Systematic Identification of Novel Fusion Transcripts

Fangqi Hu, Pius Brzoska

Abstract- Fusion transcripts are products of chromosomal rearrangements that are frequently involved in carcinogenesis. A genomic translocation event can result in the fusion of two genes. The resulting chimeric gene is transcribed into a fusion transcript which in turn is translated into a fusion protein. In order to identify fusion transcripts systematically and comprehensively, we developed a pipeline that uses the sequence alignment tool Splign to analyze mRNA sequences. We benchmarked our alignment results against well characterized fusion transcripts and tuned the alignment criteria to minimize false positives and false negative signals. Using this approach, we analyzed the entire set of human Genbank mRNAs (~250,000 accessions) on the human genome. From analyzing the alignment results, we were able to identify 1054 novel fusion transcripts that have not been described before, 260 of which have high level confidence with more stringent criteria. The newly identified fusion transcripts can be valuable for cancer research.

Keywords: fusion transcript, translocation, Splign alignment

I. Introduction

Chromosomal rearrangements are frequently involved in carcinogenesis [1]. Translocations are chromosomal rearrangements of genomic regions and can lead to the fusion of two genes which results in a chimeric mRNA product -- fusion transcript. The fusion transcripts often affect regulatory pathways and stimulate cancer cell growth and proliferation [2]. A well known example is BCR-ABL fusion protein which is found in almost all cases of chronic myeloid leukemia (CML) [3]. However, more fusion transcripts remain unidentified. So far, there’s no known public database that dedicates on collecting and updating fusion transcripts. Currently there’re two known public data sources for fusion transcripts: ChimerDB (http://genome.ewha.ac.kr/ChimerDB/) [4] and FusionGeneDB (NIH group) [5]. However, these datasets have not been updated for years, and are by no means complete. NCBI’s Genbank mRNA collection includes fusion genes which are rarely annotated as such. Many Genbank records result from high throughput cloning studies and only the cDNA sequence is deposited without further biological validation or analysis.

In order to identify systematically which Genbank mRNAs are potential fusion transcripts, we used the sequence alignment tool Splign (developed by NCBI staff) [6] to align the entire set of ~250,000 human Genbank mRNAs accessions on the human chromosomes. The fusion transcripts are identified by determining if the alignment of a transcript sequence can be clearly separated into 2 compartments, with each compartment mapping to a distinct locus on genome.

II. Procedure

A. *cDNA-Genomic Sequence Alignment Program and Parameter Settings*

Splign was downloaded from NCBI website http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi, and installed and run in a local linux server. For fusion transcripts identification, 3 parameter values are not default and set as below when running Splign:

-min_compartment_idty	20% (overall coverage)
-min_exon_idty	95%
-max_intron	1,200,000 bp

B. *Input dataset to Splign*

Query database: entire set of 251,968 human GenBank RNAs downloaded from NCBI

Subject database: entire set of full-length human chromosomes (NCBI B36 assembly)

III. Fusion Transcript Types

There are 4 fusion transcript types as described below:

A. *Fusions are between 2 different chromosomes*

A fragment of a chromosome is translocated to another chromosome resulting in a fusion transcript.

B. *Fusions are between sequences that have big span*

A fragment of a chromosome is translocated on the same chromosome to a different locus.

C. *Fusions are genomic inversions*

Part of a chromosome is spliced out, flipped, and then reconnected.

Manuscript received November 20, 2009.

Fangqi Hu and Pius Brzoska are with Life Technologies, Foster City, CA 94404 USA (phone: 650-554-3075; e-mails: fangqi.hu@lifetech.com, pius.brzoska@lifetech.com).

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

D. *Fusions are genomic reversions*

On the same chromosome, a chunk of sequences is rearranged from one place to another.

The following diagram illustrates the 4 fusion types:

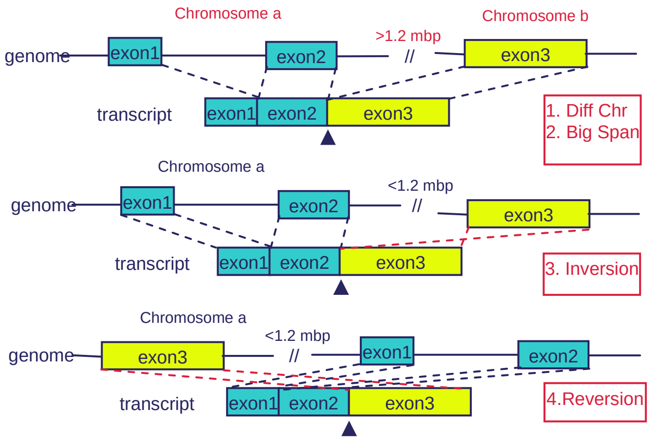

Figure 1. Four Fusion Types

IV. Criteria for Calling Fusion Transcripts

A. Minimum exon identity: 95%

B. Minimum compartment identity: 20%

C. Minimum aligned length of each compartment: 50bp

D. Maximum overlap/gap in breakpoint region on transcript: 50bp

E. Each compartment must have either L-gap (left gap) or R-gap (right gap), but not both (L-gap/R-gap can be ignored if <25bp). M-gap (middle gap), if any, must be <25bp long. (L-gap/M-gap/R-gap are patches in the left/middle/right part of the query sequence that do not align to the genome)

F. Two aligned compartments must reside on different chromosomes, or at least 1.2 million bp away on the same chromosome, or in different orientation such as inversion and reversion.

G. One and only one plausible translocation model. Query transcripts with multiple acceptable models will be excluded.

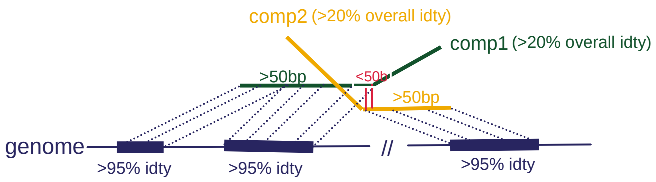

Figure 2. Criteria for calling fusion transcript

V. Results

252,000 Genbank mRNA sequences were processed using Splign and the filters described above. 1202 Genbank mRNAs were identified as fusion transcripts, of which 148 were previously described [4, 5], 1054 were novel. The fusion transcripts were analyzed using the criteria: fusion breakpoint location on transcript and genome, fusion type, percent identity, alignment length and orientation, exon-intron boundary dinucleotide sequences, parent genes.

Some fusion sequences may be artificial chimeras, created by accidental ligation of different cDNAs during the cloning procedure.

- The breakpoint from true fusion genes will usually coincide with a canonical exon boundary because the genes are likely to break in an intron.
- In contrast, the breakpoint for an artificial chimera will usually be within an exon of each gene because the fusion occurs between two cDNAs.

In order to minimize calling artificial chimeras as fusions, the following 3 additional criteria are defined to call fusion transcripts with high-level confidence:

- The exon boundaries of the breakpoint region on chromosome co-localize exactly with the exon coordinates from the parent refseq genes.
- The exon boundaries of the breakpoint region follow the consensus canonical pattern: AG<exon>GT (~98%), or AG<exon>GC (~0.7%), or AC<exon>AT (0.1%).
- The exon boundaries of the breakpoint region are supported by more than 1 Genbank accessions.

After applying these additional requirements, of the 1202 Genbank mRNAs that are identified as fusion transcripts, 369 have high-level confidence.

The following 3 tables show some examples of the identified fusion transcripts.

VI. Discussion

Splign was chosen over other commonly-used transcript-to-genome alignment tools (such as Blat, Sim4 etc) because of its unique compartmentization feature, which was leveraged in our pipeline to separate fusion transcript sequences into 2 distinct compartments when aligning to genome.

The methods and criteria described in this paper for the identification of fusion transcripts are relatively stringent in order to minimize the possibility of false positives that may otherwise arise from loose alignment settings. Of the 265 known fusion transcripts from the 2 data sources [4, 5], 148 were captured as positive, which is 56%. By further fine-tuning the parameter setting and selection method, we may rescue some false negatives.

Of the 1202 identified fusion transcripts, majority of them (77%) come from fusions between 2 different chromosomes, the remaining 23% distribute about equally among same chromosome inversions, reversions, and rearrangement over big span.

VII. Conclusions

Splign alignment tool can be used to identify fusion transcripts with modified parameter settings. About 252K full set human Genbank mRNAs were screened systematically, and 1202 fusion transcripts were identified, of which 148 are known ones, and 1054 are novel ones. 369 fusion transcripts identified have high-level confidence, of which 109 are known ones, and 260 are novel ones. The approach and method described in this paper can facilitate the discovery of novel fusion transcripts, therefore benefit cancer research and other biological studies.

References

[1] Mitelman, F., Johansson, B. & Mertens, F. (2004) *Nat. Genet.* **36,** 331–334.

[2] Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. & Stratton, M. R. “A census of human cancer genes”, 2004, Nat. Rev. Cancer 4, 177–183.

[3] Pane, F., Intrieri, M., Quintarelli, C., Izzo, B., Muccioli, G. C. & Salvatore, F. (2002) Oncogene 21, 8652–8667.

[4] N. Kim, P. Kim, S. Nam1, S. Shin, and S. Lee, “ChimerDB—a knowledgebase for fusion sequences,” in Nucleic Acids Research, 2006, Vol. 34, Database issue D21–D24.

[5] Y. Hahn, T. Bera, K. Gehlhaus, Ilan R. Kirsch, Ira H. Pastan, and B. Lee, “ Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases,” in PNAS, vol. 101, 2004, pp. 13257–13261.

[6] Yuri Kapustin, Alexander Souvorov, Tatiana Tatusova and David Lipman, “Splign: algorithms for computing spliced alignments with identification of paralogs,” in Biology Direct, 2008, **3**:20 doi:10.1186/1745-6150-3-20

TABLE I. Known Fusion Transcript AY662674 confirmed

COMPART NUM	QUERYID	SUBJECTID	IDENTITY	ALIGNLEN	QUERYSTART	QUERYSTOP	SUBJECTSTART	SUBJECTSTOP	PARENTREF	EXONNUMBER
39720	AY662674.1	NC_000011.8	1	129	1	129	3740836	3740708	NM_139132.2	9
39720	AY662674.1	NC_000011.8	1	88	130	217	3738432	3738345	NM_139132.2	10
39720	AY662674.1	NC_000011.8	0.99	97	218	314	3731214	3731118	NM_139132.2	11
35889	AY662674.1	NC_000009.10	1	191	308	498	131521328	131521518	NM_016307.3	2
35889	AY662674.1	NC_000009.10	1	179	499	677	131522696	131522874	NM_016307.3	3
35889	AY662674.1	NC_000009.10	1	127	678	804	131524317	131524443	NM_016307.3	4

This table shows that the map of transcript accession AY662674 on the genome is divided into 2 compartments: comp#39720 on chr.11, and comp#35889 on chr.9. The breakpoint position is 308-314 on transcript. The left part of the transcript has 3 exons co-localizing with refseq NM_139132 (gene NUP98) exon 9 through 11 on chr.11, and the right part of the transcript has 3 exons co-localizing with a different refseq NM_016307 (gene PRRX2) exon 2 through 4 on chr.9. This transcript was previously characterized as fusion, and now it is confirmed in this paper.

TABLE II. NOVEL FUSION TRANSCRIPT -- CHROMOSOME INVERSION

COMPNUM	QUERYID	SUBJECTID	ALIGNLEN	QUERYSTART	QUERYSTOP	SUBJECTSTART	SUBJECTSTOP	SUBJECTSTRAND	CO-LOCATION	EXONNUMBER
3921	BC004197.1	NC_000012.10	128	9	136	56168122	56168249	+	NM_004990.2	1
3921	BC004197.1	NC_000012.10	91	137	227	56169069	56169159	+	NM_004990.2	2
3921	BC004197.1	NC_000012.10	79	228	306	56169317	56169395	+	NM_004990.2	3
3921	BC004197.1	NC_000012.10	135	307	441	56169474	56169608	+	NM_004990.2	4
3921	BC004197.1	NC_000012.10	76	442	517	56169946	56170021	+	NM_004990.2	5
3921	BC004197.1	NC_000012.10	173	518	690	56170257	56170429	+	NM_004990.2	6
3921	BC004197.1	NC_000012.10	107	691	797	56170588	56170694	+	NM_004990.2	7
3921	BC004197.1	NC_000012.10	117	798	914	56178207	56178323	+	NM_004990.2	8
3921	BC004197.1	NC_000012.10	204	915	1118	56178470	56178673	+	NM_004990.2	9
3921	BC004197.1	NC_000012.10	202	1119	1320	56180371	56180572	+	NM_004990.2	10
53810	BC004197.1	NC_000012.10	75	1321	1395	56494276	56494202	-	NM_006576.2	2
53810	BC004197.1	NC_000012.10	197	1396	1592	56493473	56493277	-	NM_006576.2	3
53810	BC004197.1	NC_000012.10	109	1593	1701	56491177	56491069	-	NM_006576.2	4
53810	BC004197.1	NC_000012.10	111	1702	1812	56490976	56490866	-	NM_006576.2	5
53810	BC004197.1	NC_000012.10	203	1813	2015	56490601	56490399	-	NM_006576.2	6
53810	BC004197.1	NC_000012.10	79	2016	2094	56489940	56489862	-	NM_006576.2	7
53810	BC004197.1	NC_000012.10	99	2095	2193	56489745	56489647	-	NM_006576.2	8
53810	BC004197.1	NC_000012.10	154	2194	2347	56488598	56488445	-	NM_006576.2	9
53810	BC004197.1	NC_000012.10	101	2348	2448	56488344	56488244	-	NM_006576.2	10
53810	BC004197.1	NC_000012.10	138	2449	2586	56487777	56487640	-	NM_006576.2	11
53810	BC004197.1	NC_000012.10	159	2587	2745	56487539	56487381	-	NM_006576.2	12

This table shows that the map of BC004191 on genome is divided into 2 compartments: comp#3921 and comp#53810 on chr.12. The breakpoint position is 1320-1321 on transcript. The left part of the transcript has 10 exons co-localizing with refseq gene NM_004990 (MARS) exon 1 through 10 on chr.12, and the right part of the transcript has 11 exons co-localizing with a different refseq gene NM_006576 (AVIL) exon 2 through 12 on the opposite strand of chr.12. This is a fusion by inversion example.

TABLE III. NOVEL FUSION TRANSCRIPTS SUPPORTED BY MULTIPLE GENBANK ACCESSIONS

COMPARTNUM	QUERYID	SUBJECTID	IDENTITY	ALIGNLEN	QUERYSTART	QUERYSTOP	SUBJECTSTART	SUBJECTSTOP	ALIGNTYPE
113630	DQ204773.2	NC_000021.7	1	77	1	77	41,801,952	**41,801,876**	<exon>GA
113630	DQ204773.2	NC_000021.7		100	78	177			<R-Gap>
113631	DQ204773.2	NC_000021.7		72	1	72			<L-Gap>
113631	DQ204773.2	NC_000021.7	1	105	73	177	**38,878,742**	**38,878,638**	TT<exon>GT
96348	EU090248.1	NC_000021.7	1	73	1	73	41,801,948	**41,801,876**	<exon>GA
96348	EU090248.1	NC_000021.7		218	74	291			<R-Gap>
96349	EU090248.1	NC_000021.7		68	1	68			<L-Gap>
96349	EU090248.1	NC_000021.7	1	105	69	173	**38,878,742**	**38,878,638**	TT<exon>GT
96349	EU090248.1	NC_000021.7	1	86	174	259	38869541	38869456	AG<exon>GT
96349	EU090248.1	NC_000021.7	0.969	32	260	291	38717353	38717322	AG<exon>

The table shows that the alignment of DQ204773 on genome co-localizes with the alignment of EU090248 on the breakpoint region, although EU090248 is longer and has 2 extra exons on the right side of the fusion. They both support the same translocation event on chr.21 fusing sequences from position 41,801,876 to position 38,878,742 which is about 3 mega bases downstream. DQ204773 and EU090248 were submitted to Genbank by different submitters. DQ204773 was submitted in 2005 by a Pathology group at University of Michigan; whereas EU090248 was submitted in 2007 by a diagnostic lab in China. So this translocation event is proved by fusion transcripts from 2 different sources, making it unlikely the fusion is an artificial chimera resulting from an unintended cloning error.