ChiTaRS-3.1—the enhanced chimeric transcripts and RNA-seq database matched with protein–protein interactions

Alessandro Gorohovski, Somnath Tagore, Vikrant Palande, Assaf Malka, Dorith Raviv-Shay and Milana Frenkel-Morgenstern*

Faculty of Medicine in Galilee, Bar-Ilan University, Henrietta Szold 8, Safed 13195, Israel

Received October 02, 2016; Revised October 26, 2016; Editorial Decision October 26, 2016; Accepted October 30, 2016

ABSTRACT

Discovery of chimeric RNAs, which are produced by chromosomal translocations as well as the joining of exons from different genes by trans-splicing, has added a new level of complexity to our study and understanding of the transcriptome. The enhanced ChiTaRS-3.1 database (http://chitars.md.biu.ac.il) is designed to make widely accessible a wealth of mined data on chimeric RNAs, with easy-to-use analytical tools built-in. The database comprises 34 922 chimeric transcripts along with 11 714 cancer breakpoints. In this latest version, we have included multiple cross-references to GeneCards, iHop, PubMed, NCBI, Ensembl, OMIM, RefSeq and the Mitelman collection for every entry in the ‘Full Collection’. In addition, for every chimera, we have added a predicted chimeric protein–protein interaction (ChiPPI) network, which allows for easy visualization of protein partners of both parental and fusion proteins for all human chimeras. The database contains a comprehensive annotation for 34 922 chimeric transcripts from eight organisms, and includes the manual annotation of 200 sense-antiSense (SaS) chimeras. The current improvements in the content and functionality to the ChiTaRS database make it a central resource for the study of chimeric transcripts and fusion proteins.

INTRODUCTION

Chimeric RNAs comprise sequences deriving from more than one transcription event. Fusion can occur at either the genomic level as the result of chromosomal rearrangement, or at the RNA level when two different transcripts are combined through a complex trans-splicing process (1–24). While many chimeric transcripts have been shown to be artifacts of in vitro reverse transcription reactions (25–32), recent studies clearly demonstrate that some (mostly cancer chimeric transcripts) are translated into chimeric proteins (11,16,18). Here, we expand our previously published collection of putative chimeric transcripts (ChiTaRS) that includes chimeras whose RNA expression levels have been verified by RNA-sequencing and whose translation into protein products has been shown previously by us, using mass-spectrometry analyses (33,34) by predicted protein–protein interaction networks.

Translation of chimeric transcripts into a fusion protein has been shown to dramatically alter the protein–protein interaction (PPI) networks of the two parental proteins that comprise the fusion. We have built a computational tool for analyzing changes to the PPI networks of chimeric (or ‘fusion’) proteins, called ‘ChiPPI’ (Chimeric PPI), which we have incorporated into the ChiTaRS database, providing a pre-calculated analysis for every human fusion event (http://chitars.md.biu.ac.il, see ‘Full Collection’). Using a methodology that treats discrete protein domains as building blocks of interacting proteins, we have catalogued the protein interaction networks for all the chimeric proteins in ChiTaRS. The ChiPPI method (http://chippi.md.biu.ac.il) is unique in that it incorporates the protein domain-domain co-occurrence scores in order to identify interactors of chimeric proteins. Today, the ChiTaRS-3.1 database of ‘Chimeric Transcripts and RNA-Seq data’ is a collection of 34 922 chimeric transcripts identified by Expressed Sequence Tags (ESTs) and mRNAs from the GenBank (35), ChimerDB (26,36), dbCRID (37), TICdb (38) and the Mitelman collection of cancer fusions (39–42) for Homo sapiens, Mus musculus, Drosophila melanogaster, Rattus norvegicus, Bos taurus, Danio rerio, Saccharomyces cerevisiae and Sus scrofa organisms. All the improvements in content, accessibility, usability and functionality (explained below), place ChiTaRS-3.1 as one of the major, up-to-date resources for the study of chromosomal and trans-splicing alterations in cancer.

*To whom correspondence should be addressed. Tel: +972 72 264 4901; Fax: +972 72 264 4901; Email: milana.morgenstern@biu.ac.il

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Table 1. The major improvements and data additions in ChiTaRS-3.1 in comparison to ChiTaRS-2.1.

| Content                              | ChiTaRS-2.1                | ChiTaRS-3.1                | Relevance |
|--------------------------------------|----------------------------|----------------------------|-----------|
| The collection of chimeric transcripts | 29 500 (total), 20 753 (Homo sapiens), 6226 (Mus musculus), 2151 (D. melanogaster), 4 (Bos taurus), 8 (Rattus norvegicus), 4 (Danio rerio), 5 (S. cerevisiae), 13 (Sus scrofa) 1280 | 34 922 (total), 24 608 (Homo sapiens), 7457 (Mus musculus), 2740 (D. melanogaster), 6 (Bos taurus), 10 (Rattus norvegicus), 7 (Danio rerio), 5 (S. cerevisiae), 89 (Sus scrofa) 11 714 including 69 SaS chimeras (634 FASTA sequences of chimeras) | We extended the collection for all eight organisms by ~4500 new entries. |
| Cancer Breakpoints                    |                           |                            |           |
| Chimeric protein–protein interaction (ChiPPI) networks | No | 2081 (validated), 22 527 (predicted) | Bona-fide expression of unique cancer-restricted fusion transcripts extended by more than 10 000 new entries. |
| Manual annotation of Sense-antiSense (SaS) chimeras GeneCards, iHop, PubMed, NCBI, Ensembl, OMIM, RefSeq, Mitelman | No | 200 | We added pre-computed ChiPPI networks for every human entry in ‘Full Collection’ and ‘Breakpoints’ |
|                                      |                            |                            |           |
| IMPROVEMENTS                          |                           |                            |           |
| The major updates and improvements to the content and functionality of ChiTaRS are summarized in Table 1 and Supplementary Table S1. The improvements include: the addition of >4500 chimeric transcripts from eight organisms, and >10 000 cancer breakpoints; prediction of Chimeric protein–protein interaction (ChiPPI) networks, manual annotation of Sense-antiSense (SaS) chimeras, newly added automatic annotation and links to UniProt (43), GeneCards (44), iHop (45), GeneBank (35), Ensembl (46), OMIM (47), RefSeq (48) and the Mitelman collection (39) for every entry in the ‘Full Collection’ (Figure 1, The ChiTaRS-3.1 Interface Screen-shot). |
| Updated database content              |                           |                            |           |
| In the current 2016 update, 34 922 chimeric transcripts have been collated from eight organisms (Table 1). We have identified and annotated 11 714 cancer breakpoints from the recent study of Merten et al and from the Mitelman collection (39–42). To study all these cancer fusions (see ‘Breakpoints’ collection), we have performed manual confirmation of their veracity using the information from >7700 PubMed articles and >19 000 iHop links (Table 1 and Figure 1). Malignancies with the most frequently found fusions are Adenocarcinoma (6308 fusions, ChiTaRS-3.1), Chronic Leukemia (1140 fusions), Acute Lymphoblastic Leukemia (2078 fusions), and Acute Myeloid Leukemia (AML) (135 fusions) (Supplementary Table S2). ChiTaRS-3.1 consists of 435 chimeric transcripts and their junction sites that have been confirmed by RNA-seq datasets (the Human Body Map dataset analyses from (18)), and 77 chimeras have been confirmed by the mass-spec experiments (18,33,34). Finally, for all the Breakpoints collection, the website tool has been greatly improved to provide a user-friendly interface (see ‘Breakpoints’, and Supplementary Figure S1). |
| To make the ChiTaRS-3.1 collection the most comprehensive source of chimeric transcripts available, we regularly update the list of chimeras deriving from the GenBank collection of ESTs and mRNAs for H. sapiens (UCSC reference genome: GRCh37/hg19), M. musculus (NCBI37/mm9), D. melanogaster (BDGP R5/dm3), R. norvegicus (RGSC Rnor_6.0/rn6), B. taurus ( Baylor College of Medicine HGSC Btau4.6.1/bosTaur7), D. rerio (Sanger Institute Zv9/danRer7), S. cerevisiae (SGD April 2011 sequence/sacCer3) and S. scrofa (Broad/Pig3) (35,49). Over the past two years, we have added additional chimeric transcripts for all eight organisms: H. sapiens (22 810), M. musculus (7457), D. melanogaster (2740), R. norvegicus (10), B. taurus (6), D. rerio (7), S. cerevisiae (5), S. scrofa (89) (Table 1). |
| To provide biological context to fusion sequence data, we updated the GenBank (35), RefSeq (48) and Mitelman (39–42) cross-references for all the genes (total 170 797 cross-references). All the UniProt (43) references into ChiTaRS-3.1 have been updated and include now >17 000 unique proteins. Further, we added 29 643 Ensembl (46) cross-references for all gene names (Table 1). Finally, we added links to GeneCards (44), iHop (45), OMIM (47) and PubMed publications for 34 922 entries in the ‘Full Collection’ (Figure 1). Thus, ChiTaRS-3.1 is an easy-to-use resource for the in-depth study of fusion transcripts and proteins on a genome-wide, and multi-species level. |
| Updated database functionality         |                           |                            |           |
| The improved user interface of ChiTaRS-3.1 allows for rapid and easy analysis of evolutionary conservation of fusions, literature references and experimental data supporting fusion expression in different organisms (see ‘Compare and Analyze’). We added a separate pop-up window with an extended annotation for every entry in the Full Collection (Figure 1, see a green button of the fusion, “EU216064”), allowing easy cross-reference to other databases (listed above). |
| Annotation of sense-antisense (SaS) chimeras |                           |                            |           |
| The phenomenon of Sense-antiSense (SaS) chimeric transcripts (34) is also covered by the ChiTaRS database in this latest version. While SaSs may result in chimeric protein translation, they also represent potential inhibitors of translation though dsRNA-mediated mechanisms (34) (Supple-
Figure 1. Improved ChiTaRS-3.1 interface. The improved interface of ChiTaRS-3.1 displays information about fusion proteins, their annotations, cross-links to GeneCards, Splice graphs and ‘ChiPPI predicted’ networks.

SaSs that have been identified in any of the eight organisms in ChiTaRS-2.1 (34) are easily accessed by clicking a check-box (‘Sense-ANTIsense transcripts’) on the ‘Full Collection’ page (Supplementary Figure S2). We collected 6485 chimeric RNA transcripts found in the eight organisms that comprise sense and antisense exons of the same open reading frame, incorporating them into ChiTaRS-3.1 (Table 1). Moreover, we have added manual annotation for 200 SaS chimeras that includes predicted trans-membrane domains of the translated fusion (in six frames), number of overlapping exons at the chimeric junction site, onco-genes, and the corresponding (homologous) transcripts found in other seven organisms. Interestingly, we have identified 17 common SaS chimeras in H. sapiens, M. musculus and D. melanogaster as well as 11 evolutionary-conserved SaS chimeras in three organisms: H. sapiens, M. musculus and S. scrofa (Supplementary Figure S3). Curiously, we found that two-thirds of the genes in those SaS chimeras are evolutionary-conserved phosphatases that lose their phosphate binding sites by means of incorporation of antisense exons (data not shown). This observation appears to be in line with our previous findings that some chimeric proteins tend to have a dominant-negative function in cells (19). Thus, ChiTaRS-3.1 uniquely and comprehensively catalogs SaS chimeras, rendering the study of their evolution and function readily accessible for users, world-wide.
Figure 2. The ChiPPI protein–protein Interaction network for the AFF1/KMT2A fusion protein. (A) The initial PPI networks of parental proteins (AFF1 and KMT2A). (B) The ChiPPI network for the AFF1/KMT2A fusion (chimeraID: AM050775, ‘Full Collection’ and ‘Breakpoints’). The onco-proteins, parental proteins, potential onco-proteins, tumor suppressors and normal proteins are shown in dark-orange, light-orange, yellow, blue and green colors correspondingly. (C) All the missing interactors are shown on the network. The interactions of KMT2A with the tumor suppressors MEN1, SMARCBI and CBFB and also the interactions of AFF1 with the tumor suppressors EAF1 and SIAH1 are lost upon the fusion. (D) The network of the ‘affected’ interactions, e.g. those changed upon the fusion event. (E) All the interactions that stay unchanged upon the fusion event.

Chimeric protein–protein interaction (ChiPPI) networks

Next, we asked how fusion protein function can be best assessed using computational methods. A fusion protein typically contains discreet domains from both parental proteins. It has been shown that a fusion event can dramatically alter the protein–protein interaction (PPI) network of the parental proteins. We, therefore, designed a visual means of assessing PPI network perturbations induced by protein fusion events. To this end, we have assembled a tool for analyzing PPI networks that focuses on individual protein domains as the mediators of PPIs (http://chippi.md.biu.ac.il/). Using this tool we have fully pre-computed chimeric protein–protein interactions (ChiPPI) networks for 2081 fusion proteins, and predicted ChiPPI networks for 22527 human chimeric transcripts. The predicted ChiPPI networks have been produced by the unification of the PPI networks...
of two parental proteins of a chimera (see ‘Full Collection’). This new feather provides users with the ability to study the protein interaction networks of chimeric proteins for all cancer fusions.

ChiPPI displays PPI networks in a map that gives a broad overview of the consequences of a fusion event from a proteomic perspective. ChiPPI predicts where fusion proteins are likely to lose binding to interactors of the parental proteins. Figure 2 shows the predicted PPI network of AFF1-KMT2A fusion protein. The interactions of KMT2A with the tumor suppressors: MEN1, SMARCB1 and CBFB, as well as the interactions of AFF1 with the tumor suppressors: EAF1 and SIAH1 are lost upon the fusion. Using ChiPPI, we mapped the influence of specific fusion proteins on cellular PPI networks and on essential pathways in cancer development and progression. For example, Supplementary Table S3 shows the mapping of ChiPPI networks for the different ABL1 fusions and their alterations in the ‘betweenness centrality’, ‘clustering coefficient’, and scoring for the addition of onco-proteins to, or removal of tumor suppressors from, the PPI network. In general, we find that the PPI networks of fusion proteins often lose tumor suppressor proteins, as well as being enriched in onco-proteins. Thus, ChiPPI is highly suitable for displaying how fusion proteins contribute to the skewing of protein interaction networks as well as of signaling pathways. Particularly, this new feature provides users with the ability to study the protein interaction networks of different cancer fusions (http://chippi.md.biu.ac.il/).

CONCLUSIONS AND PERSPECTIVES

The enhanced ChiTaRS-3.1 database is a comprehensive resource dedicated to the study of chimeric alterations at the proteomic, transcriptomic, genomic level in eukaryotes. ChiTaRS and ChiPPI are recently being referenced as sources for publishable data on cancer fusions (e.g. (50)). The updated version 3.1 of the ChiTaRS database provides a vast increase in annotated and verified chimeric transcripts as compared to the previous ChiTaRS releases, and includes a significant extension of specific research-oriented features. ChiTaRS-3.1 provides extensive experimental evidence for chimeras and cancer fusions, which can be effectively applied in the planning of new experiments or for the analysis of large scale RNA-sequencing experiments. International projects like ICGC and TCGA will benefit from this database and on all incremental additions to it, for improving the process of chimera identification and validation. To conclude, the ChiTaRS-3.1 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-3.1 content will be continuously maintained and updated every six months. The database is now publicly accessible at http://chitars.md.biu.ac.il and its old version 2.1 is accessible at http://chitars.bioinfo.cnio.es/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank MPLabs LTD for the website design and support, and whose efforts make ChiTaRS-3.1 a reality. We thank our users for their consistent support and valuable feedback and our outstanding group for their priceless discussions and suggestions. We would like to thank Prof. Alfonso Valencia and his group at Centro Nacional de Investigaciones Oncológicas (CNIO), Spain, for their valuable suggestions in the overall process of designing the database and writing the manuscript.

FUNDING

PBC (VATAT) (22351, 20027) [to S.T. for Post-Doctoral Fellowship], Funding for open access charge: VATAT [VaTaT #22351 to M.F.-M. & S.T.].

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. Guigo, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E. et al. (2006) EGAASP: the human ENCODE genome annotation assessment project. Genome Biol., 7(Suppl. 1), 1–31.
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., Iderk, T., Rist, B., Eng, J., Hood, L. and Aebersold, R. (2002) Complementary profiling of gene expression at the transcriptomic and proteomic 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 (CHAVI), C.H.I.A.V.I. (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. 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. U.S.A., 105, 6864–6869.
10. Akiva, P., Toporik, A., Edelheit, S., Perez, Y., Diber, A., Shemes, R., Novik, A. and Sorek, R. (2006) Transcription-mediated gene fusion in the human genome. Genome Res., 16, 30–36.
11. Parra, G., Reymond, A., Dabbous, N., Dermitzakis, E.T., Castelo, R., Thomson, T.M., Antonarakis, S.E. and Guigo, R. (2006) Tandem chimerism as a means to increase protein complexity in the human genome. Genome Res., 16, 37–44.
12. Romani, A., Guerra, E., Tretorota, M. and Alberti, S. (2003) Detection and analysis of spliced chimeric mRNAs in sequence databases. Nucleic Acids Res., 31, c17.
13. Campbell,P.J., Stephens,P.J., Pleasance,E.D., O’Meara,S., Li,H., Santarius,T., Stebbing,L.A., Leroy,C., Edkings,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.

14. Ortiz de Mendiboul,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, e48055.

15. 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.

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. Edgren,H., Murumagi,A., Kangaspeska,S., Nicoriici,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.

18. 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.

19. Frenkel-Morgenstern,M. and Valencia,A. (2012) Novel domain combinations in proteins encoded by chimeric transcripts. *Bioinformatics*, 28, 167–174.

20. Asmann,Y.W., Necula,B.M., Kalar,K.R., Hossain,A., Baker,T.R., Carr,J.M., Davis,C., Getz,G., 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.

21. Gingeras,T.R. (2009) Implications of chimaeric non-co-linear transcripts. *Nature*, 461, 206–211.

22. Maher,C.A., Palanisamy,N., Brenner,J.C., Cao,X., Kalyana-Sundaram,S., Luo,S., Khrebtukova,I., Barrett,T.R., Grasso,C., Yu,J. et al. (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, 106, 12353–12358.

23. Maher,C.A., Kumar-Sinha,C., Cao,X., Kalyana-Sundaram,S., Han,B., Jing,X., Sam,L., Barrett,T., Palanisamy,N., Peterman,S. and Maccoss,M.J. (2009) Expediting the development of targeted SRM assays: using data from shotgun proteomics to identify disease-specific therapeutic targets in breast cancer. *Cancer Res.*, 72, 1921–1928.

24. Djebari,S., Lagarde,J., Kapranov,P., Lacroix,V., Borel,C., Mudge,J.M., Howald,C., Foissee,S., Ucla,C., Christ,J. et al. (2012) Evidence for transcript networks composed of chimeric RNAs in human cells. *PLoS One*, 7, e28213.

25. Prakash,A., Tomazela,D.M., Fureen,B., Maclean,B., Merrihew,G., Petersen,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.

26. Kim,P., Yoon,S., Kim,N., Lee,S., Ko,M., Lee,H., Kang,H. and Kim,J. (2010) ChimerDB 2.0—a knowledgebase for gene fusions updated. *Nucleic Acids Res.*, 38, D81–D85.

27. Denoef,D., Kapranov,P., Ucla,C., Frankish,A., Castelo,R., Drenkow,J., Lagarde,J., Alioto,T., Manzano,C., Christ,J. et al. (2007) Prominent use of distal 5’ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.*, 17, 746–759.

28. Houseley,J. and Tollervey,D. (2010) Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One*, 5, e12271.

29. McManus,C.J., Duff,M.O., Eipper-Mains,J. and Graveley,B.R. (2010) Global analysis of trans-splicing in Dro sophila. *Proc. Natl. Acad. Sci. U.S.A.*, 107, 12975–12979.

30. McManus,C.J., Coolon,J.D., Duff,M.O., Eipper-Mains,J., Graveley,B.R. and Wittkopp,P.J. (2010) Regulatory divergence in Dro sophila revealed by mRNA-seq. *Genome Res.*, 20, 816–825.