Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs

Marcus Kinsella1,*, Olivier Harismendy2,3, Masakazu Nakano4, Kelly A. Frazer2,3,5 and Vineet Bafna5

1Bioinformatics and Systems Biology Program, University of California San Diego, 2Moores UCSD Cancer Center, 3Department of Pediatrics, University of California San Diego, La Jolla, CA 92093, USA, 4Department of Genomic Medical Sciences, Kyoto Prefectural University of Medicine, Kyoto 602-8566, Japan and 5Institute for Genomic Medicine, University of California San Diego, La Jolla, CA 92093, USA

ABSTRACT

Motivation: Paired-end whole transcriptome sequencing provides evidence for fusion transcripts. However, due to the repetitiveness of the transcriptome, many reads have multiple high-quality mappings. Previous methods to find gene fusions either ignored these reads or required additional longer single reads. This can obscure up to 30% of fusions and unnecessarily discards much of the data.

Results: We present a method for using paired-end reads to find fusion transcripts without requiring unique mappings or additional single read sequencing. Using simulated data and data from tumors and cell lines, we show that our method can find fusions with ambiguously mapping read pairs without generating numerous spurious fusions from the many mapping locations.

Availability: A C++ and Python implementation of the method demonstrated in this article is available at http://exon.ucsd.edu/ShortFuse.

Received on November 12, 2010; revised on February 1, 2011; accepted on February 10, 2011

1 INTRODUCTION

The discovery of chimeric transcripts emerging from different and potentially distant genes has introduced another layer of complexity to the genome (Gingeras, 2009). Additionally, the importance of fusion transcripts in the genesis and progression of cancer is becoming increasingly apparent (Mitelman et al., 2004; Perner et al., 2008; Yu et al., 2010a). Fusion transcripts may be the product of trans-splicing, the joining of two different transcripts emerging from distinct and often distant genes. This is especially common among lower eukaryotes (Krause and Hirsh, 1987; Sutton and Boothroyd, 1986) where trans-splicing is part of normal transcript processing (Rajkovic et al., 1990). However, trans-splicing has also been observed in higher eukaryotes, including humans (Horiiuchi and Aigaki, 2006). Additionally, fusions may be produced by adjacent genes yielding a single, joined RNA product, creating a read-through transcript (Akiva et al., 2006). Fusion transcripts can also result from genomic rearrangement that brings together two once distant regions of the genome. Probably the best known example of this type of fusion is BCR-ABL1, a product of a chromosomal translocation (Shivvelman et al., 1985) found in many hematologic cancers and a successful drug target (Druker et al., 2001). In addition, a growing list of fusion genes are being found in both hematologic and solid tumors that are the product of genomic lesions or trans-splicing (Edwards, 2010). Thus, the study of fusion transcripts has implications clinically as well for our basic understanding of the genome.

The development of high-throughput sequencing methods such as RNA-Seq (Wang et al., 2009) has offered an opportunity to hasten a fuller characterization of the transcriptome (Carninci, 2009), including the identification of fusion transcripts. Maher et al. (2009a, b) demonstrated the potential of the technology by applying transcriptome sequencing to several tumors and cancer cell lines. Using two different sequencing protocols, they were able to detect novel fusions such as TMPRSS2-ERG (Tomlins et al., 2005) in a prostate cancer cell line and BCR-ABL1 in a leukemia cell line.

Alongside these biological discoveries has been the development of computational tools and frameworks for the detection of fusion transcripts from RNA-Seq data. Ameur et al. (2010) developed a method for joining partial alignments of single RNA-Seq reads to find splice junctions and gene fusions. Upon application of the method to a public dataset, they found hundreds of examples of transcripts that apparently spanned different chromosomes but were difficult that many were genuine fusion genes. Hu et al. (2010) created a probabilistic method for aligning RNA-Seq read pairs that uses expectation-maximization (EM) to find maximum-likelihood alignments. They showed that paired-end reads better cover splice junctions than single reads and that their method can reliably identify splice junctions. Then, by augmenting their approach with long single reads, they were able to identify 18 gene fusions in two cancer cell lines.

Common to all of these efforts has been the requirement that a fusion transcript be supported by reads that map uniquely to the genome or transcriptome. Maher and colleagues required single best-hit mappings to the genome or mapped short, 36 nt Illumina
some of the repetitiveness is attributable to known repeat families such as the Alu repeat sequence, which can be found both in 5′ and 3′-UTRs as well as occasionally in coding sequence (Yulug et al., 1995). Additionally, many genes are part of gene families or have paralogs or expressed pseudogenes and thus share sequence homology with other parts of the transcriptome. Repeats mapping to these genes or regions of these genes will often map well to other loci.

Ambiguously mapped reads are a concern for all transcriptome sequencing analyses and have previously been addressed by discarding them (Carninci et al., 2005) or by proportionately allocating them over the different positions to which they map (Faulkner et al., 2008; Mortazavi et al., 2008). However, this issue becomes more prominent for gene fusions because combinations of mappings are considered. Consider, for example, a fusion between a pair of genes, A1 and B1. It is possible that a read pair that maps to this fusion will also map to paralogs of each gene, say A2 and B2. If all of these mappings were accepted as true, then three spurious fusions would be called (Fig. 1). If the read pair was discarded because of its ambiguous mappings, evidence for the true fusion would not be detectable without allowing for multimapping reads or annotation errors, we map these discordant read pairs to the genome and transcriptome, and we greatly relax the stringency of reported mappings and allow for many mappings to be reported for each read. For the experiments in this study, we use the Bowtie flags -l 22 -e 350 -y -a -m 5000. These flags cause Bowtie to report all mappings for each read, to try as hard as possible to find valid mappings and to suppress mappings with more than two mismatches in the first 22 bases, summed quality values at all mismatched positions greater than 350 or mappings from reads with more than 5000 reportable mappings. With these less stringent mappings, we check if each pair of reads both map within the genomic bounds of a known gene or within 10 kb of each other in a region of the genome with no annotated genes. This filtering step decreases the possibility of events such as retained introns or misassemblies for pairs that span fusion junctions. In this way, rather than discarding ambiguously mapping read pairs or allowing them to overstate the number of fusions present, we find the best supported fusions by using the mappings of all the reads in the dataset, the quality of those mappings and the implied insert sizes of read pairs that span a fusion site. This allows our method to more sensitively detect gene fusions than if ambiguously mapping read pairs were discarded.

We have implemented our method on simulated data generated from fusions between genes with very high similarity to other genes to demonstrate that our method can resolve the ambiguous mappings to find the correct fusions when it is possible to do so. We then implemented it on reads derived from neoplastic and hyperplastic prostate tissue and recovered the known TMPRSS2-ERG fusion along with several read-through fusions without finding many spurious, poorly supported fusions as a result of allowing reads to have many mappings. Finally, using publicly available data from several melanoma tumors and cell lines, we find fusion events that would not be detectable without allowing for multimapping reads that span the fusion site.

2 METHODS

2.1 Discovery of putative fusions

The first step of our method is to map each read of a pair independently. We use Bowtie (Langmead et al., 2009) in single-end mode to perform this mapping against a database of RefSeq transcripts (Pruitt et al., 2007) that have been prepped with 50 nt of upstream sequence and appended with string of adenines to account for variation in transcription start site and polyadenylation, respectively. Filtering the mapping results yields a set of reportable mappings. With these less stringent mappings, we check if each pair of reads both map within the genomic bounds of a known gene or within 10 kb of each other in a region of the genome with no annotated genes. This filtering step decreases the possibility of events such as retained introns or misassemblies for pairs that span fusion junctions. In this way, rather than discarding ambiguously mapping read pairs or allowing them to overstate the number of fusions present, we find the best supported fusions by using the mappings of all the reads in the dataset, the quality of those mappings and the implied insert sizes of read pairs that span a fusion site. This allows our method to more sensitively detect gene fusions than if ambiguously mapping read pairs were discarded.

In this article, we propose a method to discover fusion transcripts that exploits ambiguously mapping RNA-Seq read pairs, does not require additional long, 75 nt or greater, single read sequencing and decreases the occurrence of mapping artifacts. We begin by mapping read pairs to the transcriptome independently without imposing any unique-mappability criterion. We then find pairs which do not map to the same gene and build a set of possible gene fusions from the mappings of each read. Next, we employ a generative model of RNA-Seq data that utilizes mapping qualities and insert size distributions to resolve any ambiguous mappings. After the convergence of the EM technique used to find maximum-likelihood transcript abundances, we perform a final partial expectation step for the discordantly mapping read pairs to find optimal fusion assignments for pairs that span fusion junctions. In this way, rather than discarding ambiguously mapping read pairs or allowing them to overstate the number of fusions present, we find the best supported fusions by using the mappings of all the reads in the dataset, the quality of those mappings and the implied insert sizes of read pairs that span a fusion site. This allows our method to more sensitively detect gene fusions than if ambiguously mapping read pairs were discarded.

In this article, we propose a method to discover fusion transcripts that exploits ambiguously mapping RNA-Seq read pairs, does not require additional long, 75 nt or greater, single read sequencing and decreases the occurrence of mapping artifacts. We begin by mapping read pairs to the transcriptome independently without imposing any unique-mappability criterion. We then find pairs which do not map to the same gene and build a set of possible gene fusions from the mappings of each read. Next, we employ a generative model of RNA-Seq data that utilizes mapping qualities and insert size distributions to resolve any ambiguous mappings. After the convergence of the EM technique used to find maximum-likelihood transcript abundances, we perform a final partial expectation step for the discordantly mapping read pairs to find optimal fusion assignments for pairs that span fusion junctions. In this way, rather than discarding ambiguously mapping read pairs or allowing them to overstate the number of fusions present, we find the best supported fusions by using the mappings of all the reads in the dataset, the quality of those mappings and the implied insert sizes of read pairs that span a fusion site. This allows our method to more sensitively detect gene fusions than if ambiguously mapping read pairs were discarded.

In this article, we propose a method to discover fusion transcripts that exploits ambiguously mapping RNA-Seq read pairs, does not require additional long, 75 nt or greater, single read sequencing and decreases the occurrence of mapping artifacts. We begin by mapping read pairs to the transcriptome independently without imposing any unique-mappability criterion. We then find pairs which do not map to the same gene and build a set of possible gene fusions from the mappings of each read. Next, we employ a generative model of RNA-Seq data that utilizes mapping qualities and insert size distributions to resolve any ambiguous mappings. After the
Then, we search for paths from the beginning of the upstream isoform to the end of the downstream isoform, which we call DLS, depth-limited search. For each isoform pair, this search is initially called with the root node that maps to the fusion junction, the inner boundary of the mapping is used to determine the exon to which the path is added, and the set of reads mapping to the exons.

We extend the generative model of Li et al. (2010) to develop a probabilistic model for generating read pairs (Fig. 4). We reason that a read pair is generated in four steps. First the transcript from which the pair will come, \( t_n \), is chosen. Then the starting point for the upstream read, \( s_n \), within that transcript is chosen, then the end point for the downstream read, \( e_n \), is chosen. Finally, errors are introduced and the final read pair is observed. As we only observe reads, we can consider transcript choice, starting position, ending position and read error to be hidden variables. The likelihood of a collection of read pairs, and specific values of the hidden variables can be expressed as a function of the true transcript nucleotide abundances:

\[
P(R|T,S,E) = \prod_{n=1}^{N} P(t_n|\theta) P(s_n|t_n) P(e_n|s_n) P(\ell_n|e_n) P(\delta_n|s_n,e_n)\]

Each term in this equation can be calculated in a straightforward way. The probability of a transcript \( t \) being chosen is the relative nucleotide abundance of that transcript, that is, the fraction of all nucleotides that are part of that transcript. Thus, \( P(t_n|\theta) = \theta \). Assuming that each base within a transcript is equally likely to be the starting point of the upstream read, the probability of a particular starting point is the inverse of the length of the transcript \( t_n \):

\[
P(s_n|t_n) = \frac{1}{L(t_n)}
\]

The choice of the ending point depends on the distribution of insert sizes used for sequencing and the starting point. We use \( P(\ell_n|s_n,e_n) \) to account for occasional abnormally long insert sizes, we allow one-tenth of read pairs to be read through fusions, the 3’ exon of the upstream gene and the 5’ exon of the downstream gene do not contribute to the reads’ implied insert sizes. This follows from our observation that these exons often appear truncated in read-through fusions. Finally, since different isoforms of the same gene mostly contain the same exons, duplicate exon sets can be generated by calling DLS on different isoforms. These duplicates are removed before proceeding to the next step.

### 2.2 Mapping to augmented reference

After the set of putative fusions are generated, the sequence for each is generated and added to the original set of transcripts from RefSeq. Then, the read pairs are mapped to this augmented reference. Unlike the previous mapping, Bowtie is used in paired-end mode and the default mapping stringencies are used except that up to 1500 possible mappings for each paired-end read are allowed. While the addition of the putative fusion sequences may result in the addition of thousands of additional transcripts to the reference, the total amount of sequence in the augmented reference remains smaller than the genome, and the mapping can still be carried out on a standard desktop computer. After mapping, we proceed, as discussed below, to ranking fusions based on coverage.

### 2.3 Model of paired-end RNA-Seq data

We extend the generative model of Li et al. (2010) to develop a probabilistic model for generating read pairs (Fig. 4). We reason that a read pair is generated in four steps. First the transcript from which the pair will come, \( t_n \), is chosen. Then the starting point for the upstream read, \( s_n \), within that transcript is chosen, then the end point for the downstream read, \( e_n \), is chosen. Finally, errors are introduced and the final read pair is observed. As we only observe reads, we can consider transcript choice, starting position, ending position and read error to be hidden variables. The likelihood of a collection of read pairs, and specific values of the hidden variables can be expressed as a function of the true transcript nucleotide abundances:

\[
P(R|T,S,E) = \prod_{n=1}^{N} P(t_n|\theta) P(s_n|t_n) P(e_n|s_n) P(\ell_n|e_n) P(\delta_n|s_n,e_n)\]

Each term in this equation can be calculated in a straightforward way. The probability of a transcript \( t \) being chosen is the relative nucleotide abundance of that transcript, that is, the fraction of all nucleotides that are part of that transcript. Thus, \( P(t_n|\theta) = \theta \). Assuming that each base within a transcript is equally likely to be the starting point of the upstream read, the probability of a particular starting point is the inverse of the length of the transcript \( t_n \):

\[
P(s_n|t_n) = \frac{1}{L(t_n)}
\]

The choice of the ending point depends on the distribution of insert sizes used for sequencing and the starting point. We use \( P(\ell_n|s_n,e_n) \) to account for occasional abnormally large insert sizes, we allow one-tenth of read pairs to be read through fusions, the 3’ exon of the upstream gene and the 5’ exon of the downstream gene do not contribute to the reads’ implied insert sizes. This follows from our observation that these exons often appear truncated in read-through fusions. Finally, since different isoforms of the same gene mostly contain the same exons, duplicate exon sets can be generated by calling DLS on different isoforms. These duplicates are removed before proceeding to the next step.

### 2.2 Mapping to augmented reference

After the set of putative fusions are generated, the sequence for each is generated and added to the original set of transcripts from RefSeq. Then, the read pairs are mapped to this augmented reference. Unlike the previous mapping, Bowtie is used in paired-end mode and the default mapping stringencies are used except that up to 1500 possible mappings for each paired-end read are allowed. While the addition of the putative fusion sequences may result in the addition of thousands of additional transcripts to the reference, the total amount of sequence in the augmented reference remains smaller than the genome, and the mapping can still be carried out on a standard desktop computer. After mapping, we proceed, as discussed below, to ranking fusions based on coverage.

### 2.3 Model of paired-end RNA-Seq data

We extend the generative model of Li et al. (2010) to develop a probabilistic model for generating read pairs (Fig. 4). We reason that a read pair is generated in four steps. First the transcript from which the pair will come, \( t_n \), is chosen. Then the starting point for the upstream read, \( s_n \), within that transcript is chosen, then the end point for the downstream read, \( e_n \), is chosen. Finally, errors are introduced and the final read pair is observed. As we only observe reads, we can consider transcript choice, starting position, ending position and read error to be hidden variables. The likelihood of a collection of read pairs, and specific values of the hidden variables can be expressed as a function of the true transcript nucleotide abundances:

\[
P(R|T,S,E) = \prod_{n=1}^{N} P(t_n|\theta) P(s_n|t_n) P(e_n|s_n) P(\ell_n|e_n) P(\delta_n|s_n,e_n)\]

Each term in this equation can be calculated in a straightforward way. The probability of a transcript \( t \) being chosen is the relative nucleotide abundance of that transcript, that is, the fraction of all nucleotides that are part of that transcript. Thus, \( P(t_n|\theta) = \theta \). Assuming that each base within a transcript is equally likely to be the starting point of the upstream read, the probability of a particular starting point is the inverse of the length of the transcript \( t_n \):

\[
P(s_n|t_n) = \frac{1}{L(t_n)}
\]

The choice of the ending point depends on the distribution of insert sizes used for sequencing and the starting point. We use \( P(\ell_n|s_n,e_n) \) to account for occasional abnormally large insert sizes, we allow one-tenth of read pairs to be read through fusions, the 3’ exon of the upstream gene and the 5’ exon of the downstream gene do not contribute to the reads’ implied insert sizes. This follows from our observation that these exons often appear truncated in read-through fusions. Finally, since different isoforms of the same gene mostly contain the same exons, duplicate exon sets can be generated by calling DLS on different isoforms. These duplicates are removed before proceeding to the next step.
Then, the E-step consists of calculating the log-likelihood weighted by these values for individual reads.

To expand the probability distribution to \( N \) read pairs, we take the product of values for individual reads.

\[
P(R, T, S | E(\theta)) = \prod_{n=1}^{N} \prod_{t \in R_n} \prod_{s \in T_n} \prod_{e \in E_n} \theta \left( t_{i, t}, s_{i, s}, e_{i, e} \right)
\]

Finally, the probability of our observed variable, the read pairs, given the transcript abundances can be calculated by summing over the values of the hidden variables.

\[
P(R|\theta) = \sum_{Z_{\text{NA}} \in \mathbb{Z}_{\text{NA}}} \sum_{Z_{\text{NF}} \in \mathbb{Z}_{\text{NF}}} \sum_{Z_{\text{NB}} \in \mathbb{Z}_{\text{NB}}} \prod_{n=1}^{N} \prod_{t \in R_n} \prod_{s \in T_n} \prod_{e \in E_n} \theta \left( t_{i, t}, s_{i, s}, e_{i, e} \right)
\]

We seek to find the set of transcript abundances, \( \theta \), that maximizes this probability by applying EM to the results of the paired-end mapping to the reference augmented with the putative fusions.

### 2.4 EM

For consistency, we use notation similar to that used by Li et al. (2010). Let \( Z_{\text{NA}} = 1 \) if \( (t_{i, t}, s_{i, s}, e_{i, e}) = (i, j, k) \). Then, as the first step of the EM algorithm, we find the expected values of \( Z_{\text{NA}} \) given the observed reads and the current estimate of \( \theta \).

\[
E_{\text{EM}}(\theta^{(t)}) = \frac{\theta'(i, j, k) \prod_{n=1}^{m} \prod_{t \in R_n} \prod_{s \in T_n} \prod_{e \in E_n} \theta \left( t_{i, t}, s_{i, s}, e_{i, e} \right)}{\sum_{n=1}^{m} \sum_{t \in R_n} \sum_{s \in T_n} \sum_{e \in E_n} \theta \left( t_{i, t}, s_{i, s}, e_{i, e} \right)}
\]

The E-step consists of calculating the log-likelihood weighted by these values.

\[
Q(\theta^{(t+1)}) = \sum_{n=1}^{m} \sum_{t \in R_n} \sum_{s \in T_n} \sum_{e \in E_n} E_{\text{EM}}(\theta^{(t)}) \log(\theta \left( t_{i, t}, s_{i, s}, e_{i, e} \right))
\]

The values for \( \theta^{(t+1)} \) are then found by finding the \( \theta \) that maximizes this function subject to the constraint \( \sum_{i=1}^{\infty} \theta(i) = 1 \) using Lagrange multipliers.

\[
\Delta = Q(\theta^{(t)}) + \lambda \left( \sum_{i=1}^{\infty} \theta(i) - 1 \right)
\]

\[
\frac{\partial \Delta}{\partial \theta(i)} = \sum_{n=1}^{m} \sum_{t \in R_n} \sum_{s \in T_n} \sum_{e \in E_n} E_{\text{EM}}(\theta^{(t)}) \frac{1}{\theta(i)}
\]

Equating all of these terms to zero, we have

\[
\theta^{(t+1)} = \frac{\sum_{n=1}^{m} \sum_{t \in R_n} \sum_{s \in T_n} \sum_{e \in E_n} E_{\text{EM}}(\theta^{(t)}) | Z_{\text{NA}} \cap Z_{\text{NF}} \cap Z_{\text{NB}} |}{\sum_{n=1}^{m} \sum_{t \in R_n} \sum_{s \in T_n} \sum_{e \in E_n} E_{\text{EM}}(\theta^{(t)}) | Z_{\text{NA}} \cap Z_{\text{NF}} \cap Z_{\text{NB}} |} \frac{1}{N}
\]

This procedure is repeated until convergence. We make the probability calculations tractable by only considering, for each read, the values of \( r, s \) and \( e \) reported by short read mapping software and assuming the probability of the read coming from any other position to be zero.

### 2.5 Calculating mappings to fusion junctions

After convergence of the EM algorithm, we have an estimate of the maximum-likelihood abundances for each transcript, including all of the putative fusion transcripts. These abundances reflect the resolution of read mapping ambiguity, as demonstrated by the successful elimination of many spurious fusions in the results below. However, they do not yet account for potential unevenness of coverage across a given transcript. In particular, they can be confounded by a fusion transcript with high coverage everywhere but the fusion site. To illustrate this issue, consider the situation illustrated in Figure 5. We have three reference transcripts: Gene A, Gene B and a fusion gene created by concatenating Genes A and B. We also have three sets of read pairs: \( N_A \) pairs that map to Gene A and the fusion gene, \( N_B \) pairs that map to Gene B and the fusion gene and \( N_F \) pairs that only map to the fusion gene. For simplicity, assume that the values of \( \epsilon(n, i, j, k) \) and \( d(\ell(n, i, j, k)) = 1 \) for each mapping of each read pair and the length of both Genes A and B is 1.

Then, the probability of the observed data is

\[
P(R|\theta) = \theta(\phi_A + \frac{1}{2} \phi^B) N_A \theta(\phi_B + \frac{1}{2} \phi^B) N_B \theta(\phi_f + \frac{1}{2} \phi^f) N_F
\]

If we further assume that \( N_A = N_B \) and therefore \( \theta_A = \theta_B \), and use the fact that the sum of the transcript abundances must be 1, we have that \( \theta_A = \frac{1}{2\theta_B} \).

Then, the probability of the observed data becomes

\[
P(R|\theta) = (1) \theta_A \left( \frac{1}{2} \phi^B \right)^{N_A} \theta_B \left( \frac{1}{2} \phi^B \right)^{N_B} \theta_f \left( \frac{1}{2} \phi^f \right)^{N_F}
\]

This expression is maximized by setting \( \theta_A \) to 1, which sets \( \theta_A = \theta_B \) to zero. So, if there is a single read pair that spans the fusion site in this scenario, all abundance is transferred to the fusion transcript regardless of how large \( N_A \) and \( N_B \) may be in relation to \( N_F \). While this example has been rather stringently defined for sake of demonstration, a similar situation occurs whenever \( N_F \) >> 0 and \( N_A >> N_F \) or \( N_B >> N_F \); an unreasonable abundance is assigned to the fusion transcript based on reads that do not map to the fusion site. In the context of seeking fusions, this means a fusion between highly expressed genes supported by a single read pair, perhaps an experimental artifact, will dominate other putative fusions in abundance. To avoid this, rather than simply using the maximum-likelihood abundances, we calculate the sum of the expected values of \( Z_{\text{NA}} \) for each fusion transcript \( t \) for read
pairs that span the fusion junction to get a probabilistically weighted count of reads supporting the fusion, \( C_i \):

\[
C_i = \frac{\sum_{j,k} E_{j,k} \left| \mathbb{E} \right|_{j,k} }{n \cdot \text{pairs spanning junction}}
\]

This retains the ambiguity resolution described above but focuses the abundance estimates on fusions.

As a final filtering step to eliminate experimental artifacts, we find the mean physical coverage, that is, the coverage counting both the reads and the insert, for the upstream and downstream genes in the fusion separately and compare each of them to the physical coverage at the fusion site. If coverage at the fusion site is less than one-twentieth of the upstream and downstream coverage, we discard the fusion as a probable artifact based on the same reasoning discussed above. We also discard fusions where all reads have the sequence of an RNA component of the spliceosome, U1 through U6, as these are likely produced artifactually as well.

3 RESULTS

3.1 Fusion transcripts generate ambiguous reads

To quantify the prevalence of ambiguously mapping read pairs and the extent to which discarding them would impact fusion discovery, we simulated gene fusions by randomly selecting a pair of transcripts from RefSeq and the exon within each transcript that would serve as the fusion breakpoint. For each fusion, we generated, with random errors based on quality scores from an existing dataset, the full set of read pairs that could span it given a constant insert size. We then mapped each of these reads and tabulated the number of read pairs with unique mappings that satisfy default Bowtie mapping criteria (Langmead et al., 2009). We repeated this for several read lengths, generating 100 000 simulated fusions for each read length, while keeping the insert between the two reads at 200 nt.

For each read length, we calculated the fraction of partially ambiguous fusions and totally ambiguous fusions, that is, fusions where some, but not all, of the reads supporting them mapped ambiguously and fusions that only generated ambiguously mapping read pairs. As expected, the fraction of ambiguous fusions declined as read length increased. At a read length of 50 nt, nearly 1 in 20 fusions would only be detectable via ambiguously mapping read pairs (Table 1, Supplementary Table S1). Even at a read length of 100 nt, over a 10-th of all fusions were able to generate an ambiguously mapping read pair. These results suggest that even as read lengths increase, a significant portion of fusions remain difficult to detect if read pairs are required to map unambiguously.

Table 1. The fraction of totally and partially ambiguous fusions for a range of read lengths

<table>
<thead>
<tr>
<th>Read length</th>
<th>% Partially ambiguous fusions</th>
<th>% Totally ambiguous fusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>30.3</td>
<td>5.7</td>
</tr>
<tr>
<td>35</td>
<td>22.4</td>
<td>5.5</td>
</tr>
<tr>
<td>40</td>
<td>17.5</td>
<td>5.1</td>
</tr>
<tr>
<td>45</td>
<td>14.9</td>
<td>4.8</td>
</tr>
<tr>
<td>50</td>
<td>13.4</td>
<td>4.5</td>
</tr>
<tr>
<td>75</td>
<td>9.4</td>
<td>3.7</td>
</tr>
<tr>
<td>100</td>
<td>7.9</td>
<td>2.9</td>
</tr>
</tbody>
</table>

3.2 Resolving ambiguous simulated fusions

To demonstrate the capability of our method to find gene fusions between highly repetitive regions of the transcriptome using multimegapread mapping read pairs, we simulated five fusion genes, outlined in Table 2, derived from possible fusions between genes that share homology with other parts of the transcriptome. Then, 10 000 pairs of 40 nt reads were generated from these five fusions using MAQ (Li et al., 2008) in simulate mode with insert size set to 200 nt. Sequencing errors and quality values were modeled from an existing dataset, and the MAQ simulation code was modified to produce a distribution of different expression levels for each transcript so performance over a range of coverage levels could be examined. As a comparison, the coverage levels used in the simulation would correspond to a range of ∼8 FPKM for MAGED4-MBD3L2 to 80 FPKM for FOXO3-EIF3CL in a 20 M read pair sequencing experiment. Thus, the simulated coverages provide a reasonable range on which to evaluate the performance of our method.

Mapping the 10 000 read pairs to RefSeq transcripts yielded 395 pairs that mapped only discordantly. As expected, all these discordantly mapping pairs mapped to multiple genomic loci and thus suggested multiple fusion candidates. Each discordant read pair is mapped, on average, to seven different pairs of genes, and in some cases mapped to as many as 22. The total number of fusion genes that would be nominated by naively accepting all discordant mappings was 56 (Supplementary Table S2).

Applying the filtering and fusion discovery process described in the Section 2.1 yielded 252 putative fusion transcripts. The high number reflects both the multiple gene pairs to which the discordant read pairs mapped and the multiple sets of exons from each gene pair that could be consistent with the discordant mappings.

Table 2. Simulated fusions

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Pair count</th>
<th>Pairs spanning fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOXO3</td>
<td>EIF3CL</td>
<td>7152</td>
<td>281</td>
</tr>
<tr>
<td>PSG2</td>
<td>PHB</td>
<td>1324</td>
<td>117</td>
</tr>
<tr>
<td>FRG1</td>
<td>USP6</td>
<td>803</td>
<td>47</td>
</tr>
<tr>
<td>SMN2</td>
<td>CSAG1</td>
<td>434</td>
<td>78</td>
</tr>
<tr>
<td>MAGED4</td>
<td>MBD3L2</td>
<td>286</td>
<td>34</td>
</tr>
</tbody>
</table>
Table 3. Sum of expected values of $Z_{ijk}$ for read pairs supporting each fusion after maximum-likelihood transcript abundance estimation

<table>
<thead>
<tr>
<th>Upstream partner</th>
<th>Downstream partner</th>
<th>Supporting read pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOXO3</td>
<td>EIF5C</td>
<td>180.3</td>
</tr>
<tr>
<td>P5G2</td>
<td>PHB</td>
<td>117.0</td>
</tr>
<tr>
<td>FOXO3</td>
<td>EIF5CL</td>
<td>100.6</td>
</tr>
<tr>
<td>SMN1</td>
<td>CSAG1</td>
<td>56.6</td>
</tr>
<tr>
<td>FG1</td>
<td>USP6</td>
<td>46.9</td>
</tr>
<tr>
<td>MAGED4</td>
<td>MBD3L2</td>
<td>34.0</td>
</tr>
<tr>
<td>SMN2</td>
<td>CSAG1</td>
<td>21.4</td>
</tr>
</tbody>
</table>

Table 4. Prostate neoplasia fusions with sum of expected $Z_{ijk}$ values

<table>
<thead>
<tr>
<th>Upstream partner</th>
<th>Downstream partner</th>
<th>Supporting read pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMPRSS2</td>
<td>ERG</td>
<td>49.0</td>
</tr>
<tr>
<td>AZGP1</td>
<td>GJC3</td>
<td>28.0</td>
</tr>
<tr>
<td>TTY14</td>
<td>NCRNA00185</td>
<td>8.0</td>
</tr>
<tr>
<td>LOC728606</td>
<td>KCTD1</td>
<td>4.0</td>
</tr>
<tr>
<td>ZNF564</td>
<td>ZNF577</td>
<td>3.0</td>
</tr>
<tr>
<td>SMA4</td>
<td>ETF2HE8</td>
<td>2.5</td>
</tr>
<tr>
<td>LOC100134368</td>
<td>NME4</td>
<td>2.0</td>
</tr>
<tr>
<td>SYJNBP</td>
<td>COX16</td>
<td>2.0</td>
</tr>
<tr>
<td>SMG5</td>
<td>PAQR6</td>
<td>2.0</td>
</tr>
<tr>
<td>PRKAA1</td>
<td>TCC33</td>
<td>2.0</td>
</tr>
<tr>
<td>LOC401588</td>
<td>CHST7</td>
<td>2.0</td>
</tr>
<tr>
<td>HARS2</td>
<td>ZMAT2</td>
<td>2.0</td>
</tr>
<tr>
<td>NQCRQ</td>
<td>LEAP2</td>
<td>2.0</td>
</tr>
<tr>
<td>GRHL2</td>
<td>SNTG1</td>
<td>2.0</td>
</tr>
<tr>
<td>KLK4</td>
<td>KLKP1</td>
<td>2.0</td>
</tr>
</tbody>
</table>

ones and retained two fusions that are indistinguishable from true fusions.

3.3 Application to a prostate tissue transcriptome data

We applied our method to two datasets derived from tissue resected from an individual with prostate cancer. The first dataset consisted of 18 027 834 pairs of 40 nt reads from neoplastic tissue. The second was 21 978 463 read pairs from adjacent hyperplastic tissue. Of the neoplasia read pairs, 18 177 had only discordant mappings and mapped to 127 102 gene pairs. Of the hyperplasia read pairs, 24 569 had only discordant mappings and mapped to 266 571 gene pairs. Application of the filtering and fusion discovery process described above yielded 887 and 746 putative fusion transcripts for neoplasia and hyperplasia, respectively. After estimating transcript abundances, only 15 fusion transcripts from the neoplasia data had at least two reads assigned to its junction site (Table 4). The top result, a fusion between TMPRSS2 and ERG, is a known recurrent fusion in prostate cancer (Tomlins et al., 2005). A novel fusion between GRHL2 and SNTG1 was also reported. These genes lie about 50 Mb apart on chromosome 8. Intriguingly, there is a short sequence shared by both sequences at the site of the fusion (Supplementary Fig. S2), potentially providing a clue to the origin of the chimera (Li et al., 2010).

Table 5. Prostate hyperplasia fusions with sums of expected $Z_{ijk}$ values

<table>
<thead>
<tr>
<th>Upstream partner</th>
<th>Downstream partner</th>
<th>Supporting read pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>AZGP1</td>
<td>GJC3</td>
<td>54.0</td>
</tr>
<tr>
<td>SPINT2</td>
<td>C1orf33</td>
<td>6.8</td>
</tr>
<tr>
<td>RPL7</td>
<td>LOC100130301</td>
<td>3.0</td>
</tr>
<tr>
<td>TMEM203</td>
<td>C1orf75</td>
<td>3.0</td>
</tr>
<tr>
<td>DHR51</td>
<td>RABGCTA</td>
<td>3.0</td>
</tr>
<tr>
<td>IRF6</td>
<td>C1orf74</td>
<td>2.0</td>
</tr>
</tbody>
</table>

In sharp contrast to the neoplasia results, the hyperplasia data showed no evidence of a fusion between TMPRSS2 and ERG (Table 5). This is consistent with the central role that the TMPRSS2-ERG fusion is suspected to play in the progression of prostate cancer (Yu et al., 2010). Beyond this critical difference, the results largely mirrored those from neoplasia. There was one novel read-through transcript reported, RPL7-LOC100130301, and multiple previously reported read-throughs: AZGP1-GJC3, SPINT2-C1orf33, DHR51-RABGCTA, TMEM203-C1orf75 and IRF6-C1orf74. The large number of potential fusions suggested by a naive examination of discordant reads, over 100 000 in each dataset, underscores the complexity of the transcriptome and the often muddled nature of experimentally derived transcriptomic sequencing data. We were gratified that our method was able to discard nearly all of these inauthentic fusions while retaining those of biological importance.

3.4 Discovery of novel ambiguous fusions

To demonstrate the ability of our method to make new discoveries, we analyzed two publicly available datasets. The first was transcriptome sequencing of a set of melanoma tumors and cell lines originally published by Berger et al. (2010). The second was sequencing of Stratagene’s Universal Human Reference RNA (UHR), a reference composed of RNA from 10 cell lines originally published by Bullard et al. (2010). Analysis of these data with our method yielded numerous fusions, including all of the fusions reported by Berger and numerous fusions known to be present in UHR including BCR-ABL1, BCAS4-BCAS3 and GAS6-RASA3 (Supplementary Table S3). In each case, the ambiguity was due to genomic duplications. Some reads mapping to the MYH6 side of the HOMEZ-MYH6 fusion also mapped to MYH6’s paralog, MYH7 (Fig. 7). The remaining ambiguous fusions were due to recent segmental duplications. The fusion between CPEB1 and RPS17 was clearly a read-through, but was confounded by the presence of another copy of RPS17 in an upstream segmental duplication (Fig. 7). KIAA1267-ARL17A was similarly made ambiguous by multiple copies of ARL17. The fusions between PRP5K1-CATSPER2 and TRIM16L-FBXW10 were confounded by mappings to CATSPER2P1 and TRIM16-LCDRT1. The sequence of each fusion is available in Supplementary Fig. S3. These findings confirm that additional fusions can be detected in tumors when ambiguously mapping read pairs are included in the analysis.
Table 6. Fusions found in previously published datasets that are either partially or completely supported by ambiguously mapping read pairs

<table>
<thead>
<tr>
<th>Fusion</th>
<th>Samples</th>
<th>Supporting read pairs</th>
<th>Ambiguous read pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOMEZ-MYH6</td>
<td>UHR</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>KIAA1267-ARL17A</td>
<td>M010403</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>UHR</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>CPEB1-RPS17</td>
<td>M980409</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>MeWo</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>PIPPSK1-ACTSPER2</td>
<td>M010403</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>M990802</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>TRIM16L-FBXW10</td>
<td>M010403</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Fig. 6. The fusion between HOMEZ and MYH6. Three mate pairs support this fusion, but two also map to a fusion between HOMEZ and MYH7.

Fig. 7. The fusion between CPEB1 and RPS17. A copy of RPS17 lies 2000 bases downstream of CPEB1, but another copy lies 400 kb downstream, as well.

4 DISCUSSION

In this article, we have demonstrated a method to use discordantly and often ambiguously mapping RNA-Seq read pairs to identify fusion transcripts. In doing so, we bring the increasingly sophisticated methods employed to estimate transcript abundance in the presence of multimapping reads to the problem of fusion discovery. In contrast to previously proposed methods for fusion identification that focus on reads that map to the junction between two genes (Ameur et al., 2010), our method estimates fusion transcript abundances by considering physical coverage over the entire length of the proposed fusion. In addition, it employs several filters to minimize experimental artifacts. Finally, it does not require that any single read sequence hit the point of fusion. Instead, it uses implied insert sizes and known exon boundaries to determine the most likely point of fusion. This would be a liability if a fusion transcript contained partial exons, but reported fusions to date suggest that a vast majority of fusions do indeed involve the joining of whole exons from different genes, the breakpoints occurring in introns and the splice sites remaining unchanged (Hahn et al., 2004).

Several avenues for future development are apparent from this work. Here, we chose to use RefSeq transcripts as the reference against which reads are mapped. This allowed us to avoid the issue of reads that map to splice junctions because the splice junction sequence would be contiguous in the transcript sequence. However, it prevents us from identifying transcripts that are produced by novel or aberrant splicing, which is common in cancer (Rajani et al., 2009), or are significantly altered by RNA editing (Skarda et al., 2009). It may be fruitful to combine the approach described here with methods that identify splice junctions and expressed regions of the genome de novo (Ameur et al., 2010; Trapnell et al., 2009). Additionally, fusion transcript discovery shares many parallels with the problem of resolving genomic rearrangements, especially the challenges of repetitive sequence. The adaptation of the methods developed here to genomic sequencing may prove useful in this related field.

ACKNOWLEDGEMENTS

We thank Eric Topol for his support and the clinical team at The Scripps Translational Science Institute for sample collections.

Funding: National Institutes of Health [grant numbers RO1-HG004962, 5U54HL108460, 1R21CA152613-01, CTSA-1U54RR025204] and the California Institute for Regenerative Medicine [grant number DR1-01430].

Conflict of Interest: none declared.

REFERENCES

Fusion detection using ambiguous read pairs


