BMC Genomics

BioMed Central

Database

Open Access

HYBRIDdb: a database of hybrid genes in the human genome

Dae-Soo Kim[1], Jae-Won Huh[2] and Heui-Soo Kim*[1,2]

Address: [1]PBBRC, Interdisciplinary Research Program of Bioinformatics, College of Natural Sciences, Pusan National University, Busan 609-735, Korea and [2]Division of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Korea

Email: Dae-Soo Kim - kds2465@pusan.ac.kr; Jae-Won Huh - primate@pusan.ac.kr; Heui-Soo Kim* - khs307@pusan.ac.kr

* Corresponding author

Published: 23 May 2007

BMC Genomics 2007, **8**:128 doi:10.1186/1471-2164-8-128

Received: 26 September 2006

Accepted: 23 May 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/128

© 2007 Kim et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Hybrid genes are candidate risk factors for human tumors by inducing mutation, translocation, inversion, or rearrangement of genes. The occurrence of hybrid genes may also have given rise to new transcripts during hominid evolution.

Description: HYBRIDdb is a database of hybrid genes in humans. This system encompasses the bioinformatics analysis of mRNA, EST, cDNA, and genomic DNA sequences in the INDC databases, and can be used to identify hybrid genes. We searched for hybrid genes among the 28,171 genes listed in the NCBI database, and analyzed their structural patterns in the human genome. The 2,344 gene pairs were detected as hybrid forms of transcriptional products. We classified the hybrid genes into two groups: chromosomal-mediated translocation fusion transcripts and transcription-mediated fusion transcripts.

Conclusion: The HYBRIDdb database will provide genome scientists with insight into potential roles for hybrid genes in human evolution and disease.

Background

Hybrid genes are created by trans-splicing, sense/antisense transcription, genome rearrangement, or intergenic splicing between two genes [1-5]. The creation of new genes is a potential risk factor for tumor development, yet may also promote diversification by inducing gene substitution, translocation, inversion, or rearrangement [1,6-8]. As a result, hybrid genes have the ability to be either harmful or advantageous to humans [9,10].

Cancer genes have somatic mutations similar to chromosomal translocations that result in hybrid transcripts by apposing one gene to the regulatory regions of another. For example, a hybrid transcript is created by genomic rearrangement of the mixed-lineage leukemia (MLL) gene at 11q23 and the septin family (SEPT6) gene at Xq24 in acute leukemia patients [11]. These rearrangements result in fusions with at least 40 other genes, resulting in expression of hybrid proteins with leukemogenic activity [12,13]. Occasionally hybrid gene formations can also contribute to genomic diversity. Normal forms of *UEV* proteins are located in the nuclei of cells, while *Kua* proteins are distributed in endomembranes. The abnormal fusion transcript of these proteins, however, has new enzymatic activity that is associated with cytoplasmic structures [9,14,15]. Thus, *UEV-Kua* gene fusion may be a critical step toward creating a protein with a novel function. The result of splicing out the intergenic region is reported to be a new mechanism of intergenic splicing [15,17]. For example, the *HHLA1-OC90* fusion transcript is expressed by a heterologous HERV-H LTR promoter that is highly active in a teratocarcinoma cell line [18]. Inter-

genic splicing also occurs between the *SSF1* and $P2Y_{11}$ genes on human chromosome 19 [19]. The *SSF1*-$P2Y_{11}$ gene fusion product is 5.6 kb mRNA in length and results in the addition of a potential ATP binding site in *SSF1* [19]. Trans-splicing joins two independently transcribed mRNA sequences at canonical exon-exon borders. Although this is a wide- spread phenomenon among lower eukaryotes, only a few isolated cases have been reported in mammals. In the hybrid *CYP3A* transcripts, *CYP3A43* exon 1 is joined to distinct sets of *CYP3A4* or *CYP3A5* exons at canonical splice sites [5].

Although advanced cytogenetic banding experiments reveal important information about gene rearrangement mechanisms, experimental screening is costly, time consuming, and tedious. The translocation-mediated hybrid transcript using bioinformatic tools was examined [20,21]. Kapranov et al. [14], Parra et al. [15], and Akiva et al. [16] have also analyzed the intergenic splicing-mediated gene fusion mechanism only in the human genome. Therefore, we describe the database, HYBRIDdb, which was designed to detect all of the human hybrid genes (chromosomal-mediated translocation, intergenic splicing-mediated, and few trans-splicing hybrid genes) from publicly available transcript sequences for the understanding of the complex gene catalog in normal and abnormal human tissues. We systematically identified hybrid genes from human sequences and discovered some unique features. Included in this database is a comprehensive list of hybrid genes created by trans-splicing, intergenic splicing, and genomic rearrangement between two human genes.

Construction and content

Data set

The human transcript (mRNA, EST, cDNA) and human genome sequences were downloaded from the NCBI database. Mobile elements in the human genome sequences were identified by RepeatMasker [22], and transposable element consensus sequences were identified by Repbase Update [23]. Useful transcript information from tissues and pathology samples was obtained from NCBI genbank.

In silico *identification of transcriptional hybrid genes*

To characterize the phenomenon of hybrid genes present in the human genome, we clustered RefSeq mRNA onto the human genome sequence (NCBI Build 35.1) using the SIM4 program. RefSeq mRNA was obtained from the NCBI Genbank database. If RefSeq mRNA sequences overlapped, only the longest was considered. After aligning the mRNAs using the SIM4 program [24], we removed naturally overlapping gene pairs based on the coordinates of RefSeq in the human genome sequence. We filtered out connecting sequences with high scoring alignment in both genes. Then, human mRNA, cDNA, and EST sequences in the NCBI GenBank were aligned to human RefSeq using the MegaBLAST program [25]. Alignments having >97% sequence identity and a minimum length of 100 bp were used in this study. We searched hybrid transcript sequences having high-scoring alignments in both genes. To remove the cDNA and mRNA sequences which were suspected to have DNA contaminations, we selected the sequences which have minimum two different sources of transcript (mRNA, cDNA, EST). And also they should be spliced canonically and share at least one splice site with each of the two separate genes. We extracted the human transcript sequences (mRNA, cDNA, EST) which were aligned to two different RefSeq mRNA sequences. For the confirmation of hybrid transcript candidate, we aligned our candidate sequences to genomic DNA sequence by using the SIM4 program [24] and the alignment around the fusion point was manually inspected. To discard the false-positive results of alignment error derived by the homologous gene family, we filtered out connecting sequences having high-scoring alignments in both genes. We extracted the position information of the exon and genome sequences to be matched. Based on this information, the location of the hybrid transcripts and exons were calculated from their position on the genome. We searched for canonically spliced sequences connecting these two transcripts that share at least one splice site with each of the genes. To avoid the use of DNA-contaminated cDNA sequences, we demanded that the sequences connecting these two transcripts should be canonically spliced and share at least one splice site with each of the two separate genes. This procedure identified 2,344 pairs of the 28,171 genes as hybrid transcript candidates, including hybrids transcript created by chromosomal translocation (Figure 1A) or intergenic splicing between two genes (Figure 1B).

Utility and discussion

User interface

HYBRIDdb is a biological database that uses a MySQL management system to transfer the data from a primary database. The HYBRIDdb database can be accessed through a CGI-Python base web interface that includes one retrieval section, referred to as the transcriptional hybrid genes section (Figure 2). HYBRIDdb provides detailed information regarding the exploration of a specific hybrid gene of interest.

Transcriptional hybrid genes interface

The transcriptional hybrid gene was created by joining the transcript portions of two different genes in the human genome. Access to the database can be obtained in three ways. First, users may search for genes of interest using the HUGO symbol, and will retrieve sequences and detailed information from the NCBI data bank. Secondly, users

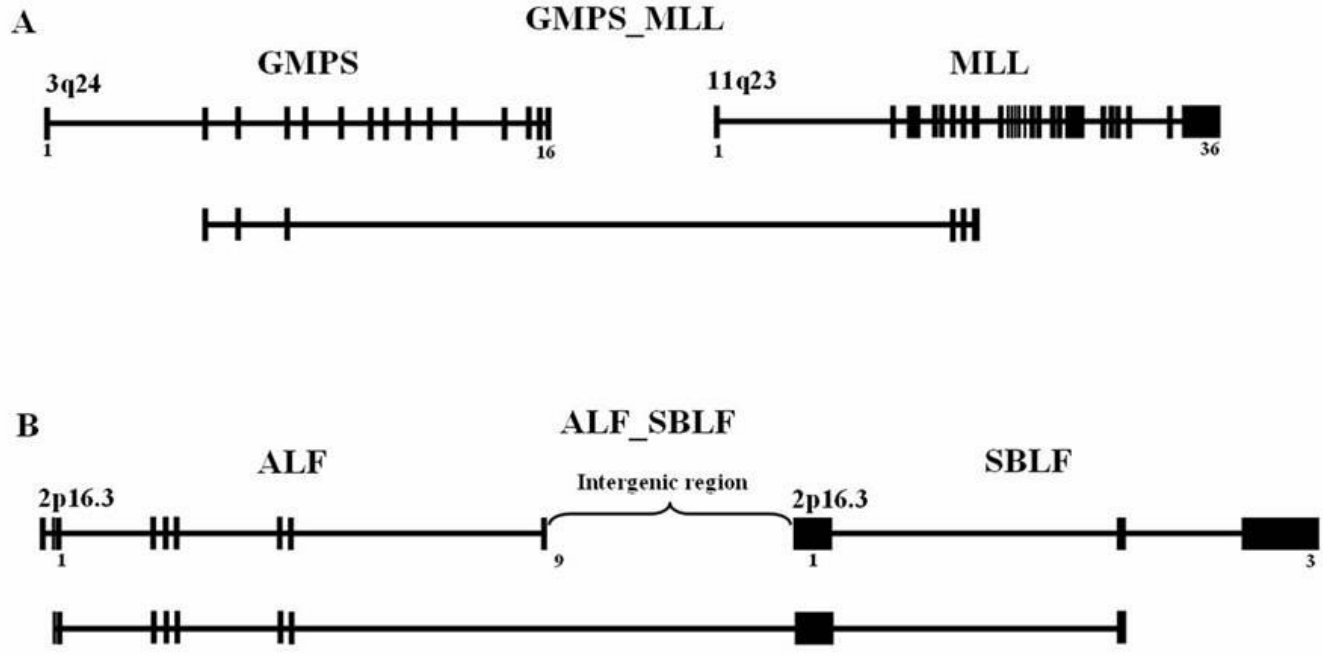

Figure 1
Schematic representation of transcriptional hybrid gene. A model for transcriptional hybrid gene. The hybrid genes, GMPS_MLL(A) and ALF_SBLF (B), are represented. GMPS_MLL hybrid transcript was created by genome rearrangement. ALF_SBLF was a hybrid transcripts containing exons intergenic splicing between two genes. Boxes represent the exons, and bold line indicates the introns.

may search interesting gene names by clicking on a list of genes on the view page according to their chromosome numbers. Moreover, it is possible for users to view the results of this search by clicking on genomic loci. Thirdly, users can view the results of this search by clicking on pathology information or tissue information, and they can also acquire mRNA sequences from the NCBI data bank for further study.

The graphic viewer shows hybrid gene events in the human genome that are represented by the exon-intron splicing structure of mRNAs/ESTs, and functional analysis from the conserved domain database using RPS-BLAST [26]. The results page also includes tissue, pathology, and organ information about the target gene. Importantly, users can see detailed tissue, pathology, and organ information about the target gene in the table displayed on the results page.

Conclusion

HYBRIDdb is an integrated database for genome-wide hybrid genes in humans. This system can identify hybrid genes containing hybrid transcripts created by chromosomal-mediated translocation and intergenic splicing-mediated gene fusion. HYBRIDdb is constantly being updated with new human gene databases from available sources. We also plan to supplement this database with hybrid genes from other mammalian species so that they can be directly compared with hybrid genes from humans. Our work should provide insight into roles for hybrid genes in human evolution and disease.

Availability and requirements

HYBRIDdb is publicly available at the URL http://www.primate.or.kr/hybriddb. Questions and comments are welcomed through the site.

Abbreviations

BLAST – Basic Local Alignment Search Tool

CGI – Common Gateway Interface

EST – Expressed Sequence Tag

HUGO – Human Genome Organization

INSDC – International Nucleotide Sequence Databases

NCBI – National Center for Biotechnology Information

RPS-BLAST – Reversed Position Specific Blast

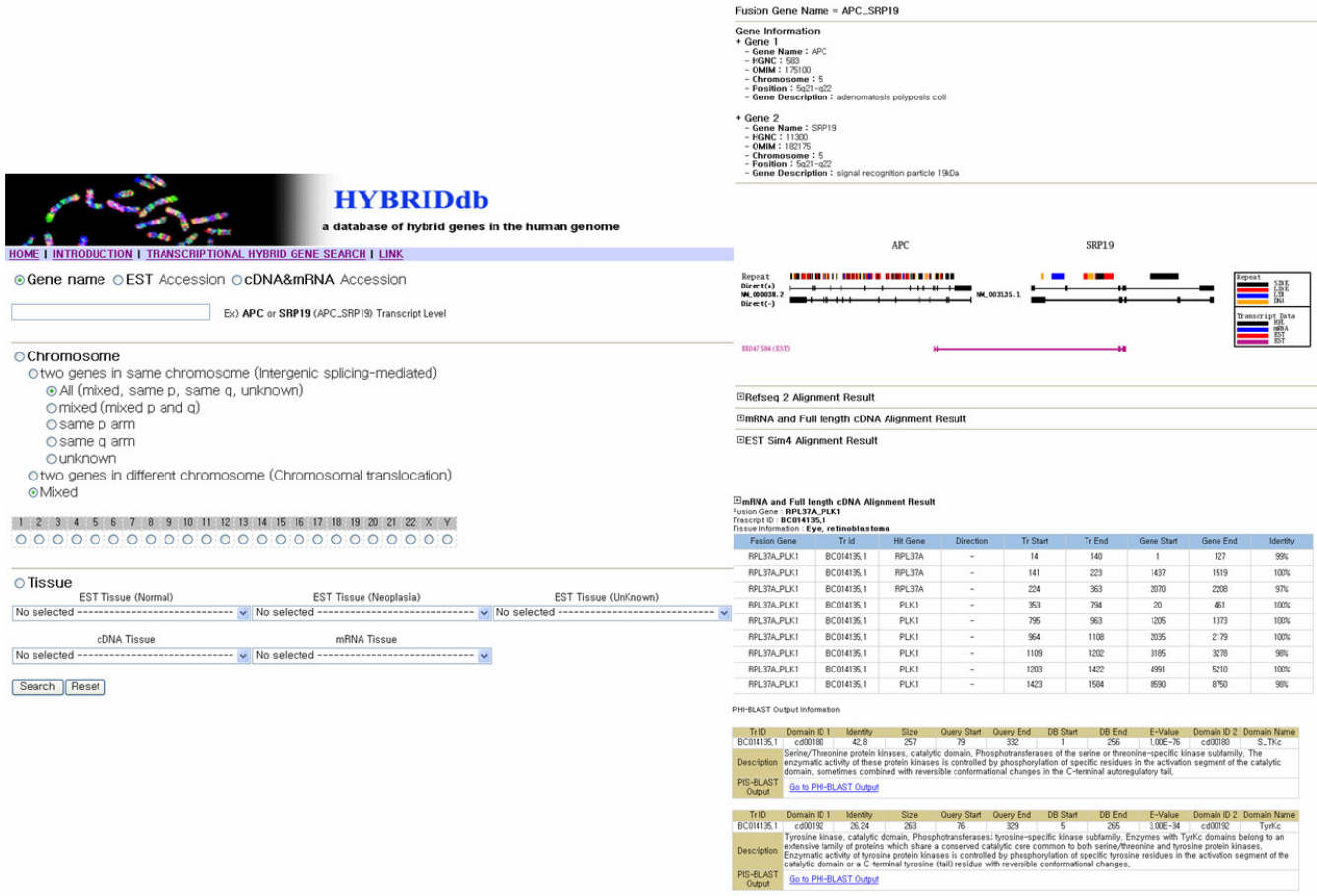

Figure 2

A snapshot of HYBRIDdb interface. The hybrid genes web retrieval interface and a sample from the results pages of HYBRIDdb. The web interface provides access to the database contents in three ways. The results pages are listed in a tabular format which provides evidence for hybrid formation, as well as information about the alignment of hybrid gene pairs within the human genome. The graphic viewer shows hybrid gene formation events that are represented by the exon-intron splicing structure of mRNAs/ESTs/cDNA. And also, the results page includes transcript information about the hybrid gene pairs and tissue, pathology, and organ information about the target gene.

Authors' contributions

DS Kim analyzed contents of this paper and wrote the manuscript. JW Huh contributed the manuscript correction and continuous discussion. HS Kim participated in its analysis and provided essential direction.

Acknowledgements

This study was supported by a grant from the National R&D Program for Cancer Control, Ministry of Health & Welfare, Republic of Korea (0620150-1). We thank to UJ Jo for his technical assistance.

References

1. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nature Rev Cancer* 2004, **4:**177-183.
2. Mitelman F, Johansson B, Mertens F: **Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer.** *Nature Genet* 2004, **36:**331-334.
3. Romani A, Guerra E, Trerotola M, Alberti S: **Detection and analysis of spliced chimeric mRNAs in sequence databanks.** *Nucleic Acids Res* 2003, **31:**e17.
4. Dahary D, Elroy-Stein O, Sorek R: **Naturally occurring antisense: transcriptional leakage or real overlap?** *Genome Res* 2005, **15:**364-368.
5. Finta C, Zaphiropoulos PG: **Intergenic mRNA molecules resulting from trans-splicing.** *J Biol Chem* 2002, **277:**5882-5890.
6. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
7. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285:**751-753.
8. Peltonen L, McKusick VA: **Dissecting human disease in the postgenomic era.** *Science* 2001, **291:**1224-1229.
9. Long M: **A new function evolved from gene fusion.** *Genome Res* 2000, **10:**1743-1756.
10. Courseaux A, Nahon JL: **Birth of two chimeric genes in the hominidae lineage.** *Science* 2001, **291:**1293-1297.
11. Kadkol SS, Bruno A, Oh S, Schmidt ML, Lindgren V: ***MLL-SEPT6* fusion transcript with a novel sequence in an infant with**

acute myeloid leukemia. *Cancer Genet Cytogenet* 2006, **168:**162-167.
12. Collins EC, Rabbitts TH: **The promiscuous MLL gene links chromosomal translocations to cellular differentiation and tumour tropism.** *Trends Mol Med* 2002, **8:**436-442.
13. Eguchi M, Eguchi-Ishimae M, Greaves M: **The role of the MLL gene in infant leukemia.** *Int J Hematol* 2003, **78:**390-401.
14. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR: **Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays.** *Genome Res* 2005, **15:**987-997.
15. Thomson TM, Lozano JJ, Loukili N, Carrio R, Serras F, Cormand B, Valeri M, Diaz VM, Abril J, Burset M, *et al.*: **Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene.** *Genome Res* 2000, **10:**1743-1756.
16. Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, Antonarakis SE, Guigo R: **Tandem chimerism as a means to increase protein complexity in the human genome.** *Genome Res* 2006, **16:**37-44.
17. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, Sorek R: **Transcription-mediated gene fusion in the human genome.** *Genome Res* 2006, **16:**30-36.
18. Kowalski PE, Freeman JD, Mager DL: **Intergenic splicing between a HERV-H endogenous retrovirus and two adjacent human genes.** *Genomics* 1999, **57:**371-379.
19. Communi D, Suarez-Huerta N, Dussossoy D, Savi P, Boeynaems JM: **Cotranscription and intergenic splicing of human P2Y11 and SSF1 genes.** *J Biol Chem* 2001, **276:**16561-16566.
20. Hahn YS, Bera TE, Gehlhaus K, Kirsch IR, Pastan IH, Lee BK: **Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases.** *Proc Natl Acad Sci USA* 2004, **101:**13257-13261.
21. Kim N, Kim P, Nam S, Shin S, Lee S: **ChimerDB-a knowledgebase for fusion sequences.** *Nucleic Acid Res* 2006, **34:**D21-D24.
22. **RepeatMasker** [http://repeatmasker.genome.washington.edu]
23. Jurka J: **Repbase update: a database and an electronic journal of repetitive elements.** *Trends Genet* 2000, **16:**418-420.
24. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: **A computer program for aligning a cDNA sequence with a genomic DNA sequence.** *Genome Res* 1998, **8:**967-974.
25. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences.** *J Comput Biol* 2000, **7:**203-214.
26. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral