ChiTaRS 2.1—an improved database of the chimeric transcripts and RNA-seq data with novel sense–antisense chimeric RNA transcripts

Milana Frenkel-Morgenstern, Alessandro Gorohovski, Dunja Vucenovic, Lorena Maestre and Alfonso Valencia

1Structural Biology and BioComputing Program, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain. and 2Monoclonal Antibodies Unit, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain.

Received September 16, 2014; Revised November 4, 2014; Accepted November 4, 2014

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 2.1 database of chimeric transcripts and RNA-Seq data (http://chitars.bioinfo.cnio.es/) is the second version of the ChiTaRS database and includes improvements in content and functionality. Chimeras from eight organisms have been collated including novel sense–antisense (SAS) chimeras resulting from the slip-page of the sense and anti-sense intragenic regions. The new database version collects more than 29 000 chimeric transcripts and indicates the expression and tissue specificity for 333 entries confirmed by RNA-seq reads mapping the chimeric junction sites. User interface allows for rapid and easy analysis of evolutionary conservation of fusions, literature references and experimental data supporting fusions in different organisms. More than 1428 cancer breakpoints have been automatically collected from public databases and manually verified to identify their correct cross-references, genomic sequences and junction sites. As a result, the ChiTaRS 2.1 collection of chimeras from eight organisms and human cancer breakpoints extends our understanding of the evolution of chimeric transcripts in eukaryotes as well as their functional role in carcinogenic processes.

INTRODUCTION

Chimeric RNAs may be produced by the joining of exons from different genes either through a complex splicing process or as the result of chromosome rearrangement (1–23). Thus, two loci on different chromosomes may produce chimeras through a genomic rearrangement event or through trans-splicing (21,24). Additionally, read-through transcription of two adjacent genomic loci may result in chimaera synthesis (10,11,25–27). While many chimeras have been shown to be artifacts of the in vitro reverse transcription reaction (28–31), there is sufficient data demonstrating that some chimeras are translated into chimeric proteins (18). Here we establish an extended collection of putative chimeric transcripts whose existence are supported at different levels by experimental data, including tissue specific expression levels of chimeric RNAs and protein products (18,32).

Our ChiTaRS database of ‘Chimeric Transcripts and RNA-Seq data’ is a collection of chimeric transcripts identified by Expressed Sequence Tags (ESTs) and mRNAs from the GenBank (33), ChimerDB (26,34), dbCRID (35), TICdb (36) and other databases for humans, mouse and flies (37). Our pipeline for finding chimeric transcripts is shown on Supplementary Figure S1 (Supplementary Material). Here we present the updated ChiTaRS 2.1 database of more than 29 000 chimeric transcripts in eight organisms; the database incorporates major additions in content and functionality. The ChiTaRS database is currently used to study the identity and incidence of specific fusions of transcripts that may result in a chimeric RNA with novel biological function. In the original ChiTaRS database (32), there was some experimental data included, such as RNA-seq, and mass spectrometry identification of peptides formed by the translation of the chimeric RNA transcripts. In the current version, we extend the experimental data evidence and...
Figure 1. A putative chimera composed of RAD9A (RAD9A homolog A) and PPP1CA (protein phosphatase 1). (A) A chimera found among human ESTs. (B) A mouse chimera.

the organism coverage by chimeras from eight organisms: *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Rattus norvegicus*, *Bos taurus*, *Sus scrofa*, *Danio rerio* and *Saccharomyces cerevisiae*. Furthermore, the new database version includes a novel type of particularly interesting sense–antisense chimeric transcripts, together with their experimental confirmation by the RNA-seq reads.

Cancer fusions resulting from chromosomal translocations, deletions or inversions are well characterized in cancer (38–48). Fusion proteins increase the complexity of the proteome in many types of cancers with the production of novel proteins (18). In other cases they can produce non-coding regulatory RNAs or interfere with other genomic regions (39–43). Although gene fusions can be detected by the RNA-seq technique, for many fusions the correct junction sequences have yet to be determined, and there are many inconsistencies between different databases, including the corresponding annotations in GenBank (33). Therefore, we have initiated a curation effort to collate information on cancer fusions from GenBank (33), UniProt (49), the Mitelman database (47,50) and the Atlas of Genetics and Cytogenetics in Oncology and Haematology (http://atlasgeneticsoncology.org/) and to run our chimeric transcript analytical methodology in order to determine the correct junction sites of these fusions. First, we automatically collected all the fusions from UniProt including their description and corresponding GenBank-ids and then we have verified those entries manually in order to find cancer breakpoints references in GenBank and other database. Next, we run our automatic procedure to identify chimeric junction sites for all the entries using the genomic sequence of the breakpoints. Finally, we produced the manual verification and identification of the junction sites for all 610 breakpoints from the Mitelman collection having the GenBank-id and for all 818 breakpoints without GenBank-id. Thus, ChiTaRS-2.1 incorporates the largest collection of cancer breakpoints and their junction sequences and it includes 1428 (about 800 new) annotated cancer fusions in different types of cancers. We added the corresponding fusion junction sites and the genomic sequences for all the breakpoints (See ‘Breakpoints’ and ‘Downloads’).

In ChiTaRS-2.1, we also collected an additional type of chimeric RNA transcripts, the ‘read-through’ chimera, that begins upstream of gene 1 and ends at the termination site of adjacent gene 2. Such chimeras have been detected in various cancer and normal cells. Read-through chimeric transcripts are not included in other datasets like ChimerDB (26,34), TICdb (36) or dbCRID (35), and are thus unique to ChiTaRS-2.1. To view ‘Read-through’ chimeras we added a check-box on the ‘Full Collection’ page. All the entries in ChiTaRS-2.1 can be accessed from the UniProt Knowledgebase system (UniProtKB) that collates information on individual proteins from laboratories world-wide, including 2870 fusions proteins (and parental proteins) listed in UniProt (51). Chimeric RNAs and proteins have become a powerful tool for researchers over the past few years since
Figure 2. The most frequent junction motifs of SAS chimeras are incorporate palindromic sequences. (A) Two palindromic motifs found for human SAS chimeras. (B) Motifs of the mouse SAS chimeras. (C) Motifs of the fly SAS chimeras.
Figure 3. A new interface with enhanced query capacity and support information has been added to the ChiTaRS-2.1 database.

they can be used as cancer markers as well as putative targets for the development of new drugs. Thus, the current ChiTaRS-2.1 database represents a basic starting point for identifying cancer fusions, for studying chimeric transcripts, for analyzing New-Generation-Sequencing results and for investigating the biological processes underlying the phenomenon of cancer fusions.

IMPROVEMENTS

Ten updates and improvements to the content and functionality of ChiTaRS are summarized in Table 1. Major improvements include: addition of chimeric transcripts from eight organisms, to the ability to compare and analyze chimeras from different organisms, links to PubMed references by means of an iHop online text-mining routine and a new category of chimeric transcripts: the sense–antisense chimeras.

Updated database content

In the 2014 update, 29,164 chimeras and 1,428 cancer breakpoints have been collected from eight organisms. The number of chimeras identified in each species is presented in Table 2. For all the 1428 cancer breakpoints produced by 1090 human genes, we have performed manual confirma-
tion of their veracity using sequence information and experimental data from 6941 articles. In addition, 333 chimeric transcripts and their junction sites were confirmed by in-house RNA-seq including our previous results (19). Finally, four chimeric transcripts for the ATP1A1 gene, three from human and one from mouse, were extensively verified by means of RT-qPCR, PCR, cloning and sequencing procedures, in order to confirm their expression levels in six tissue samples from two organisms (human and mouse) (Supplementary Figure S2, Supplementary Material). Therefore, the ChiTaRS 2014 update includes experimental support for 337 transcripts, 1.9× more than in the original ChiTaRS database, which had support for 175 chimeras (Table 1).

We identified chimeric transcripts from the GenBank (33) collection of ESTs and mRNAs for *H. sapiens* (UCSC reference genome: GRCh37/hg19), *M. musculus* (NCBI37/mm9) and *D. melanogaster* (BDGP R5/dm3) *R. norvegicus* (RGSC Rnor_6.0/rn6), *B. taurus* (Baylor College of Medicine HGSC Btau4.6.1/bosTau7), *D. rerio* (Sanger Institute Zv9/danRer7), *S. cerevisiae* (SGD April 2011 sequence/sacCer3) and *S. scrofa* (Broad/Pig3). The ESTs and mRNA sequences were mapped to their corresponding reference genomic sequences using the UCSC BLAT program (52). We included a chimera if the first and the second sequence tracts of the chimera had a minimum identity of 95%, a minimum length of 50 nt, and if these two tracts could not be mapped linearly to the reference genome.

In ChiTaRS-2.1, we have added an analysis and comparison of the junction sites, rank and consistency between different chimeric transcripts (18) in all eight studied organisms. This new feature provides users the ability to study the evolution of chimeric transcripts and conservation of the junction sites for any chimera, including the 2337 chimeras conserved between human and mouse. A new improved interface allows users to ‘Compare and Analyze’ chimeras from different organisms (see a link at the Top Menu of the database webpage: http://chitars.bioinfo.cnio.es/). To illustrate the power of this new utility, we applied it to identify a putative chimera composed of RAD9A (RAD9A homolog A) and PPP1CA (protein phosphatase 1), present in both human and mouse ESTs (Figure 1A and B). In human, this chimera is encoded by the same strand as a read-through of the RAD9A and PPP1CA genes (Figure 1A). However, the transcript in mouse may be considered as sense–antisense (‘SAS’) chimera (see below), since the two genes incorporated in the chimeras are encoded by the opposite strands of the overlapping genes (Figure 1B). ChiTaRS-2.1 has the ‘Junction Search’ feature that may be applied for the junction sites analysis of all eight organisms using the alignment and the E-value found by the FASTA program (53). To conclude, our database provides unexplored datasets of evolutionarily conserved chimeric transcripts in eukaryotes and enables the study of their functional role in cellular processes.

**Table 1.** Major improvements as provided in the ChiTaRS-2.1 database

| Features                        | ChiTaRS version 1.0 | ChiTaRS version 2.1 |
|---------------------------------|---------------------|---------------------|
| Species                         | 3 species           | 8 species           |
| Number of chimeric transcripts  | 16 261              | 29 164              |
| Chimeras validated by more than two | 175                | 337                 |
| RNA-seq reads spanning the junction site |                   |                     |
| Cancer breakpoints              | 1286                | 1428                |
| Manually verified breakpoints   | 456                 | 48 586              |
| UniProt cross-references        | NA                  | 2229                |
| Sense–antisense chimeras        | NA                  | 6044                |
| iHop cross-links                | NA                  | 48 586              |
| Comparison and analysis of species | Not Available      | Available           |
| SpliceGraphs                    | 8000                | 8232                |
quences to PER1 was confirmed in this case; it reduced the expression level of the WT-PER1 protein and affected the overall response of a patient to the chemotherapy drugs (55). Therefore, the SAS chimeras in ChiTaRS-2.1 is a unique collection that allows to study the effect of antisense transcripts in cancers. In ChiTaRS-2.1, there are 69 SAS chimeras confirmed by RNA-seq reads spanning the junction sites (see ‘Full Collection’).

### New RNA-seq evidence for the expression of chimeras

To establish the veracity of all the chimeric transcripts in ChiTaRS-2.1, we produced RNA-seq libraries of three human cancer cell lines: MCF7 (breast cancer), LNCAP (prostate cancer), VCAP (prostate cancer) and one fly cell line MBN (timorous blood *Drosophila* cell line). The datasets have 85 million (M) paired-end reads of 50 nt per sample. The reads mapping to the template chimeras was carried out following the previously described procedure (18). For the MCF7, LNCAP, VCAP and MBN cell lines, we required at least five RNA-seq reads covering the chimeric junction site with only a maximum of two mismatches allowed (Table 2). This requirement is more restricted than one used in our previous studies (18) in order to decrease a number of artifacts. As a result, we confirmed the presence of 333 chimeras: 297 in human, 8 in mouse, 28 in fly (see ‘Full Collection’). These 297 chimeras include 175 previously reported cases, 89 new ones expressed in MCF7, VCAP and LNCAP, and 69 SAS chimeras confirmed by RNA-seq reads. Interestingly, an inter-chromosomal fusion, *NDUFAF2-MAST4*, in VCAP, identified previously by ChimeraScan (56), was identified in our sample, since we detected five junction-spanning paired-end reads for this chimera. Such examples in the database demonstrate that our methodology is sufficiently sensitive for the analysis of the expression of putative chimeras. We analyzed all the chimeras expressed in MCF7 (118 transcripts), finding that they include known cancer breakpoints, sense–antisense chimeric transcripts and read-through chimeras from our new database ChiTaRS-2.1 (see ‘Full Collection’). The chimeras are generally highly expressed in comparison to a normal breast tissue (Supplementary Material, Supplementary Figure S2, in reads assigned per kilobase of target per million mapped reads (RPKM), *P* < 0.05). As such, the new version of ChiTaRS contains the highest number of chimeric transcripts known today and the largest collection of experimental evidences for the expression of chimeras. All the datasets in ChiTaRS-2.1 can be retrieved from ‘Downloads’.

### Functionality improvements

To improve the data access and analyses of the information on chimeric transcripts contained in ChiTaRS, a new interface with enhanced query capacity and support information have been added (Figure 3). Every ChiTaRS-2.1 entry is associated to a genomic position in the UCSC browser, which appears in a new pop-up window and includes downloadable files incorporating all the transcription start/stop sites, the genomic, chromosomal and strand location (Figure 3). Publications associated with each of the two genes in every chimera can be easily accessed using an automated PubMed search, and all the retrieved references can be downloaded using the ‘Save Text’ option (See ‘Full Collection’ and Figure 3). To improve the visual association of chimeric transcripts with gene function, we have added a link to the iHOP family (57–60) of web services (www.iHop-net.org/) for every gene in the ChiTaRS 2.1 database. The *iHOP, Information Hyperlinked over Proteins* (57), engine provides information on gene function, potential gene–gene relation in networks of genes, as an intuitive way of screening the millions of abstracts in PubMed for relevant publications (Figure 3). This improvement provides users with an easy means of exploring and combining information for each parental gene of a chimera.

### CONCLUSIONS AND PERSPECTIVES

The current update of the ChiTaRS-2.1 database represents a 1.9-fold increase of chimeric transcripts as compared to the initial ChiTaRS release, and includes a significant extension of specific research-oriented features. ChiTaRS-2.1 provides extensive experimental evidence for chimeras and cancer fusions, and this information can be considered instrumental for planning new experiments or for the analysis of large scale RNA-seq experiments. The database will be updated every six months to include the growing number of chimeras published. International projects like ICGC and TCGA will benefit from this database and on all incremental additions to the database, for improving the process of chimera identification and validation in cancer research. To conclude, the ChiTaRS-2.1 database is designed to advance the field of Cancer Research as well as our understanding of the phenomenon of chimeric transcripts and its evolution in eukaryotes.

### AVAILABILITY

The ChiTaRS-2.1 content will be continuously maintained and updated every six months. The database is now publicly accessible at http://chitars.bioinfo.cnio.es/ and the old version of the database is accessible at http://chitars-old.bioinfo.cnio.es/.

### Table 2. SAS chimeras identified in different organisms

| Species               | *H. sapiens* | *M. musculus* | *D. melanogaster* | *R. norvegicus* | *B. taurus* | *D. rerio* | *S. cerevisiae* | *S. scrofa* |
|-----------------------|--------------|---------------|-------------------|----------------|-------------|------------|----------------|-----------|
| Number of chimeric transcripts | 20 740       | 6224          | 2151              | 8              | 4           | 4          | 5              | 14        |
| Sense–antisense chimeras | 3998         | 1713          | 323               | 1              | 0           | 0          | 2              | 7         |

*Species* include *H. sapiens* (human), *M. musculus* (mouse), *D. melanogaster* (*Drosophila*), *R. norvegicus* (rat), *B. taurus* (bovine), *D. rerio* (zebrafish), *S. cerevisiae* (yeast) and *S. scrofa* (pig).
SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank the following contributors whose work makes the ChiTaRS-2.1 possible: MPLabs LTD for the website design and its support, Kovid BioAnalytics LTD for the manual verification of all the cancer breakpoints in ChiTaRS-2.1, the ACGT Inc. for the RT-qPCR, cloning and sequencing experiments, the Genomics Unit at CNIO for RNA extractions, library preparation and the RNA sequencing results. We thank our users for their consistent support and valuable feedback and our outstanding group for their priceless discussions and suggestions.

FUNDING

Miguel Servet (FIS: CP11/00294) [to M.F.-M. for staff scientists]. Funding for open access charge: NHGRI-ENCODE [HG00455-04]; Blueprint European Union project [282510]; Spanish Government [BIO2007-66855]; Spanish National Bioinformatics Institute (INB-ISCIII), Genecode/ENCODENHGRINCIIE [HG00455-04].

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), S1–S11.
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 Knizler, 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 (CHAVI), C.H.A.V.I. (2010) Screening the human exome: a comparison of large number of additional exons in ENCODE regions. Genome Res., 107, 12975–12979.
7. 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.
8. Denoed, F., Kapranov, P., Lacroix, V., Borel, C., Mudge, J.M., Howald, C., Foissac, S., Ucla, C., Christ, J. et al. (2012) Evidence for transcript networks composed of chimeric RNAs in human cells. PLoS One, 7, e28215.
9. Prakash, A., Tomazela, D.M., Frewen, B., Maclean, B., Merrihew, G., Peterman, S. and Maccoss, M.J. (2009) Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. J. Proteome Res., 8, 2733–2739.
10. 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.
11. Denoed, F., Kapranov, P., Ucla, C., Frankish, A., Castelo, R., Drenkow, J., Lagarde, J., Alioto, T., Manzano, C., Christ, J. et al. (2007) Evidence for transcript networks composed of chimeric RNAs in human cells. PLoS One, 7, e28215.
12. Prakash, A., Tomazela, D.M., Frewen, B., Maclean, B., Merrihew, G., Peterman, S. and Maccoss, M.J. (2009) Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. J. Proteome Res., 8, 2733–2739.
13. 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.
14. Prakash, A., Tomazela, D.M., Frewen, B., Maclean, B., Merrihew, G., Peterman, S. and Maccoss, M.J. (2009) Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. J. Proteome Res., 8, 2733–2739.
15. 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.
16. Denoed, F., Kapranov, P., Ucla, C., Frankish, A., Castelo, R., Drenkow, J., Lagarde, J., Alioto, T., Manzano, C., Christ, J. et al. (2007) Evidence for transcript networks composed of chimeric RNAs in human cells. PLoS One, 7, e28215.
17. Prakash, A., Tomazela, D.M., Frewen, B., Maclean, B., Merrihew, G., Peterman, S. and Maccoss, M.J. (2009) Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. J. Proteome Res., 8, 2733–2739.
18. 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.
19. Prakash, A., Tomazela, D.M., Frewen, B., Maclean, B., Merrihew, G., Peterman, S. and Maccoss, M.J. (2009) Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. J. Proteome Res., 8, 2733–2739.
20. 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.
21. Prakash, A., Tomazela, D.M., Frewen, B., Maclean, B., Merrihew, G., Peterman, S. and Maccoss, M.J. (2009) Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. J. Proteome Res., 8, 2733–2739.
22. 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.
23. Prakash, A., Tomazela, D.M., Frewen, B., Maclean, B., Merrihew, G., Peterman, S. and Maccoss, M.J. (2009) Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. J. Proteome Res., 8, 2733–2739.
24. Prakash, A., Tomazela, D.M., Frewen, B., Maclean, B., Merrihew, G., Peterman, S. and Maccoss, M.J. (2009) Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. J. Proteome Res., 8, 2733–2739.
