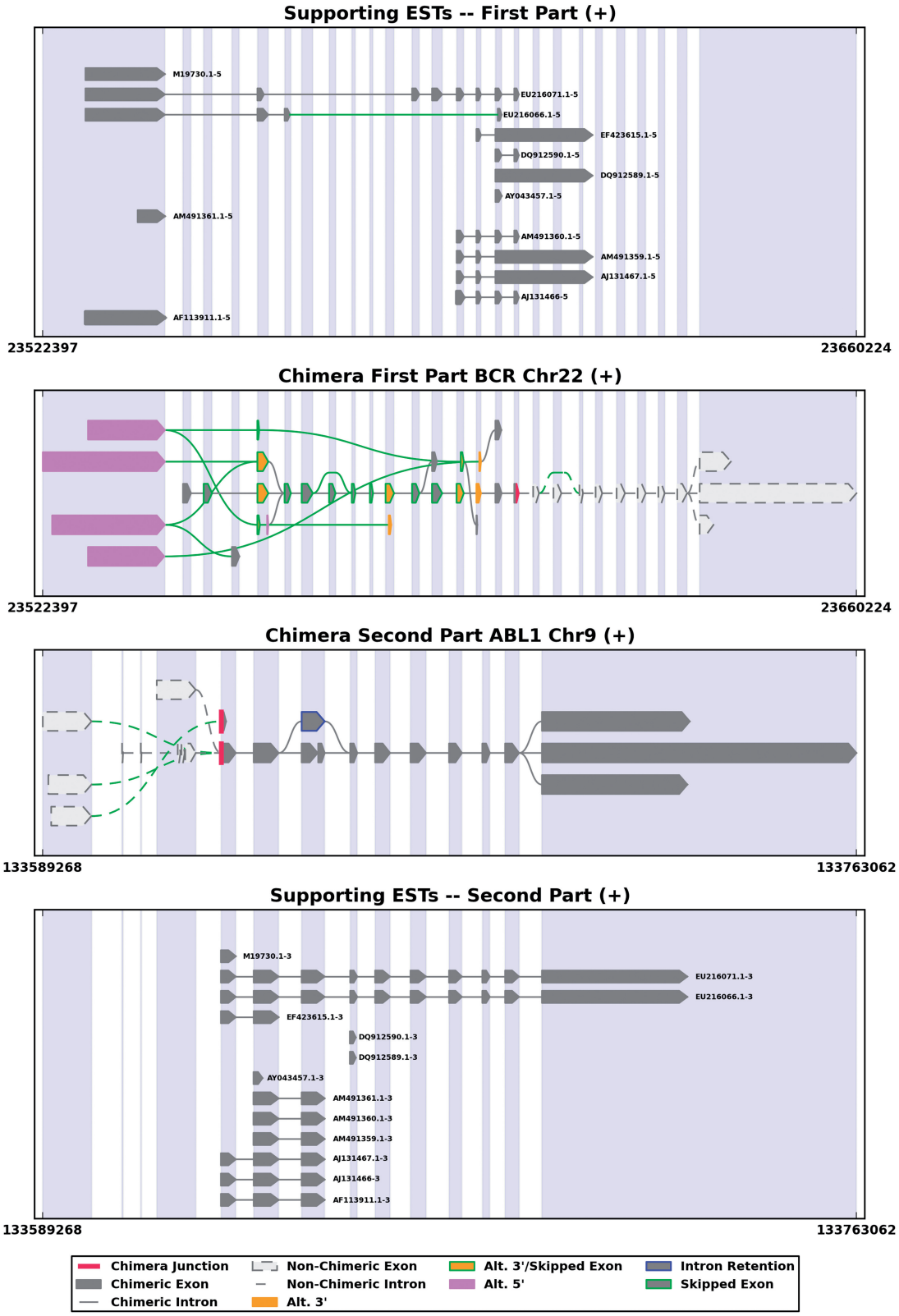

Figure 3. An example of 15 chimeric transcripts involving the BCR and ABL1 genes (ESTids = 'M19730.1', 'DQ912590.1', 'AM491360.1', 'EF423615.1', 'AM491359.1', 'DQ912589.1', 'AF113911.1', 'AY043457.1', 'M25946.1', 'AY789120.1', 'AM491361.1', 'EU216071.1', 'EU216066.1', 'AJ131467.1' and 'AJ131466.1'), presented by SpliceGrapher, with a consistent junction awarded a ranking of 5. This figure depicts the known splicing patterns for the two genes involved (middle two panels) in the chimera along with the 15 ESTs found in the ChiTaRS database that provide evidence for the chimera (top and bottom panels). Exons that participate in the chimera are highlighted in dark grey, and the location of the chimeric junction is highlighted in red. To score the junction consistency, we selected all 15 chimeras mapping to this gene pair. For each chimera, we have at our disposal the genomic location of its junction: end (gene 1) and start (gene 2). We calculated a distance between all pairs of chimeras based on

(continued)

gene, protein and transcript databases, which may represent a source of duplication in the databases. In ChiTaRS, we use a specific table to map the synonymous gene names to a unique record, using the NCBI Entrez gene name as a key. We have currently performed four updates to the ChiTaRS database after manual verification of the entries and cancer breakpoints. Each update is verified automatically for the synonymous gene name so that it is unique for both genes incorporated into the chimeras. Thus, all entries currently appearing in ChiTaRS have unified gene names, and as a result, searches can be performed based on gene names and synonyms (under 'Full Collection & Search', Figure 2).

Ranking of chimeric junction consistency

One of the key novelties of our database is the calculation and ranking of chimeric junction consistency. The ChiTaRS database contains transcripts that are chimeras of two genes, and in some cases, there is evidence these two genes may participate in many chimeras. The junction consistency ranking is a measure of how many times the same junction between the same genes has been found in chimeric transcripts. Thus, if the junction site is at the same genomic location of two genes incorporated in chimeras with a difference of no more than 1000 nt (an empirical number, can be changed in the 'Search' options), the junction rank is high (for more details, see Figure 3). The junction consistency in ChiTaRS is a particularly important experimental feature that may be of interest to verify the existence of highly ranked chimeras in cells by polymerase chain recation, reverse transcriptase-quantitative polymerase chain reaction or other techniques, thereby reducing the chance of dealing with chimeras that are mere artifacts.

Visualization of chimeras by SpliceGrapher

A bonus feature of ChiTaRS is the visualization of chimeric transcripts, and their genomic context, including the junction site. The visualization figures were produced using the SpliceGrapher package, which was designed to predict splice graphs for a gene by combining evidence from RNA-Seq data, annotated gene models and EST alignments (52). To produce splice graphs for chimeras, we first used GMAP (57) to align all available chimeric sequences to their reference genome (*Homo sapiens* version GGRCh37.63, *Drosophila melanogaster* version BDGP R5/dm3 and *Mus musculus* version NCBI37/mm9), and subsequently, SpliceGrapher was used to convert the resulting alignments into splice graphs (52). Finally, we used SpliceGrapher's visualization modules to integrate the ESTs and gene models into figures that illustrate chimeric splicing. Each figure shows how the ESTs align across two genes, making it possible to envisage the potential transcripts that could arise from each chimera (Figure 3).

The human and mouse chimeras

The ChiTaRS database provides evidence of chimeric transcripts and their mapping by the RNA-seq reads from three higher eukaryotes: human, mouse and fruit fly. The database is very robust and allows investigating the transcripts that incorporate the same orthologous genes in different organisms. An interesting example is the human chimera, ChimeraID = 'AW882230', and mouse chimera, ChimeraID = 'CF577921'. These chimeras both incorporate the PTMS gene (parathymosin, which may mediate the immune function) and are confirmed by RNA-seq reads in each organism. Therefore, these RNA-seq data support that the ability to form chimeras is a conserved feature of genomic loci. ChiTaRS takes the first step in exploring this premise because one of its main future goals is related to the study of the evolution of chimeric transcripts.

The 'Contact Us' webpage

The 'Contact Us' page describes a way to submit new chimeras and fusion transcripts, which have been detected by other groups, published, or found using alternative software or data sets. All requests to include data in the ChiTaRS database will be inspected and verified manually before uploading.

Downloads

The ChiTaRS database not only provides extended 'Search' options, but also offers the possibility to download all the database tables and the data sets in a very user-friendly manner. The full human, mouse and fruit fly collections include information about the two genes incorporated into the chimeras, the sequence identity and the positions of the junction sites. In addition, the freely available RNA-seq results, all the unmapped RNA-seq reads and mass spectrometry results are downloadable for each organism. For easy access to the most important fusions, we produced separate files for the published fusions (1) as well as the fusions identified in a prostate cancer by high-throughput RNA sequencing (23,24).

Figure 3. Continued

these coordinates. The distance simply corresponds to the difference between the two starts of gene 1 and the difference between the two ends of gene 2. Then we selected the chimera with shortest distance to all others as the reference chimera. If another chimera of the same gene 1 and gene 2 had a distance of <1000 nt to the reference chimera, we decided that the junction is consistent and incremented the rank by 1. In the special case where two chimeras had strictly the same mapping positions, we selected only one, assuming that the duplication could be due to artifacts. In the example from the figure, the reference chimera is EU216071.1. Twelve chimeric ESTs among 15 (except chimeras 'AM491361.1', 'AF113911.1', 'M19730.1') are consistent with the junction site of EU216071.1; the rank of these 12 chimeras is 5. The junction consistency and rank may show that potential breakpoints are not artifacts, and indeed, the BCR and ABL1 chimeras have a breakpoint for the Philadelphia translocation t(9;22)(q34;q11) in chronic myelogenous leukemia.

CONCLUSIONS AND FUTURE PLANS

ChiTaRS is an extended database of chimeric transcripts selected from GenBank from three organisms: human, mouse and fruit fly. The database features, such as the junction consistency and the ranking, allow rapid discovery of genes from different chimeras and of chimeras that share the same junction site. In addition, the chimeras derived from the three organisms provide an evolutionary tool to study chimeric transcripts across different organisms that involve the same genes. The RNA-Seq data should serve as a basis for further experimental confirmation of candidate chimeric transcripts. Moreover, the expression level of transcripts, as obtained from RNA-seq reads in different organisms and tissues, offers important information regarding the expression of chimeric transcripts, in particular, tissue specificity and function. Our ChiTaRS database is already of great use for experimental and evolutionary studies of chimeric transcripts, and for the annotation of chimeras in the International Cancer Genome Consortium (ICGC) project studing Chronic Lymphocytic Leukemia (CLL) project [in collaboration with the ICGC consortium (26,58)].

In summary, the ChiTaRS database encompasses all chimeric transcripts confirmed in humans and potentially translated into chimeric proteins (19). Our prediction is that the functions of chimeric proteins are substantially different from those of the original native proteins. Indeed, chimeric proteins sometimes contain different protein domains (20), or they are found in distinct cellular compartments or specific tissues associated with disease or cancer. We intend to continue expanding and annotating the ChiTaRS database with chimeric transcripts confirmed by RNA-seq reads and through the existence of the corresponding chimeric proteins, the latter preferably confirmed by mass spectrometry experiments. Our database should prove useful to biologists characterizing normal and cancer-associated chimeric transcripts and their corresponding proteins, and more generally, to researchers interested in gene expression and evolution, both physiological and pathological.

ACKNOWLEDGEMENTS

The authors thank Begoña Aguado, Alberto Rastrojo, Jaime Prilusky, Roderic Guigo and David Pisano for valuable discussions. The authors also thank authors of ChimerDB, dbCrid TICdb and the Mitelman database of the Cancer Aberrations for making available many human chimeric transcripts and cancer breakpoints and MPLabs LTD for a graphical design of the ChiTaRS home page.

FUNDING

Miguel Servet (FIS) grant (to M.F-M.); Obra Social laCaixa grant (to K.I.); National Science Foundation ABI [0743097 to A.B-H. and M.R.]. Funding for open access charge: NHGRI-NIH ENCODE grant [HG00455-04]; Blueprint European Union project [282510]; Spanish Government grant [BIO2007-66855]; Spanish National Bioinformatics Institute (INB-ISCIII), Genecode/ENCODE NHGRI-NIH grant [HG00455-04]; Plan Cancer 2009-2013, ERC Advanced Grant Sisyphe, Investissements d'avenir en Bioinformatique.

Conflict of interest statement. None declared.

REFERENCES

1. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigó,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
2. Guigó,R., Flicek,P., Abril,J.F., Reymond,A., Lagarde,J., Denoeud,F., Antonarakis,S., Ashburner,M., Bajic,V.B., Birney,E. *et al.* (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.*, **7(Suppl. 1)**, S2.1–S2.31.
3. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
4. Griffin,T.J., Gygi,S.P., Ideker,T., Rist,B., Eng,J., Hood,L. and Aebersold,R. (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics*, **1**, 323–333.
5. Velculescu,V.E., Zhang,L., Zhou,W., Vogelstein,J., Basrai,M.A., Bassett,D.E., Hieter,P., Vogelstein,B. and Kinzler,K.W. (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.
6. Cirulli,E.T., Singh,A., Shianna,K.V., Ge,D., Smith,J.P., Maia,J.M., Heinzen,E.L., Goedert,J.J., Goldstein,D.B. and Center for HIV/AIDS Vaccine Immunology (CHAVI) (2010) Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biol.*, 11, R57.
7. Finta,C. and Zaphiropoulos,P.G. (2002) Intergenic mRNA molecules resulting from trans-splicing. *J. Biol. Chem.*, **277**, 5882–5890.
8. Kapranov,P., Drenkow,J., Cheng,J., Long,J., Helt,G., Dike,S. and Gingeras,T.R. (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.*, **15**, 987–997.
9. Djebali,S., Lagarde,J., Kapranov,P., Lacroix,V., Borel,C., Mudge,J.M., Howald,C., Foissac,S., Ucla,C., Chrast,J. *et al.* (2012) Evidence for transcript networks composed of chimeric RNAs in human cells. *PLoS One*, **7**, e28213.
10. Di Segni,G., Gastaldi,S. and Tocchini-Valentini,G.P. (2008) Cis- and trans-splicing of mRNAs mediated by tRNA sequences in eukaryotic cells. *Proc. Natl Acad. Sci. USA*, **105**, 6864–6869.
11. Akiva,P., Toporik,A., Edelheit,S., Peretz,Y., Diber,A., Shemesh,R., Novik,A. and Sorek,R. (2006) Transcription-mediated gene fusion in the human genome. *Genome Res.*, **16**, 30–36.
12. Parra,G., Reymond,A., Dabbouseh,N., Dermitzakis,E.T., Castelo,R., Thomson,T.M., Antonarakis,S.E. and Guigó,R. (2006) Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.*, **16**, 37–44.
13. 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.
14. 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.
15. Ortiz de Mendíbil,I., Vizmanos,J.L. and Novo,F.J. (2009) Signatures of selection in fusion transcripts resulting from chromosomal translocations in human cancer. *PLoS One*, **4**, e4805.
16. Li,H., Wang,J., Ma,X. and Sklar,J. (2009) Gene fusions and RNA trans-splicing in normal and neoplastic human cells. *Cell Cycle*, **8**, 218–222.

17. Li,H., Wang,J., Mor,G. and Sklar,J. (2008) A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science*, **321**, 1357–1361.
18. Edgren,H., Murumagi,A., Kangaspeska,S., Nicorici,D., Hongisto,V., Kleivi,K., Rye,I.H., Nyberg,S., Wolf,M., Borresen-Dale,A.L. *et al.* (2011) Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol.*, **12**, R6.
19. Frenkel-Morgenstern,M., Lacroix,V., Ezkurdia,I., Levin,Y., Gabashvili,A., Prilusky,J., Del Pozo,A., Tress,M., Johnson,R., Guigo,R. *et al.* (2012) Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res.*, **22**, 1231–1242.
20. Frenkel-Morgenstern,M. and Valencia,A. (2012) Novel domain combinations in proteins encoded by chimeric transcripts. *Bioinformatics*, **28**, i67–i74.
21. Asmann,Y.W., Necela,B.M., Kalari,K.R., Hossain,A., Baker,T.R., Carr,J.M., Davis,C., Getz,J.E., Hostetter,G., Li,X. *et al.* (2012) Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. *Cancer Res.*, **72**, 1921–1928.
22. Gingeras,T.R. (2009) Implications of chimaeric non-co-linear transcripts. *Nature*, **461**, 206–211.
23. 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.
24. Maher,C.A., Palanisamy,N., Brenner,J.C., Cao,X., Kalyana-Sundaram,S., Luo,S., Khrebtukova,I., Barrette,T.R., Grasso,C., Yu,J. *et al.* (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA*, **106**, 12353–12358.
25. Houseley,J. and Tollervey,D. (2010) Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One*, **5**, e12271.
26. Quesada,V., Conde,L., Villamor,N., Ordóñez,G.R., Jares,P., Bassaganyas,L., Ramsay,A.J., Beà,S., Pinyol,M., Martínez-Trillos,A. *et al.* (2012) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.*, **44**, 47–52.
27. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **39**, D945–D950.
28. Roeder,I., Horn,M., Glauche,I., Hochhaus,A., Mueller,M.C. and Loeffler,M. (2006) Dynamic modeling of imatinib-treated chronic myeloid leukemia: functional insights and clinical implications. *Nat. Med.*, **12**, 1181–1184.
29. Tang,M., Foo,J., Gonen,M., Guilhot,J., Mahon,F.X. and Michor,F. (2012) Selection pressure exerted by imatinib therapy leads to disparate outcomes of imatinib discontinuation trials. *Haematologica*, **97**, 1553–1561.
30. Sutherland,G.T., Janitz,M. and Kril,J.J. (2011) Understanding the pathogenesis of Alzheimer's disease: will RNA-Seq realize the promise of transcriptomics? *J. Neurochem.*, **116**, 937–946.
31. Hall,P.A., Reis-Filho,J.S., Tomlinson,I.P. and Poulsom,R. (2010) An introduction to genes, genomes and disease. *J. Pathol.*, **220**, 109–113.
32. Aparicio,S.A., Caldas,C. and Ponder,B. (2000) Does massively parallel transcriptome analysis signify the end of cancer histopathology as we know it? *Genome Biol.*, **1**, REVIEWS1021.
33. Costa,V., Angelini,C., De Feis,I. and Ciccodicola,A. (2010) Uncovering the complexity of transcriptomes with RNA-Seq. *J. Biomed. Biotechnol.*, **2010**, 853916.
34. Guffanti,A., Iacono,M., Pelucchi,P., Kim,N., Soldà,G., Croft,L.J., Taft,R.J., Rizzi,E., Askarian-Amiri,M., Bonnal,R.J. *et al.* (2009) A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics*, **10**, 163.
35. Choi,Y.L., Takeuchi,K., Soda,M., Inamura,K., Togashi,Y., Hatano,S., Enomoto,M., Hamada,T., Haruta,H., Watanabe,H. *et al.* (2008) Identification of novel isoforms of the EML4-ALK transforming gene in non-small cell lung cancer. *Cancer Res.*, **68**, 4971–4976.
36. 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.
37. Wang,X.S., Prensner,J.R., Chen,G., Cao,Q., Han,B., Dhanasekaran,S.M., Ponnala,R., Cao,X., Varambally,S., Thomas,D.G. *et al.* (2009) An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. *Nat. Biotechnol.*, **27**, 1005–1011.
38. Kannan,K., Wang,L., Wang,J., Ittmann,M.M., Li,W. and Yen,L. (2011) Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. *Proc. Natl Acad. Sci. USA*, **108**, 9172–9177.
39. Herai,R.H. and Yamagishi,M.E. (2010) Detection of human interchromosomal trans-splicing in sequence databanks. *Brief. Bioinform.*, **11**, 198–209.
40. Mitelman,F., Mertens,F. and Johansson,B. (2005) Prevalence estimates of recurrent balanced cytogenetic aberrations and gene fusions in unselected patients with neoplastic disorders. *Genes Chromosomes Cancer*, **43**, 350–366.
41. 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.
42. 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.
43. 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.
44. Benson,D.A., Karsch-Mizrachi,I., Clark,K., Lipman,D.J., Ostell,J. and Sayers,E.W. (2012) GenBank. *Nucleic Acids Res.*, **40**, D48–D53.
45. 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.
46. Li,X., Zhao,L., Jiang,H. and Wang,W. (2009) Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes. *J. Mol. Evol.*, **68**, 56–65.
47. Kim,P., Yoon,S., Kim,N., Lee,S., Ko,M., Lee,H., Kang,H. and Kim,J. (2010) ChimerDB 2.0–a knowledgebase for fusion genes updated. *Nucleic Acids Res.*, **38**, D81–D85.
48. Kim,N., Kim,P., Nam,S., Shin,S. and Lee,S. (2006) ChimerDB–a knowledgebase for fusion sequences. *Nucleic Acids Res.*, **34**, D21–D24.
49. 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.
50. Novo,F.J., de Mendíbil,I.O. and Vizmanos,J.L. (2007) TICdb: a collection of gene-mapped translocation breakpoints in cancer. *BMC Genomics*, **8**, 33.
51. 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.
52. Rogers,M.F., Thomas,J., Reddy,A.S. and Ben-Hur,A. (2012) SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol.*, **13**, R4.
53. Dreszer,T.R., Karolchik,D., Zweig,A.S., Hinrichs,A.S., Raney,B.J., Kuhn,R.M., Meyer,L.R., Wong,M., Sloan,C.A., Rosenbloom,K.R. *et al.* (2012) The UCSC Genome browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.
54. Kent,W.J. (2002) BLAT- the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
55. Robertson,G., Schein,J., Chiu,R., Corbett,R., Field,M., Jackman,S.D., Mungall,K., Lee,S., Okada,H.M., Qian,J.Q. *et al.* (2010) De novo assembly and analysis of RNA-seq data. *Nat. Methods*, **7**, 909–912.

56. Mitelman,F., Johansson,B. and Mertens,F. (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.
57. Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.
58. Puente,X.S., Pinyol,M., Quesada,V., Conde,L., Ordóñez,G.R., Villamor,N., Escaramis,G., Jares,P., Beà,S., González-Díaz,M. *et al.* (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, **475**, 101–105.