ChimerDB - Database of Chimeric Sequences in the GenBank

Namshin Kim¹,², Seokmin Shin², Kwang-Hwi Cho³ and Sanghyuk Lee¹,*

¹Division of Molecular Life Sciences, Ewha Womans University, Seoul 120-750, Korea
²School of Chemistry, Seoul National University, Seoul 151-747, Korea
³Department of Bioinformatics and CAMDRC, Soongsil University, Seoul 156-743, Korea

Abstract

Fusion proteins resulting from chimeric sequences are excellent targets for therapeutic drug development. We developed a database of chimeric sequences by examining the genomic alignment of mRNA and EST sequences in the GenBank. We identified 688 chimeric mRNA and 20,998 chimeric EST sequences. Including EST sequences greatly expands the scope of chimeric sequences even though it inevitably accompanies many artifacts. Chimeric sequences are clustered according to the ECgene ID so that the user can easily find chimeric sequences related to a specific gene. Alignments of chimeric sequences are displayed as custom tracks in the UCSC genome browser. ChimerDB, available at http://genome.ewha.ac.kr/ECgene/ChimerDB/, should be a valuable resource for finding drug targets to treat cancers.

Keywords: chimeric sequence, chromosomal translocation, trans-splicing, multi-locus transcription

Introduction

Chromosomal translocation is frequently observed in many hematologic and solid tumors (Mitelman, 2000). It can affect the gene expression by disrupting the promoter region of the gene or by joining the gene with enhancer elements like immunoglobulin or T-cell receptor genes (Croce et al., 1984). However, chromosomal translocation in tumors frequently creates a chimeric mRNA sequence encoding a fusion protein that interferes the normal regulating pathways (Mitelman, 2000). The most famous example is the fusion protein BCR-ABL, which is the target protein of the drug Gleevec treating CML, chronic myeloid leukemia (Mauro and Druker, 2001).

CML is associated in most cases with a chromosomal translocation between chromosomes 9 and 22 that creates the Philadelphia chromosome. The BCR gene in chr22 is fused with the gene ABL in chr9, so called the t(9;22)(q34;q11) translocation. The tyrosine kinase activity of ABL is constantly activated by the BCR gene (GTPase activator) in the fusion protein, resulting in the rapid cellular mitosis and inability of the cell to perform apoptosis. Gleevec inhibits the tyrosine kinase ability of the BCR-ABL fusion. Gleevec is revolutionary because it proved that molecular targeting works in treating cancer if the target is correctly chosen. In AML (acute myeloid leukemia) patients, the RUNX1 gene (also known as AML1 or CBFA2) is among the most frequent targets of chromosomal rearrangements in human leukemias (Mikhail et al., 2004). More than 20 cases have been observed by fluorescence in situ hybridization (FISH) analysis (Mikhail et al., 2002).

Given the success of Gleevec and the recent explosion of genomic data, it is quite plausible to search the GenBank database looking for chimeric sequences. Fusion proteins from chimeric mRNAs may serve as excellent drug targets in other tumors. Function and expression analyses of resultant fusion proteins are essential to find better targets as can be seen in the example of the BCR-ABL case, the combination of activator and kinase domains.

Chimeric sequences can be generated from other mechanisms. Trans-splicing can join two independently transcribed mRNA sequences at canonical exon-exon borders. Trans-splicing was first detected in vitro (Solnick, 1985), but subsequently shown to occur in vivo in eukaryotes (Bonen, 1993) and mammals (Caudella et al., 1998). Even though natural trans-splicing of pre-mRNAs has been regarded to be a rare event in mammals, human estrogen receptor-α gene is shown to be trans-spliced in addition to alternative cis-splicing (Flouriot et al., 2002).

Long transcription across neighboring genes that normally act as independent transcription units has been demonstrated in several cases. Cotranscription and intergenic splicing lead to a fusion mRNA. Even though
its role is not known, it certainly increases the diversity of the exon complement of the participating genes, which may contribute to gene evolution (Communi et al., 2001; Finta and Zaphiropoulos, 2000).

Alberti and coworkers developed a screening procedure (ISTReS, in silico trans-splicing retrieval system) to identify heterologous, spliced mRNAs with potential origin from chromosomal translocation, mRNA trans-splicing, and multi-locus transcription (Romani et al., 2003). mRNA sequences are aligned onto the genome by BLAST, and applied several criteria for filtering and validation. EST sequences were discarded since the transcript direction is not annotated in many cases.

EST sequence data increases the transcriptome diversity significantly. For example, 61% of transcript variants due to alternative splicing are supported only by EST sequences. Furthermore, gene expression pattern can be inferred by examining the cDNA libraries comprising EST sequences. The NCI's CGAP (cancer genome anatomy project) is a major source of public EST sequences, and 59% of human cDNA libraries in the dbEST are related to cancers. Therefore, we believe that EST sequence data is too valuable to be discarded especially when the purpose of searching for chimeric sequences is finding drug targets to treat cancers. The problem of uncertain read direction can be alleviated by using the genome-based EST clustering procedures as in the UniGene or the ECgene.

In this paper, we describe a new publicly available database of chimeric sequences – ChimerDB – and the associated navigator tools. Its coverage is substantially larger than the ISTReS result because all mRNA and EST sequences in the GenBank are included.

Methods

Datasets

We used the July 2003 human reference sequence (UCSC version hg16) that is based on NCBI Build 34. The genome sequence was downloaded from the UCSC genome center. mRNA and ESTs sequences are obtained from NCBI GenBank release 138 (October 15, 2003). ESTs in gbestNN.seq and mRNAs in gbhtcNN.seq and gbpriNN.seq were extracted. ECgene version 1.1 based on hg16 is used throughout this work. Reference sequences were not used since their build procedure specifically removes chimeric sequences.

Algorithm

mRNA and EST sequences were aligned onto the human genome using the BLAT program (Kent, 2002). We used the alignments whose percent identity is over 93%. BLAT output may contain many overlapping hits. Segments whose overlap \( \leq 16 \) bp are treated as independent alignments. The longest alignment is chosen among dependent alignments. We kept chimeras with two independent segments since other chimeras were likely to derive from the random co-ligation of unrelated cDNA fragments. To retain reliable chimeric sequences, we imposed three additional conditions – (1) the small segment should cover over 20% of the sequence, (2) the large segments should cover less than 80% of the sequence, and (3) aligned sequences from the two segments should be over 80% of the entire sequence. Segments in the same chromosome should be separated farther than 1 Mbp. Any alignment on chromosome M and random assembly was discarded. As a final step, each segment should have more than 30 bp overlap with an ECgene gene model. Since the ECgene is based on clustering EST sequences of high quality, this removes unreliable chimeras from intergenic or intronic region of the genome. Resultant chimeric sequences are grouped according to the ECgene ID of constituent segments.

Results and Discussion

ChimerDB Website

The contents of the chimeric sequence database can be accessed at the ChimerDB website, http://genome.ewha.ac.kr/ECgene/ChimerDB. Figure 1 is the summary table of chimeric sequences. Chromosome 22 contains 573 ECgene clusters overlapping with at least one chimeric sequence. The table shows number of chimeric mRNA and EST sequences. It should be noted that many chimeric relationships are supported by EST sequences only as can be seen in the second and third columns. Since the ECgene contains many putative genes with no known RefSeq, we provide the number of RefSeq, mRNA, and EST sequences to aid user's decision on reliability of the gene. Clicking the ECgene ID shows more detailed information of the ECgene cluster.

The BCR gene (H22C1443) contains 12 chimeric mRNA and 3 chimeric ESTs, and they generate four different fusion genes as listed in the last two columns. In addition to the well-known ABL1 gene, we have three fusion partners – SLU7, MTHFS, and FGFR1. Clicking on 'R123' shows detailed information on each chimeric sequence as shown in Figure 2. The table summarizes how the constituent genes align on the chimeric sequence. It also shows which genes are joined in each
chimeric sequence. The alignment information is helpful in assessing the reliability of the chimeric sequence. For example, two EST sequences (BF380221, BF094682) giving the BCR-SLU7 fusion gene, contain unaligned nucleotides in the range of 162-187 bp. Both ESTs are from the same library of uterus-tumor tissue. All chimeric mRNA sequences generate the BCR-ABL1 fusion gene, and most of them (except AB069693) have contiguous alignments between the two segments.

We also provide a viewer for genomic alignment of the chimeric sequences. Clicking on the ECgene ID (H22C1443) in Figure 2 opens a genome browser as shown in Figure 3. It shows the ECgene structure and alignments of chimeric sequences as custom tracks.

Fig. 1. Summary page of chimeric relationships in chromosome 22. Chimeric sequences are clustered according to overlapping ECgene ID. For example, the entry R123 contains chimeric sequences related to the ECgene H22C1443, which comprises 3 RefSeq, 17 mRNA, and 348 EST sequences. It represents the BCR gene on chromosome 22. The second and third column means that there are 12 chimeric mRNA and 3 chimeric EST sequences overlapping with the ECgene H22C1443. Clicking on ECgene ID show the summary page of corresponding ECgene. The last column shows the counterpart genes constituting the chimeric sequence in case official gene symbols are available.

Fig. 2. Detailed information on chimeric sequences related to the ECgene H22C1443 (BCR). We have 12 chimeric mRNA and 3 chimeric EST sequences, and the table summarizes the alignment of all segment pairs. For example, two EST sequences (BF380221, BF094682) giving the BCR-SLU7 fusion gene, contain unaligned nucleotides in the range of 162-187 bp. Both ESTs are from the same library of uterus-tumor tissue. All chimeric mRNA sequences generate the BCR-ABL1 fusion gene, and most of them (except AB069693) have contiguous alignments between the two segments.

We also provide a viewer for genomic alignment of the chimeric sequences. Clicking on the ECgene ID (H22C1443) in Figure 2 opens a genome browser as shown in Figure 3. It shows the ECgene structure and alignments of chimeric sequences as custom tracks.

---

**ChimerDB : Chimeric Sequence Database for chr22**

**Table 1: Summary for R123 Chimeras in chr22**

<table>
<thead>
<tr>
<th>Chimeric ID</th>
<th>#Chimeric mRNA</th>
<th># Chimer EST</th>
<th>ECgene ID</th>
<th>EcGene Symbol</th>
<th>Partner List</th>
</tr>
</thead>
<tbody>
<tr>
<td>R123</td>
<td>3</td>
<td>1</td>
<td>H22C1443</td>
<td>BCR</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2: Sequence Information**

<table>
<thead>
<tr>
<th>BenBank Accession</th>
<th>Seq. Type</th>
<th>Seq. Length</th>
<th>Pathological</th>
<th>Pathway</th>
<th>Seq. Alignment</th>
<th>Overlapping EcGene</th>
<th>Gene Symbol</th>
<th>Chimeric Partner Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>BF380221</td>
<td>EST</td>
<td>501</td>
<td>Neoplasia</td>
<td></td>
<td>185-554</td>
<td>H22C1443</td>
<td>BCR</td>
<td>SLU7</td>
</tr>
<tr>
<td>BF094682</td>
<td>EST</td>
<td>501</td>
<td>Neoplasia</td>
<td></td>
<td>185-554</td>
<td>H22C1443</td>
<td>BCR</td>
<td>SLU7</td>
</tr>
</tbody>
</table>

**Table 3: Chimeric Information**

<table>
<thead>
<tr>
<th>BenBank Accession</th>
<th>Seq. Type</th>
<th>Seq. Length</th>
<th>Pathological</th>
<th>Pathway</th>
<th>Seq. Alignment</th>
<th>Overlapping EcGene</th>
<th>Gene Symbol</th>
<th>Chimeric Partner Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>BF380221</td>
<td>EST</td>
<td>501</td>
<td>Neoplasia</td>
<td></td>
<td>185-554</td>
<td>H22C1443</td>
<td>BCR</td>
<td>SLU7</td>
</tr>
<tr>
<td>BF094682</td>
<td>EST</td>
<td>501</td>
<td>Neoplasia</td>
<td></td>
<td>185-554</td>
<td>H22C1443</td>
<td>BCR</td>
<td>SLU7</td>
</tr>
</tbody>
</table>

---
Fig. 3. Genome browser for viewing genomic alignment of chimeric sequences. ECgene models and chimeric alignments are added as custom tracks in the UCSC human genome browser. The ID of each chimeric sequence contains brief information on alignment. For example, ‘AJ298917[m,107-324;324]’ implies that this sequence is an mRNA (324 bp long) whose nucleotides ranging 107-324 bp align onto the BCR gene as displayed in the browser. Putative breakpoints are indicated with block arrows.

ECgene structure includes transcript variants due to alternative splicing. Genomic alignment of the chimeric sequence often reveals the location of breakpoint in chromosomal translocation. We find three candidate breakpoints for the BCR-ABL fusion mRNA, indicated as block arrows in the picture. Chromosomal translocation in the Philadelphia chromosome seems to happen frequently near the third breakpoint since several spliced chimeric mRNAs appear in that region. Another benefit of genomic alignment is that we can examine the character of the fusion boundary. The enlarged view in the inset picture shows that the two ESTs inferring the BCR-SLU7 have different alignment from other chimeric sequences. The last exon aligns in the intron region of the gene, which strongly suggests artifactual origin of two chimeric sequences. Clicking on the partner gene (e.g. ABL1 gene, H9C11979) in Figure 2 shows similar tables and pictures for the partner gene.

Statistics of ChimerDB

We identified 688 chimeric mRNA and 20,998 chimeric EST sequences. The number of chimeric mRNAs is approximately 0.6% of 118,034 BLAT-aligned mRNAs. Chimeric ESTs comprise ~0.4% of 4,838,878 BLAT-aligned EST sequences. The ratio of chimeric ESTs is smaller than that of mRNA probably because
short EST sequences have less chance of containing the fusion boundary of the chimeric sequence. Chimeric ESTs consist of 4,235 5' ESTs, 4,848 3' ESTs, and 11,915 ESTs without read direction. The raw database, NCBI's dbEST, contains 2.47 million 5' ESTs, 1.78 million 3' ESTs, and 1.18 million undirected ESTs. Portion of ESTs without read direction is rather high in chimeric sequences, which may imply that chimeric sequences would contain more artifactual sequences.

We found 22,834 ECgene clusters overlapping with any of 21,686 chimeric sequences. 797 ECgenes overlap with both mRNA and EST chimeras. 473 ECgenes overlap with chimeric mRNA sequences, and 21,564 ECgene clusters are related to chimeric EST sequences with no mRNA chimeras. Even though substantial portion of chimeric sequences are expected from the artifacts in cDNA library construction, it is obvious that EST data would contribute significantly to the content of chimeric sequence database.

Total number of the ECgene pairs in the ChimerDB is 75,455. We have 1,685 pair relationships being supported by mRNA chimeras, 147 of which are also supported by chimeric EST sequences. Remaining 73,764 pair relationships are supported only by chimeric EST sequences. This may imply that majority of EST chimers are from artifacts or that the coverage by EST sequences is much wider. It is too early to draw any conclusion at this point.

It should be noted that 25 chimeric mRNAs, identified by Alberti and coworker in 2003, are removed from the GenBank database now. Only 30 chimeric mRNAs among 55 chimeric RefSeq's remain in the database. This suggests that NCBI's RefSeq project is putting substantial efforts to filter out artifactual chimeric sequences. Even chimeric mRNAs should be carefully examined for their validity.

Conclusion

We constructed a database of chimeric sequences by examining the genomic alignment of mRNA and EST sequences in the GenBank. Including EST sequences greatly expands the scope of chimeric sequences even though it inevitably includes many artifactual sequences. Instead of discarding all EST sequences, our strategy is to filter out artifacts at later stages. Currently, we are implementing another filtering scheme that selects chimeric sequences whose junction appears at exon-exon border as in trans-splicing. Function and expression analyses of resultant fusion proteins are also in progress to identify better targets for example, the combination of activator and kinase domains in the BCR-ABL fusion.

Nevertheless, the database and tools provided at the ChimerDB site should be a valuable resource for finding drug targets to treat cancers.

Acknowledgements

We are grateful to the UCSC genome center for making such wonderful resources available to the public. This work was supported by the Ministry of Science and Technology of Korea through the bioinformatics research program of MOST NRDP (M1-0217-00-0027) and the Korea Science and Engineering Foundation through the center for cell signaling research at Ewha Womans University.

References


