The structure of state of art gene fusion-finder algorithms

Marco Beccuti[1], Matteo Carrara[2], Francesca Cordero[1], Susanna Donatelli[1] and Raffaele A Calogero[2]

(1) Department of Computer Science, University of Torino, Torino, Italy;

(2) Department of Biotechnology and Health Sciences, University of Torino, Torino, Italy;

Abstract

Fusion genes, also known as chimeras, play important roles in tumorigenesis and cancer progression. Then, their role become crucial in the areas of biomarkers and therapeutic targets investigation. High-throughput sequencing technologies combined with sophisticated bioinformatics tools might facilitate the discovery of such aberrations. A significant number of bioinformatics algorithms have been developed to detect fusion genes. Detection strategies are quite variegated, then we inspect the strategy of 18 fusion finder algorithm to understand how these tools call chimeras.

1 Introduction

The joining of DNA of two genes, by translocation or inversion, gives rise to gene fusions resulting in hybrid proteins, also know as chimera/fusion products, or in the deregulation of the transcription of one gene by the cis regulatory elements (enhancers) of another. Gene fusions are an important class of cancer aberrations as in the case of BCR-ABL fusion found in nearly all chronic myeloid leukemia patients [1]. Chimeras can be categorized into two classes: intergenic and transgenic fusion transcripts[2] as reported in Figure 1 (A) and (B), respectively. Intergenic fusion transcripts refer to a splicing event between adjacent genes in the same chromosome, while transgenic fusion transcripts gives rise from splicing event involved exons of two genes located in different chromosomes.

High-throughput sequencing technologies facilitate the characterization of the aberrant landscape of human cancers [3, 4], pushing the modern medicine to the development of personalized treatment of cancer patients.

Recently many computation approaches for the detection of chimera, taking advantage of the remarkable throughput of the new RNA-seq technologies [5], have been developed. The recent review of Wang [6] lists a total of 23 different fusion detection tools published between 2009 and 2012. The paper of wag and coworkers also considered 24 papers involving the identification of fusion products in cancer published in the same period. The same paper reports that the chimera detection methodologies used in the 24 papers take advantage only in two cases of the 23 reported tools: FusionSeq and defuse, used to detect fusions in one and two papers, respectively.

It seems quite odd that only few biologically important works did not considered the use of the available fusion detection tools. Carrara and coworkers [7] recently compared the behavior of eight fusion-detection tools (FusionHunter, FusionMap, FusionFinder, MapSplice, deFuse, Bellerophontes, ChimeraScan, and TopHat-fusion) and they highlighted that these tools are able to detect chimeras, but the number of false positive contaminating the results produced make the validation of true fusions a real challenge. In this paper the authors highlight that further improvement in fusion-finder algorithms is essential. In this review we examine the structure of 18 fusion detection published algorithm to identify critical steps in the chimera call procedure that might need further refinements.

2 Results

In this review we consider the following 18 tools: Bellerophontes [8], BreakFusion [9], Breakpointer [10], ChimeraScan [11], deFuse [12], EBARDenovo [13], Ericscript [14], FusionAnalyser [15], FusionFinder [16], FusionHunter [17], FusionMap [18], FusionSeq [19], LifeScope [20], MapSplice [21], ShortFuse [22], SnowShoes-FTD [23], SOAPfuse [24], TopHat-Fusion [25]; which, at the best of our knowledge, are the current state of the art

chimera detection tools.

Before describing these fusion finder algorithms, we recall some terms used in the rest of this paper. RNA-seq experiments generate a huge set of short reads which can be in two forms: *single-end* or *paired-end*. In the former case the sequencer reads only one of the two DNA fragment strand; while in the latter case both the forward and reverse strands of DNA fragment are sequenced, giving rise to a couple of mates, called paired-end read.

During the identification of the fusion boundaries (the positions where the nucleotide coordinates corresponding to the breakpoint of both genes involved in the fusion are discovered) it is possible to classify each read as either *spanning* or *encompassing*, reported in Figure 2. Spanning reads, derived from single-end or paired-end experiments, overlaps with a fusion product. Encompassing read, requires instead paired-end format, harbor a fusion boundary so that each read of the mate maps on a different gene of the fused gene couple. All gene fusion finder algorithms are composed by two phases: first, a mapping step is required to align the reads with respect to the reference specified by the algorithm. Then, a set of filters based on several biological or technical indications will be applied to reduce the set of putative fusion products. The considered tools can be classified according to their alignment strategies into four different macro-groups: *Whole paired-end*, *Paired-end + fragmentation*, *Direct fragmentation* and *statistical read distribution* as summarized in Table 2. The first two techniques requires paired-end reads since they exploit encompassing reads during the first alignment phase, while the last two can be applied on both the read formats.

The *whole paired-end* approach consists in two mapping steps. In the first step the reads are aligned to a reference using mapping tools as Bowtie [26] and BWA [27] and considering a limited number of mismatches. This step can be rise to some "discordant alignment", which occur when both mates have a unique alignments in the reference, but some features do not match the assumption of paired-end design. For example, the mate orientations are not correct or the distance among them do not match with the experiment advises. The discordant alignments are used to generate a set of putative fusion products which will be used as reference for the second alignment step where the unmapped reads are rescued. The resulting putative fusions are passed as input to a filtering step. The tools falling into this category are Breakfusion, Ericscript, deFuse, FusionAnalyser, FusionHunter, FusionSeq, ShortFuse, SOAPfuse and SnowShoes-FTD.

The *paired-end + fragmentation* approach proposes a first phase of mapping similar to the ones used in *whole paired-end* case. Then the second phase of alignment which exploits read fragmentation is performed. In details, the unmapped reads are fragmented and remapped with respect to the reference depending on the tool requirements. Note that using a fragmentation approach these algorithms are able to detect a higher number of junction-spanning reads which simplified the detection of the fusion junctions. Again, the putative fusions are passed as input to the filter step. Bellerophontes, ChimeraScan, LifeScope, TopHat-Fusion are part of this category.

Unlike the previous categories, the *direct fragmentation* approach generates immediately a set of fragmented reads, which will be used in the mapping phase on the reference genome. This approach makes use of the detection strength on the fusion junctions. Also the algorithms of this category propose a set of filters useful to select the real fusion products. The tools that can be classified in this macro-group are EBARDenovo, FusionFinder, FusionMap and MapSlice.

Finally, another approach is the *statistical read distribution* which identifies putative fusion products exploiting both local non-uniform read distribution and mapping signatures containing misalignment at the boundaries of insertions/deletions or more complex structural variants. Then, each putative fusion product is validated using the unmapped reads derived in the previous step. Only the putative fusion products associated with a number of reads greater than a threshold are selected as input to the filtering step.The only tool following this approach is Breakpointer.

All tools implement a final filter step to reduce and to validate the discovered fusion products. Table 2 summarizes the set of filters used by each fusion finder algorithm, described in the following.

Paired-End Information Filters verify the correct distance between the tags of a pair to validate the alignment on a fusion. This distance depends on the protocols used for the library preparation, and the tools can either take this information as input or infer it from the first alignment step. In both cases, reads mapping on the putative fusions at an excessive distance are filtered out. The tools including filters of this class are Bellerophontes, ChimeraScan, deFuse, FusionFinder and SOAPFuse.

Anchor Length Filters use the concept of "Anchor

length" (i.e. the number of nucleotides overlapping each side of a fusion junction) to evaluate the quality of junction-spanning reads associated with a fusion junction. Junction-spanning reads having at least one of the two anchor lengths below a threshold are interpreted as possible artifacts caused by mismatches or sequence similarity, and are removed. FusionHunter, ChimeraScan and TopHat-Fusion take advantage of this class of filters.

Read Through Transcripts Filters try to discover and remove RNA molecules formed by exons of adjacent genes, usually generated when the gene end is not recognized during the RNA elongation phase. Bellerophontes, FusionHunter, FusionAnalyser, FusionFinder, FusionMap, SnowShoes-FTD and TopHat-Fusion are the tools which use this class of filters.

Junction Spanning Reads Filters remove all the fusion products not supported by a number of spanning reads greater than a specified threshold. This class of filters is found in Bellerophontes, FusionHunter, FusionMap, ShortFuse, SnowShoes-FTD, SOAPFuse and TopHat-Fusion.

PCR Artifact Filters try to discover and remove all duplicated reads generated by the PCR amplification process, by the identification of clusters of reads of the same length with an identical alignment on the reference. Bellerophontes, FusionHunter, BreakPointer, EricScript, FusionMap and ShortFuse have an implementation of this class of filters.

Homology Filters remove all the putative fusions having a high number of reads on homologous or repetitive regions which can lead to multiple alignments. The tools belonging to this class of filters are Ericscript, FusionAnalyser, FusionFinder, FusionSeq, SnowShoes-FTD, SOAPFuse and TopHat-Fusion.

Scoring Filters compute for each fusion a corresponding quality based on different metrics (e.g., entropy, base quality, etc.) so that all the candidates with quality lower than a threshold are discarded. BreakPointer, Ericscript, FusionMap, FusionSeq and ShortFuse are the tools which use a specific scoring method to filter putative fusion products.

Reads Quality Filters act on the available reads, actively removing all the paired reads with a score below a threshold, reducing the possibility of ambiguous alignments due to low sequencing quality. Reads qualities filter are used in FusionMap, FusionSeq, LifeScope and SnowShoes-FTD.

Encompassing Reads Filters remove all putative fusion products with a number of encompassing read pairs below a threshold. The three tools including this filtering step are Bellerophontes, FusionAnalyser and SnowShoes-FTD.

Blacklist Filters remove fusions comprising genes present in a list of non-interesting regions which can be either user-defined or fixed, depending on the tool. FusionAnalyser, FusionMap and FusionSeq include a blacklisting feature.

Statistics Filters analyze the reads distribution on the putative fusion products and compute statistical evaluation with respect to the general read distribution on the genome, to decide whether the fusion should be filtered or not. BreakFusion, Ericscript, MapSplice and ShortFuse include a statistical evaluation of the fusion products.

In addition, we identified a total of eleven, tool-specific filters which do not fall in the previous categories. Bellerophontes includes a step removing ambiguous reads. Ericscript has a filtering step based on the sequence homology of the fusion junction. FusionFinder removes putative fusion products containing antisense sequences. FusionSeq requires the expression of the putative fusion to be comparable with the general expression obtained from the sequencing. LifeScope introduces a graph, called Junction Evidence Graph, to represent the fusion products and their junctions and to evaluate the confidence level of each called fusion. MapSplice filters products not containing canonical junctions as well as removing products with introns of unusual length. ShortFuse removes all the reads that align on transcripts for spliceosome components. SnowShoes-FTD checks and possibly filters fusion products on the basis of the orientation of the genes involved and also removes fusions with an excessive number of putative junction points. SOAPfuse adds a step of trimming on the reads that fail to align, in an attempt to rescue them.

3 Discussion

This paper proposes an overview on the main fusion finder algorithms published in literature. In all algorithms inspected the first step concerns the usage of mapping algorithms, ie. Bowtie, BWA, etc. The objective of this algorithms is to select reads that support putative fusion events by discordant reads that have a coherent mapping with known gene annotation. A set of filters based on several biological or technical features follow the mapping step. The filter application

Table 2: Filters implemented by each fusion finder algorithms considered.

Tool name	Readthrough	Homologous Rep.	Junction Spanning	PCR A Artifact	Score	Pair Distance	Reads Quality	Statistics	Anchor Length	Encompass	Blacklist	Additional Specific	
Bellerophontes	X		X	X		X				X		Ambiguous Reads	
FusionHunter	X		X	X					X				
BreakFusion								X					
Breakpointer				X	X								
ChimeraScan						X			X				
deFuse						X							
EBARDenovo													
Ericscript		X		X	X			X				Junction Homology	
FusionAnalyser	X	X								X	X		
FusionFinder	X	X				X						Antisense	
FusionMap	X		X	X	X		X				X		
FusionSeq		X			X		X				X	Comparison Chimera Expression with General Expression	
LifeScope							X					Junction Evidence Graph	
MapSplice								X				Canonical Junctions	Introns Length
ShortFuse			X	X	X			X					Reads From Spliceosome components
SnowShoes-FTD	X	X	X				X			X		Fusion Genes Orientation	excessive putative junc point
SOAPfuse		X	X			X						Read Trimming	
TopHat-Fusion	X	X							X				

Table 1: Fusion finder algorithms classification according to their alignment strategies.

Tool Name	**Macro-Group**
Bellerophontes	Paired-end + Fragmentation
BreakFusion	Whole Paired-end
Breakpointer	Statistical Information Exploiting
ChimeraScan	Paired-end + Fragmentation
deFuse	whole paired-end
EBARDenovo	Direct Fragmentation
Ericscript	Whole Paired-end
FusionAnalyser	Whole Paired-end
FusionFinder	Direct Fragmentation
FusionHunter	Whole Paired-end
FusionMap	Direct Fragmentation
FusionSeq	Whole Paired-end
LifeScope	Paired-end + Fragmentation
MapSplice	Direct Fragmentation
ShortFuse	Whole Paired-end
SnowShoes-FTD	Whole Paired-end
SOAPfuse	Whole Paired-end
TopHat-Fusion	Paired-End + Fragmentation

can reduce the set of putative fusion products to those that could be real fusion products.

At the present time, there is no complete evaluation between tools on the same dataset. Only partial comparisons are available, typically in the papers proposing a new tool, against a subset of the algorithms considered in this review. Some of the evaluations are even in contrast among them, which is not surprising due to the lack of a common available benchmarks. For example EricScript offers a comparison of its performance in terms of CPU time and Area Under the Curve, a measure that estimates the accuracy of each algorithms to discriminate true and false positives. The comparisons have been done on synthetic dataset, against ChimeraScan, DeFuse, FusionMap and ShotFuse, and on real datasets, against DeFuse. Also SOAPFuse offers a detailed comparison an a real dataset. The authors compare its algorithm with respect to DeFuse, TopHatFusion, FusionHunter, ChimeraScan and SnowShoes in terms of CPU time, memory usage and detection of known fusions.

In Carrara et al [7] we observed for all examined tools a very high number of false positive. Unfortunately, the specificity of the tool is not reported in most of the original papers, but there is certainly a factor that may hinder the applicability of the tools in many contexts. The different performances of fusion finder algorithms could be imputed to the application of the filters that each tool applies. Probably a good choice is to provide a clearer

separation between alignment and filter phase to offer a modularity usage of these tools. Thus, there is still some work to be done in the area of chimeras detection expecially concerning the definition of common benchmarks and increased specificity.

4 Acknowledgments

This study was funded by grants from the Epigenomics Flagship Project EPIGEN, MIUR-CNR; FP7-Health-2012-Innovation-1 NGS-PTL Grant no. 306242.

References

[1] Rowley J.D., "Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining", *Nature **243***, pp. 290–293 (**1975**).

[2] Magrangeas F, Pitiot G, Dubois S, Bragado-Nilsson E, Cherel M, Jobert S, Lebeau B, Boisteau O, Lethe B, and Mallet J et al, "Cotranscription and intergenic splicing of human galactose-1-phosphate uridylyltransferase and interleukin-11 receptor alpha-chain genes generate a fusion mrna in normal cells", *The Journal of biological chemistry **273***, pp. 16005–16010 (**1998**).

[3] Maher C.A., Kumar-Sinha C., Cao X., Kalyana-Sundaram S., Han B., Jing X., Sam L., Barrette T., Palanisamy N., and Chinnaiyan A.M., "Transcriptome sequencing to detect gene fusions in cancer", *Nature **458***, pp. 97–101 (**2009**).

[4] Maher C.A., Palanisamy N., Brenner J.C., Cao X., Kalyana-Sundaram S., Luo S., Khrebtukova I., Barrette T.R., Grasso C., Yu J., Lonigro R.J., Schroth G., Kumar-Sinha C., and Chinnaiyan A.M., "Chimeric transcript discovery by paired-end transcriptome sequencing", *Proceedings of the National Academy of Sciences of the United States of America **106***, pp. 12353–12358 (**2009**).

[5] Ozsolak F. and Milos P.M., "Rna sequencing: advances, challenges and opportunities", *Nature reviews Genetics **12***, pp. 87–98 (**2011**).

[6] Wang Q., Xia J., Jia P., Pao W., and Zhao Z., "Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives", *Briefings in bioinformatics* (**2012**).

[7] Carrara M., Beccuti M., Lazzarato F., Cavallo F., Cordero F., Donatelli S., and Calogero R.A., "State-of-the-art fusion-finder algorithms sensitivity and specificity", *BioMed research international* (**2013**).

[8] Abate F., Acquaviva A., Paciello G., Foti C., Ficarra E., Ferrarini A., Delledonne M., Iacobucci I., Soverini S., Martinelli G., and Macii E., "Bellerophontes: an rna-seq data analysis framework for chimeric transcripts discovery based on accurate fusion model", *Bioinformatics **28***, pp. 2113–2121 (**2012**).

[9] Chen K., Wallis J.W., Kandoth C., Kalicki-Veizer J.M., Mungall K.L., Mungall A.J., Jones S.J., Marra M.A., Ley T.J., Mardis E.R., Wilson R.K., Weinstein J.N., and Ding L., "Breakfusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data", *Bioinformatics **28***, pp. 1923–1924 (**2012**).

[10] Sun R., Love M.I., Zemojtel T., Emde A.K., Chung H.R., Vingron M., and Haas S.A., "Breakpointer: using local mapping artifacts to support sequence breakpoint discovery from single-end reads", *Bioinformatics **28***, pp. 1024–1025 (**2012**).

[11] Iyer M.K., Chinnaiyan A.M., and Maher C.A., "Chimerascan: a tool for identifying chimeric transcription in sequencing data", *Bioinformatics **27***, pp. 2903–2904 (**2011**).

[12] McPherson A., Hormozdiari F., Zayed A., Giuliany R., Ha G., Sun M.G., Griffith M., Heravi Moussavi A., Senz J., Melnyk N., Pacheco M., Marra M.A., Hirst M., Nielsen T.O., Sahinalp S.C., Huntsman D., and Shah S.P., "defuse: an algorithm for gene fusion discovery in tumor rna-seq data", *PLoS Computational Biology **7*** (**2011**).

[13] Chu H.T., Hsiao W.W., Chen J.C., Yeh T.J., Tsai M.H., Lin H., Liu Y.W., Lee S.A., Chen C.C., Tsao T.T., and Kao C.Y., "Ebardenovo: highly accurate de novo assembly of rna-seq with efficient chimera-detection", *Bioinformatics **29***, pp. 1004–1010 (**2013**).

[14] Benelli M., Pescucci C., Marseglia G., Severgnini M., Torricelli F., and Magi A., "Discovering chimeric transcripts in paired-end rna-seq data by using ericscript", *Bioinformatics* ***28***, pp. 3232–3239 (**2012**).

[15] Piazza R., Pirola A., Spinelli R., Valletta S., Redaelli S., Magistroni V., and Gambacorti-Passerini C., "Fusionanalyser: a new graphical, event-driven tool for fusion rearrangements discovery", *Nucleic Acids Research* ***40***, pp. e123 (**2012**).

[16] Francis R.W., Thompson-Wicking K., Carter K.W., Anderson D., Kees U.R., and Beesley A.H., "Fusionfinder: a software tool to identify expressed gene fusion candidates from rna-seq data", *PloS One* ***7***, pp. e39987 (**2012**).

[17] Li Y., Chien J., Smith D.I., and Ma J., "Fusionhunter: identifying fusion transcripts in cancer using paired-end rna-seq", *Bioinformatics* ***27***, pp. 1708–1710 (**2011**).

[18] Ge H., Liu K., Juan T., Fang F., Newman M., and Hoeck W., "Fusionmap: Detecting fusion genes from next-generation sequencing data at base-pair resolution", *Bioinformatics* ***27***, pp. 1922–1928 (**2011**).

[19] Sboner A., Habegger L., Pflueger D., Terry S., Chen D.Z., Rozowsky J.S., Tewari A.K., Kitabayashi N., Moss B.J., Chee M.S., Demichelis F., Rubin M.A., and Gerstein M.B., "Fusionseq: a modular framework for finding gene fusions by analyzing paired-end rna-sequencing data", *Genome Biology* ***11***, pp. R104 (**2010**).

[20] Sakarya O., Breu H., Radovich M., Chen Y., Wang Y.N., Barbacioru C., Utiramerur S., Whitley P.P., Brockman J.P., Vatta P., Zhang Z., Popescu L., Muller M.W., Kudlingar V., Garg N., Li C.Y., Kong B.S., Bodeau J.P., Nutter R.C., Gu J., Bramlett K.S., Ichikawa J.K., Hyland F.C., and Siddiqui A.S., "Rna-seq mapping and detection of gene fusions with a suffix array algorithm", *PLOS Computational Biology* ***8***, pp. e1002464 (**2012**).

[21] Wang K., Singh D., Zeng Z., Coleman S.J., Huang Y., Savich G.L., He X., Mieczkowski P., Grimm S.A., Perou C.M., MacLeod J.N., Chiang D.Y., Prins J.F., and Liu J., "Mapsplice: accurate mapping of rna-seq reads for splice junction discovery", *Nucleic Acids Research* ***38***, pp. e178 (**2010**).

[22] Kinsella M., Harismendy O., Nakano M., Frazer K.A., and Bafna V., "Sensitive gene fusion detection using ambiguously mapping rna-seq read pairs", *Bioinformatics* ***27***, pp. 1068–1075 (**2011**).

[23] Asmann Y.W., Hossain A., Necela B.M., Middha S., Kalari K.R., Sun Z., Chai H.S., Williamson D.W., Radisky D., Schroth G.P., Kocher J.P., Perez E.A., and Thompson E.A., "A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines", *Nucleic Acids Research* ***39***, pp. e100 (**2011**).

[24] Jia W., Qiu K., He M., Song P., Zhou Q., Zhou F., Yu Y., Zhu D., Nickerson M.L., Wan S., Liao X., Zhu X., Peng S., Li Y., Wang J., and Gao G., "Soapfuse: an algorithm for identifying fusion transcripts from paired-end rna-seq data", *Genome Biology* ***14***, pp. R12 (**2013**).

[25] Kim D. and Salzberg L., "Tophat-fusion: an algorithm for discovery of novel fusion transcripts", *Genome Biology* ***12***, pp. R72 (**2011**).

[26] Langmead B., Trapnell C., Pop M., and Salzberg SL., "Ultrafast and memory-efficient alignment of short dna sequences to the human genome", *Genome Biology* ***10*** (**2009**).

[27] Li H. and Durbin R., "Fast and accurate short read alignment with burrows-wheeler transform", *Bioinformatics* ***25***, pp. 1754–1760 (**2009**).

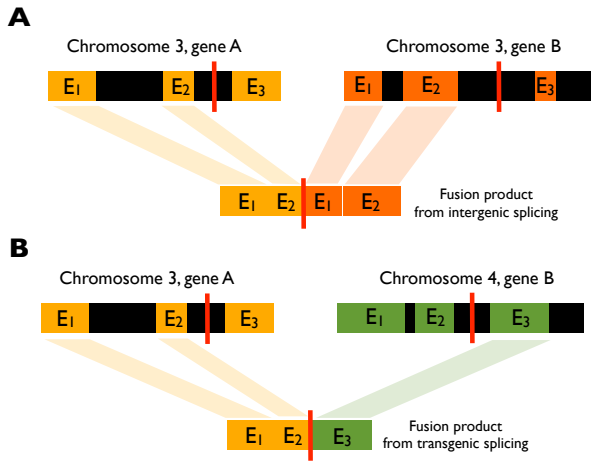

Figure 1: The fusion products are classified into two categories: intergenic fusion transcripts given from the biding of exons (E_n) of two gene on the same chromosome, panel A; and transgenic fusion transcripts obtained by the join of exons of two gens in different chromosomes, panel B.

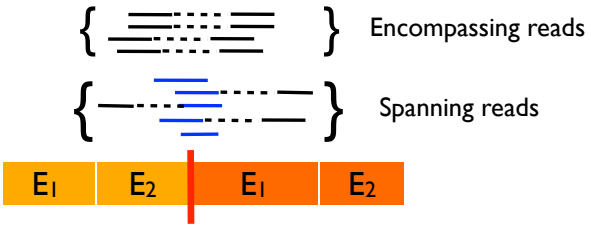

Figure 2: Representation of spanning and encompassing reads. The encompassing reads are represented only by paired-end reads where the mates (continuous lines) harbors the fusion boundary. The spanning reads are represented in both single-end, the read overlaps the fusion product, or paired-end read, only one mate (the blue one) covers the fusion boundary.