Wang et al. Diagnostic Pathology (2015) 10:131
DOI 10.1186/s13000-015-0310-4

DIAGNOSTIC PATHOLOGY

DATABASE Open Access

FusionCancer: a database of cancer fusion genes derived from RNA-seq data

Yunjin Wang[1], Nan Wu[2], Jiaqi Liu[2], Zhihong Wu[2,3*] and Dong Dong[1*]

Abstract

Background: Fusion genes are chimeric results originated from previous separate genes with aberrant functions. The resulting protein products may lead to abnormal status of expression levels, functions and action sites, which in return may cause the abnormal proliferation of cells and cancer development.

Results: With the emergence of next-generation sequencing technology, RNA-seq has spurred gene fusion discovery in various cancer types. In this work, we compiled 591 recently published RNA-seq datasets in 15 kinds of human cancer, and the gene fusion events were comprehensively identified. Based on the results, a database was developed for gene fusion in cancers (FusionCancer), with the attempt to provide a user-friendly utility for the cancer research community. A flexible query engine has been developed for the acquisition of annotated information of cancer fusion genes, which would help users to determine the chimera events leading to functional changes. FusionCancer can be accessible at the following hyperlink website: http://donglab.ecnu.edu.cn/databases/FusionCancer/

Conclusion: To the best of our knowledge, FusionCancer is the first comprehensive fusion gene database derived only from cancer RNA-seq data.

Introduction

Fusion gene is a chimera product due to the consolidation of two separate genes, which can occur as a consequence of chromosomal structural changes, such as inversion, deletion, amplification or inter-chromosomal/intra-chromosomal translocation [1]. Gene fusion can bring dramatic expression changes compared to previous separate genes because of the regulatory domain displacement, and the resulting chimeric protein-coding transcript will either lose its original function or work into a scabbed protein with functions descended from both its ancestors [2]. As a consequence of accumulation of hereditary variations, cancer can also be the result of gene fusion, especially the fusions related to kinases or transcription controllers [3]. Considering their prevalence and common characteristics across diverse human cancer types, gene fusions are always regarded as a distinct class of 'mutations'. For example, the recurrent *EML4-ALK* fusion event in lung cancer have been identified [4], and play important role in tumor metastasis. Previously, the fusion events were detected mainly based on RT-PCR, which is not suitable for massively identify fusion genes in cancers.

Although the occurrence of gene fusion events in solid tumor has long been noted, the importance has been realized due to the emergence of next-generation sequencing technology, such as transcriptome sequencing (RNA-seq) [3]. RNA-seq permits genome-wide novel transcript analysis, and spurred gene fusion discovery from diverse human cancers, including prostate, breast, lung and bladder carcinoma [4–7], etc. Up to date, huge amount of cancer RNA-seq data have been available, which provided us an opportunity to comprehensively identify the fusion genes. Meanwhile, algorithms and pipelines provided great convenience for the gene fusion detection [8–11], and many software have been developed with high sensitivity and specificity [12–15]. One obvious benefit of gene fusion discovery based on RNA-seq data is the potential to detect novel gene fusion events.

* Correspondence: wuzh3000@126.com; ddong.ecnu@gmail.com
[2]Department of Orthopaedic Surgery, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China
[1]Institute of Molecular Ecology and Evolution, SKLEC & IECR, East China Normal University, Shanghai, China
Full list of author information is available at the end of the article

Several databases of gene fusion have already been issued, such as chimerDB [16] and HYBRIDdb [17]. They are either manually curated database of published literatures or by mapping EST sequences to the human genome to find cancer-related fusion genes in human, which lead to a lower coverage of gene fusions. In this work, we retrieved recently published RNA-seq data in 15 kinds of human cancer, and comprehensively detected the gene fusion events using four gene fusion detection methods. A user-friendly database, FusionCancer, was developed with the attempt to facilitate the cancer gene fusion researches. FusionCancer is provided with an integrated web-based utility, which made our predicted gene fusion events easily accessible to cancer research community.

Data Content

RNA-seq method can allow identification of gene fusions in individual cancer samples and facilitate comprehensive characterization of cellular transcriptome. A huge amount of RNA-seq based cancer transcriptome data are available, which provided valuable resources for us to comprehensively identify gene fusion events in cancers. As shown in Fig. 1, we searched NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra) database [18] for single-end (SE) and pair-end (PE) RNA-seq data in diverse cancer types with the following search terms: 'cancer' , 'carcinoma' and 'RNA-seq'. The data included in this work have to meet the following criteria: 1) the length of sequencing reads is larger than 36 bp; 2) the data is cancer-related. Finally, a total of 591 cancer samples, published between 2008 and 2014, were selected for further processing (Table 1).

The samples included in our database focused on 15 types of cancers. All sequencing raw data were downloaded, and a software named fastq-dump in the SRA Toolkit version 2.3.2 Linux package [19] was obtained from SRA Software page to convert sra format to fastq format with default parameters. We removed the low quality sequencing reads prior to analyzing these data. Two criteria were used in this step: 1) removing reads with adaptors; 2) removing reads with unknown 'N' bases. All subsequent analyses to detect fusion genes were based on these filtered sequencing reads.

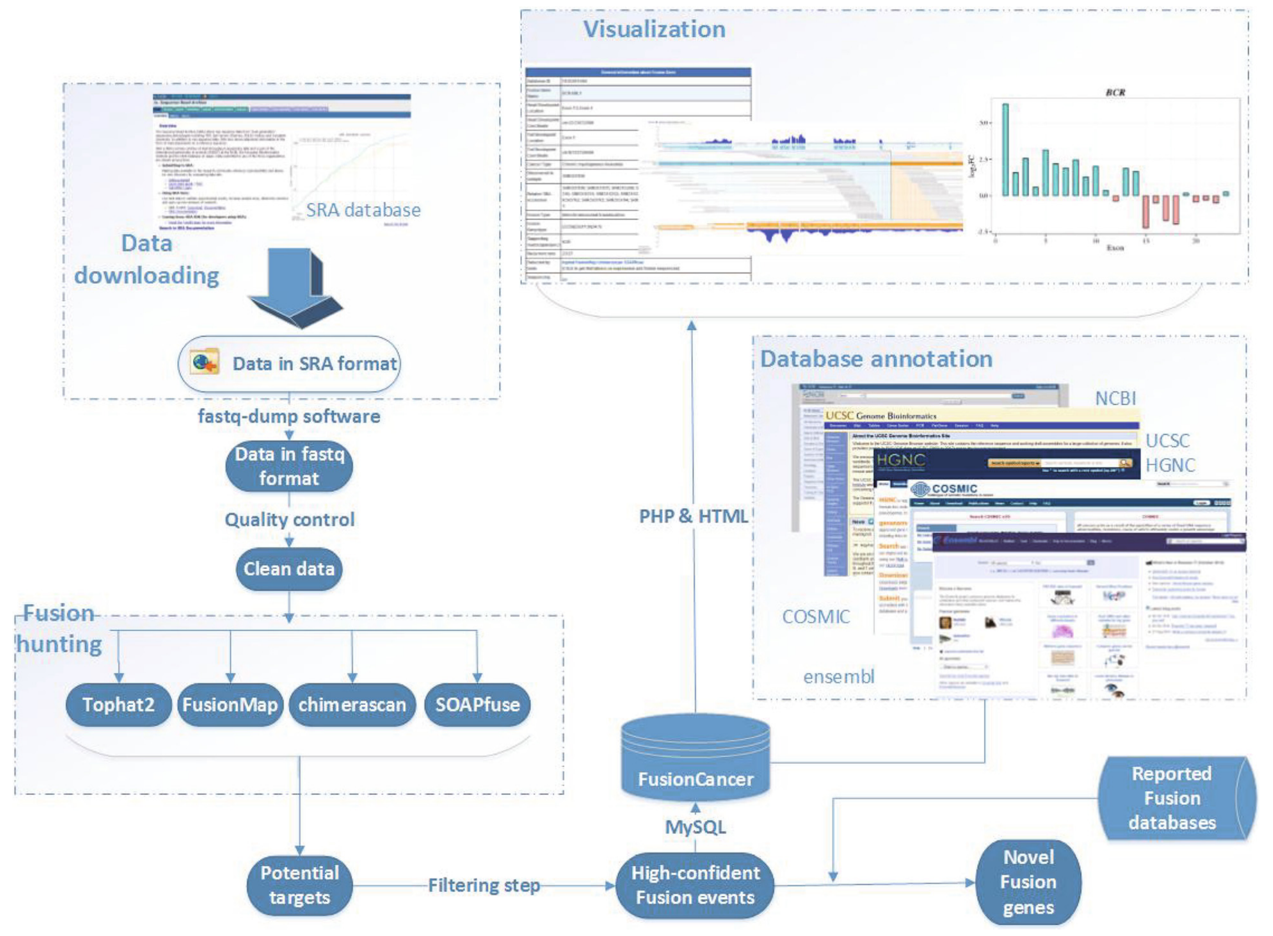

Fig. 1 The data acquisition and database construction pipeline

Table 1 List of cancer types and number of RNA-seq dataset included in our work

Cancer types	Number of datasets
Acute Myeloid Leukemia	13
Bladder Cancer	7
Burkitt Lymphoma	52
Cervical Cancer	15
Chronic Myelogenous Leukemia	53
Colon Cancer	18
Hepatocellular carcinoma	22
Intraocular Melanoma	8
Lung Cancer Non-small Cell	18
Lung Cancer	100
Melanoma	27
Neuroblastoma	77
Ovarian Cancer	12
Parathyroid Cancer	7
Prostate Cancer	162
Total	591

Fusion gene detection methods have experienced a rapid development due to the emergence of next-generation sequencing technology. In this work, four popular fusion gene detection software (Tophat2 [14], FusionMap [12], SOAPfuse [15], and chimerascan [13] were employed in our pipeline. The former two software can process both single-end and paired-end datasets, while the latter two can only deal with single-end data. We aligned all short reads to the human genome (UCSC hg19), and identified a preliminary set of fusion genes by selecting all the gene-gene pairs. Next, in-build filtering steps were performed, and fusion events were retained if they meet at least one of the following criteria: 1) fusion event formed by two distant genes (with a distance larger than 100000); 2) with a recurrence rate () larger than 0.2; 3) identified by at least two software; 4) containing at least 10 supporting reads. At last, a total of 11,839 gene fusion events were identified based on at least one software, and only 137 fusion genes were identified by all four software. The exon-level expressions of fusion genes and wild-type parts of the fusion genes were calculated by Reads Per Kilobase of transcript per Million mapped reads (RPKM). In addition, COSMIC [20] and chimerDB [16] databases have stored some previously documented known fusion genes in different cancer types. We downloaded 288 fusion genes from COSMIC and chimerDB databases, among which 209 fusion events can be found in our FusionCancer database. So, we implemented these information into FusionCancer.

Database implementation

The FusionCancer database is implemented with PHP, MySQL on a Red Hat Linux system, and provides several common gateway interface scripts to process user's input to search the database. A schematic diagram FusionCancer organization is shown in Fig. 1, and FusionCancer can be accessible at the following hyperlink website: http://donglab.ecnu.edu.cn/databases/FusionCancer/

Retrieve data

FusionCancer provides two ways to query the database: one is to search fusion genes of interests by keyword, the other is to browse database by cancer types or chromosomes. FusionCancer can be accessed with gene symbols, and return a list of fusion genes, coupled with biological implications, chromosome information and nucleotide sequences. Three kinds of keywords (Gene symbol, Database ID and Gene pairs) can be selected to search fusion gene results. Moreover, users can select fusion genes identified by a specific software. One the other hand, all fusion genes can be viewed through the Browse DB section by choosing a specific cancer type or chromosome.

BLAST

To help users perform sequence similarity analysis, BALST was provided in the database. A maximum size of 50 k sequence file with fasta format is required. All fusion sequences at transcription level discovered by four software were used as BLAST database. And users can perform sequence alignment using BLASTn searching form and appropriate parameters listed in the BLAST page.

Download

All datasets are available in this section. These datasets contain all predicted fusion genes by four software, accompanied with the annotated information and chimeric transcript sequences.

Conclusion

RNA-seq is a recently developed way to the transcriptome profiling that uses massively parallel RNA-sequencing technology [21]. The ability of RNA-seq to analyze the whole transcriptome in an unbiased fashion makes it an attractive technology to measure the dysregulation in cancers. Furthermore, it also allows identification of gene fusion in individual cancer samples. With the attempt to provide a more comprehensive cancer fusion gene resource, we compiled recently published cancer RNA-seq data and identified all possible gene fusion events using four popular software. We presented an easily accessible database, offering access to those identified fusion genes. The integration of cancer fusion genes can enhance the role of FusionCancer as an essential resource for cancer fusion gene analysis. To the best of our knowledge, FusionCancer is the first repository centralizing cancer fusion genes identified from RNA-seq datasets. The database not only provides a large resource for cancer researches, but also supplies a platform for tumor specific individual biomarker analysis.

Future direction

With the development of next-generation sequencing, sequencing costs will drop substantially. More and more cancer sequencing data would be available in the near future, and these data can provide us valuable resource. Moreover, it will be facilitated by the development of improved bioinformatics procedures for the detection of fusion genes from RNA-seq data. Future directions include an incorporation of more cancer fusion genes with higher accuracy.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YW, ZW, NW, JL carried out this work and DD drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work is supported by the National Natural Science Foundation of China to Dong Dong (Grant No. is 31200956).

Author details

[1]Institute of Molecular Ecology and Evolution, SKLEC & IECR, East China Normal University, Shanghai, China. [2]Department of Orthopaedic Surgery, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China. [3]Department of Central Laboratory, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China.

Received: 7 January 2015 Accepted: 2 June 2015
Published online: 28 July 2015

References

1. Edwards PA. Fusion genes and chromosome translocations in the common epithelial cancers. J Pathol. 2010;220:244–54.
2. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer. 2007;7:233–45.
3. Shyr D, Liu Q. Next generation sequencing in cancer research and clinical application. Biological procedures online. 2013;15:4.
4. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature. 2007;448:561–6.
5. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005;310:644–8.
6. Robinson DR, Kalyana-Sundaram S, Wu YM, Shankar S, Cao X, Ateeq B, et al. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. Nat Med. 2011;17:1646–51.
7. Williams SV, Hurst CD, Knowles MA. Oncogenic FGFR3 gene fusions in bladder cancer. Hum Mol Genet. 2013;22:795–803.
8. Chung GT, Lung RW, Hui AB, Yip KY, Woo JK, Chow C, et al. Identification of a recurrent transforming UBR5-ZNF423 fusion gene in EBV-associated nasopharyngeal carcinoma. J Pathol. 2013;231:158–67.
9. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol. 2011;12:R6.
10. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, et al. Transcriptome sequencing to detect gene fusions in cancer. Nature. 2009;458:97–101.
11. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, et al. Chimeric transcript discovery by paired-end transcriptome sequencing. Proceedings of the National Academy of Sciences of the United States of America 2009;106:12353–12358.
12. Ge H, Liu K, Juan T, Fang F, Newman M, Hoeck W. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics. 2011;27:1922–8.
13. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011;27:2903–4.
14. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.
15. Jia W, Qiu K, He M, Song P, Zhou Q, Zhou F, et al. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biol. 2013;14:R12.
16. Kim N, Kim P, Nam S, Shin S, Lee S. ChimerDB - a knowledgebase for fusion sequences. Nucleic Acids Res. 2006;34:D21–4.
17. Kim DS, Huh JW, Kim HS. HYBRIDdb: a database of hybrid genes in the human genome. Bmc Genomics. 2007;8:128.
18. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database, C. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40:D54–56.
19. Acland A, Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2014;42:D7–D17.
20. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011;39:D945–950.
21. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.