BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 16 2010, pages 1950–1957
doi:10.1093/bioinformatics/btq336

Sequence analysis

Advance Access publication June 23, 2010

A probabilistic framework for aligning paired-end RNA-seq data

Yin Hu[1], Kai Wang[1], Xiaping He[2], Derek Y. Chiang[2], Jan F. Prins[3] and Jinze Liu[1,*]

[1]Department of Computer Science, University of Kentucky, Lexington, KY, [2]Department of Genetics and [3]Department of Computer Science, University of North Carolina, Chapel Hill, NC, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: The RNA-seq paired-end read (PER) protocol samples transcript fragments longer than the sequencing capability of today's technology by sequencing just the two ends of each fragment. Deep sampling of the transcriptome using the PER protocol presents the opportunity to reconstruct the unsequenced portion of each transcript fragment using end reads from overlapping PERs, guided by the expected length of the fragment.

Methods: A probabilistic framework is described to predict the alignment to the genome of all PER transcript fragments in a PER dataset. Starting from possible exonic and spliced alignments of all end reads, our method constructs potential splicing paths connecting paired ends. An expectation maximization method assigns likelihood values to all splice junctions and assigns the most probable alignment for each transcript fragment.

Results: The method was applied to 2×35 bp PER datasets from cancer cell lines MCF-7 and SUM-102. PER fragment alignment increased the coverage 3-fold compared to the alignment of the end reads alone, and increased the accuracy of splice detection. The accuracy of the expectation maximization (EM) algorithm in the presence of alternative paths in the splice graph was validated by qRT–PCR experiments on eight exon skipping alternative splicing events. PER fragment alignment with long-range splicing confirmed 8 out of 10 fusion events identified in the MCF-7 cell line in an earlier study by (Maher *et al.*, 2009).

Availability: Software available at http://www.netlab.uky.edu/p/bioinfo/MapSplice/PER

Contact: liuj@cs.uky.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 4, 2010; revised on June 17, 2010; accepted on June 21, 2010

1 INTRODUCTION

High-throughput sequencing technologies are providing unprecedented visibility into the mRNA transcriptome of a cell. In cancer, alternative splicing and gene fusion events (Berger *et al.*, 2010; Maher *et al.*, 2009) are common changes observed in the mRNA transcriptome. Cancer-specific splicing events are promising biomarkers and targets for diagnosis, prognosis and treatment purposes. Recently, several computational methods (Au *et al.*, 2010; Trapnell *et al.*, 2009) have been developed to identify splicing events using RNA-seq data. These methods align RNA-seq reads to the reference genome rather than to a transcript database, making it possible to identify novel splicing events via gapped alignment of reads to the genome.

New protocols and sequencing methods have expanded the length and type of RNA-seq reads, enabling more accurate characterization of the splices present in the transcriptome. A *single read* may constitute 35–100 consecutive nucleotides of a fragment of an mRNA transcript. The *paired-end read* (PER) protocol sequences two ends of a size-selected fragment of an mRNA transcript and reports the results as a pair. In our experiment, for example, the expected size of mRNA fragments are around 182 bp ($\pm$40 bp).[1] Both ends of the fragment are sequenced to at least 35 bp in length.

This article focuses on predicting the alignment of an entire PER fragment, starting from the alignments of its end reads and using the alignments of other overlapping PER end reads to predict an overall alignment consistent with the expected length of the fragment. Since a PER fragment can be longer than single reads sequenced with today's RNA-seq technology, achieving such alignments may significantly increase the effective transcriptome coverage. Longer alignments also decrease alignment ambiguity in regions with genome repeats.

A unique challenge in *PER fragment alignment* is that the expected distance between the two end reads within the transcript fragment, known as *mate-pair distance*, can be very different from distance between the two end reads when aligned to the genome. This can happen when the two ends fall in different exons, so that their separation in genomic coordinates includes one or more intervening introns that are not present in the transcript (Fig. 1a). This effect is illustrated as a long tail in the mate-pair distance distribution when aligned on the genome (Fig. 1b). Resolving the discrepancy between the expected mate-pair distance and the paired-end separation on the genome is not trival. RNA-seq aligners including TopHat (Trapnell *et al.*, 2009) and SpliceMap (Au *et al.*, 2010) align PERs using heuristics. When the distance between end alignments is substantially longer than the expected mate-pair distance, TopHat reports the closest end alignment for a PER, while SpliceMap considers PERs with ends mapped within 400 000 bp on the genome. While both heuristics have meaningful biological motivations, neither method predicts nor validates the PER alignment. Since both approaches discard PER alignments that span a very long interval or cross chromosomes, neither of them is capable of finding long-range splicing or gene fusion events.

In this article, we propose a new probabilistic framework for aligning RNA-seq PERs to a reference genome, without relying on transcript databases. Our goal is to discover both short-range splice

*To whom correspondence should be addressed.

[1]The fragment length is typically around 200 bp but may vary according to different PER protocols.

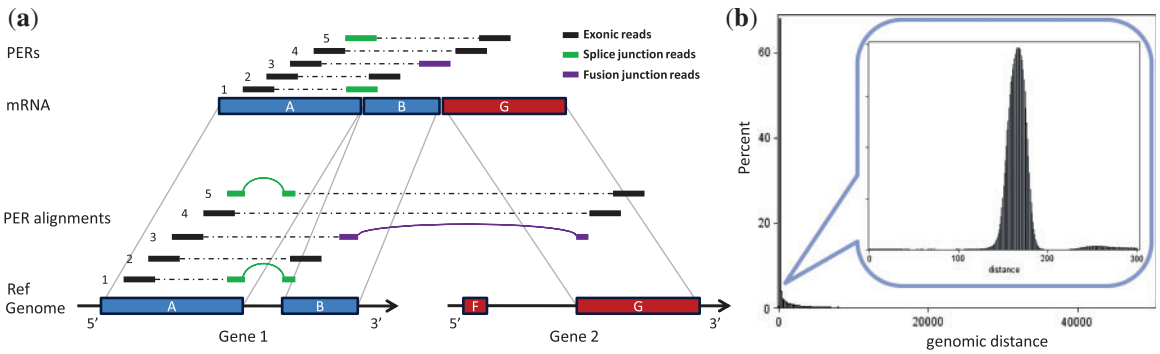


Fig. 1. (**a**) A fragment of an mRNA transcript exhibiting gene fusion between exon B in Gene 1 and exon G in Gene 2 is sampled by six PERs. The alignment of the transcript to the reference genome as well as the alignment of the PERs to the genome is shown. The unsequenced segments of PERs cannot readily be aligned to the genome because of unknown intervening splicing events including, in this case, the fusion junction. (**b**) An example of the distribution of distance in genomic coordinates between PER alignments generated from 2×35 bp PER data. While the majority of distances fall within the normal distribution for mate-pair distance on mRNA fragments, a significant portion of the distances are far beyond the expected range, indicating potential splicing events.

junctions and long-range splice/fusion junctions through accurate mapping of PER end reads as well as the unsequenced middle portion. Our approach starts by building a compact splice graph to represent all putative splicing events, regardless of the intron sizes, derived from individual end read alignments. An expectation maximization algorithm is then applied to identify the most probable path in the graph that connects the two ends of a PER based on the empirical distribution of the mate-pair distances. This in turn is used to infer the significant splice junctions.

Our approach was applied to RNA-seq datasets of 2×35 bp PER reads from MCF-7 and SUM-102, two well-known breast cancer cell lines. PER fragment alignment increased the coverage 3-fold compared to the alignment of the end reads alone, and increased the accuracy of splice detection. The accuracy of the EM algorithm in the presence of alternative paths in the splice graph was validated by qRT–PCR experiments on eight exon skipping alternative splicing events. PER fragment alignment with long-range splicing confirmed 8 out of 10 fusion events identified in the MCF-7 cell line in an earlier study by (Maher *et al.*, 2009).

2 MAPPING INDIVIDUAL READS

The alignment of RNA-seq PERs starts with the alignment of their individual end reads. MapSplice (Wang *et al.*, 2010) was used to map these end reads to the reference genome, generating both the read alignment and putative splice and fusion junctions.

MapSplice finds both exonic and spliced alignments of RNA-seq reads to a reference genome without any dependence on annotations or structural features of the genome. MapSplice operates by partitioning RNA-seq reads into short segments (18–25 bp) that are aligned directly to the reference genome. Segments that can be aligned in this fashion are likely to be transcribed from exonic regions. Segments that cannot be aligned in the first step may contain a splice junction that is located by a search extending from aligned neighboring segment(s). In general, each segment may end up with multiple alignments that exceed some alignment quality threshold σ. A merge phase constructs candidate alignments for each read, by combining consistent alignments of its segments. Splice junctions are given a confidence value by considering the quality and diversity of all candidate alignments that include the junction. Finally, the candidate alignments for a read are restricted to those in which the overall alignment quality and the confidence of any included splice junctions exceeds σ. The alignment of end reads by MapSplice was performed globally, i.e. without constraints on the proximity or strand of the mate-pair alignments in genomic coordinates. Given a PER (x_α, x_β), the alignments of x_α and x_β fall into one of the following four categories.

1. x_α and x_β are mapped onto the same chromosome and the same strand, and the mapped distance on the genome is close to their expected mate-pair distance (as shown in Fig. 2a).
2. x_α and x_β are mapped onto the same chromosome and the same strand with a distance much longer than the expected mate-pair distance. This indicates the x_α and x_β span distinct exons (as shown in Fig. 2b). When the distance is larger than 50 000 bp, the two reads are assumed to be from different genes. Similar rules were used by (Maher *et al.*, 2009).
3. x_α and x_β are mapped onto different chromosomes. This indicates a potential *trans*-chromosome fusion event (as shown in Fig. 2c).
4. x_α and x_β are mapped onto different strands, either of a same chromosome or of different chromosomes. This indicates a potential *trans*-strand chimeric event (as shown in Fig. 2d).

In the first category, the alignment of a PER fragment can easily be determined since their separation is concordant with the expected mate-pair distances. For the remaining categories, the alignment of the complete PER fragment requires knowledge of the intervening exons and splicing structure to reconstruct plausible alignments. The set of splice junctions can be inferred from the spliced alignment of PER end reads. Reads 1 and 3 in Figure 1a are examples of splice junction reads, while Read 5 is an example of a fusion junction read. However, due to alternative splicing, multiple splicing paths may exist from x_α to x_β. Furthermore, the mapping of individual end reads may have multiple alignments to the genome due to repeats and homologous genes. To address these problems, we propose a maximum likelihood approach to disambiguate the PER alignments, detailed in Section 3.

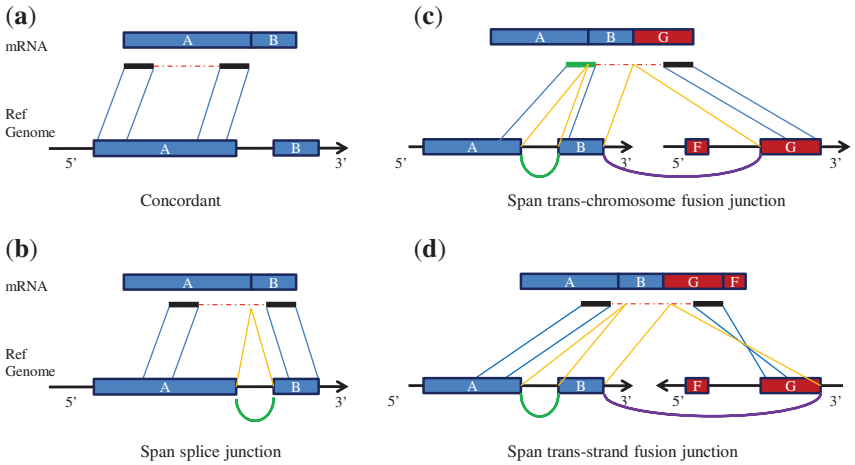


Fig. 2. An illustration of a PER fragment alignment to the reference genome. The mRNA transcript is shown at the top, the PER sequence is shown in the middle and the alignment of the PER to the genome is shown at the bottom. Four cases are shown: (**a**) concordant with mRNA alignment distance; (**b**) crossing a splice junction; (**c**) crossing *trans*-chromosome fusion junction; and (**d**) crossing *trans*-strand chimeric junction.

3 PROBABILISTIC FRAMEWORK

3.1 Graphical model and notations

The spliced alignments of individual end reads result in a putative set of splice and fusion junctions. These junctions can be used to build a splice graph $G=\langle V,E\rangle$ to reflect the relation between the genome and transcript fragments. Within the splice graph G, each node $v\in V$ corresponds to a base on the reference genome. The nodes are connected by directed edges in the direction of the transcription. There exist two types of directed edges. The first type represents the connections between two adjacent bases on the same chromosome. The second type of edge corresponds to splice or fusion junctions, and skips around the spliced-out portion of the genome.

Let D be the set of RNA-seq PERs. Let x_α and x_β be the two end reads of transcript fragment x, $\langle x_\alpha,x_\beta\rangle\in D$. We denote the unsequenced segment of x as x_γ. Therefore, the entire PER fragment of x is the concatenation of x_α, x_γ and x_β, and must be arranged in precisely this order, i.e. $x=\langle x_\alpha,x_\gamma,x_\beta\rangle$. Figure 3 illustrates the alignment of a PER based on the constructed splice graph. We are interested in predicting the alignment of entire fragment x including unsequenced x_γ as well as x_α and x_β.

Let Π_x^α and Π_x^β be the sets of valid alignments of end reads x_α and x_β, respectively. The set of putative end read alignments of a PER contains all the unique combinations of the mapped locations of x_α and x_β, i.e.

$$\Pi_x^{\alpha,\beta} = \{\langle\pi_x^\alpha,\pi_x^\beta\rangle|\pi_x^\alpha\in\Pi_x^\alpha,\pi_x^\beta\in\Pi_x^\beta\}.$$

Determining the alignment of x_γ is not straightforward since it is not sequenced. Its alignment might be predicted given the mapping of the end reads π_x^α and π_x^β and the splicing paths connecting them. We use $\Pi_x^{\gamma|\alpha,\beta}$ to denote the set of candidate alignments of x_γ given π_x^α and π_x^β, each of which corresponds to a unique concatenation of exonic regions by following a particular splicing path. A putative alignment of a PER x, π_x, therefore, is equivalent to an acyclic path that starts with the first base of π_x^α, passes π_x^γ and ends with the last base of π_x^β. Formally, given the set of end read alignments $\Pi_x^{\alpha,\beta}$, the set of candidate alignments of x, Π_x, is

$$\Pi_x=\{\pi_x|\pi_x^{\alpha,\beta}\in\Pi_x^{\alpha,\beta},\pi_x^\gamma\in\Pi_x^{\gamma|\alpha,\beta}\}.$$

Problem definition: let $\Pi=\{\pi_x|\langle x_\alpha,x_\beta\rangle\in D\}$ be the set of candidate fragment alignments for all PERs in D.1 Our goal is to determine an alignment for each PER, $\hat{\Pi}$, that maximizes the likelihood of the alignment of all the PERs in D, i.e.

$$\hat{\Pi}=\arg\max_{\Pi}\prod_{x\in D}P(x|\Pi). \quad (1)$$

3.2 Probability definitions

Probability of a PER: the probability of a PER x is determined by its end read alignments $\Pi_x^{\alpha,\beta}$. By summing up the probability that a read alignment $\pi_x^{\alpha,\beta}\in\Pi_x^{\alpha,\beta}$ is the true alignment $\hat{\pi}_x^{\alpha,\beta}$ at each candidate alignment in $\Pi_x^{\alpha,\beta}$, the probability of x can be computed as

$$\begin{aligned}P(x) &= \sum_{\pi_x^{\alpha,\beta}\in\Pi_x^{\alpha,\beta}} P(x,\hat{\pi}_x^{\alpha,\beta}=\pi_x^{\alpha,\beta})\\ &= \sum_{\pi_x^{\alpha,\beta}\in\Pi_x^{\alpha,\beta}} P(x|\hat{\pi}_x^{\alpha,\beta}=\pi_x^{\alpha,\beta})\cdot P(\hat{\pi}_x^{\alpha,\beta}=\pi_x^{\alpha,\beta}).\end{aligned}$$

Here, $P(\hat{\pi}_x^{\alpha,\beta}=\pi_x^{\alpha,\beta})$ is the expected probability that x is aligned to $\pi_x^{\alpha,\beta}$. It is estimated at the expectation step of EM algorithm described in Section 3.3. The probability of x's alignment given $\pi_x^{\alpha,\beta}$, $P(x|\hat{\pi}_x^{\alpha,\beta}=\pi_x^{\alpha,\beta})$, is determined by

- the probability of the accurate alignments for both end reads, x_α and x_β;
- the probability of the alignment for unsequenced portion, x_γ.

Mathematically, assuming the assessment of x_α, x_β and x_γ are independent,

$$\begin{aligned}P(x|\hat{\pi}_x^{\alpha,\beta}=\pi_x^{\alpha,\beta}) &= P(x_\alpha|\hat{\pi}_x^\alpha=\pi_x^\alpha)\cdot P(x_\beta|\hat{\pi}_x^\beta=\pi_x^\beta)\\ &\quad\cdot P(x_\gamma|\hat{\pi}_x^{\alpha,\beta}=\pi_x^{\alpha,\beta}).\end{aligned}$$

We first determine $P(x_\gamma|\hat{\pi}_x^{\alpha,\beta}=\pi_x^{\alpha,\beta})$, the probability of x_γ given $\pi_x^{\alpha,\beta}$. x_γ is the unsequenced portion of x. Its alignment, π_x^γ , would be one of the putative splicing paths connecting π_x^α and π_x^β, assuming the necessary splice junctions are present. Since the length of π_x^γ corresponds to the *mate-pair distance*, for each putative alignment π_x^γ, the probability $P(x_\gamma|\pi_x^{\alpha,\beta,\gamma})$ may be determined by the length of π_x^γ in the empirical distribution of the mate-pair distances $\mathcal{N}_d$. Here, we denote it as $P_d(\pi_x^\gamma)$. Therefore, the probability of x_γ given end read alignment $\pi_x^{\alpha,\beta}$ can be expressed as

$$\begin{aligned}&P(x_\gamma|\hat{\pi}_x^{\alpha,\beta}=\pi_x^{\alpha,\beta})\\ &= \sum_{\pi_x^\gamma\in\Pi_x^{\gamma|\alpha,\beta}} P(x_\gamma,\hat{\pi}_x^\gamma=\pi_x^\gamma|\hat{\pi}_x^{\alpha,\beta}=\pi_x^{\alpha,\beta})\\ &= \sum_{\pi_x^\gamma\in\Pi_x^{\gamma|\alpha,\beta}} P(x_\gamma|\hat{\pi}_x=\pi_x^{\alpha,\beta,\gamma})\cdot P(\hat{\pi}_x^\gamma=\pi_x^\gamma|\hat{\pi}_x^{\alpha,\beta}=\pi_x^{\alpha,\beta})\\ &= \sum_{\pi_x^\gamma\in\Pi_x^{\gamma|\alpha,\beta}} P_d(\pi_x^\gamma)\cdot P(\hat{\pi}_x^\gamma=\pi_x^\gamma|\hat{\pi}_x^{\alpha,\beta}=\pi_x^{\alpha,\beta})\end{aligned}$$

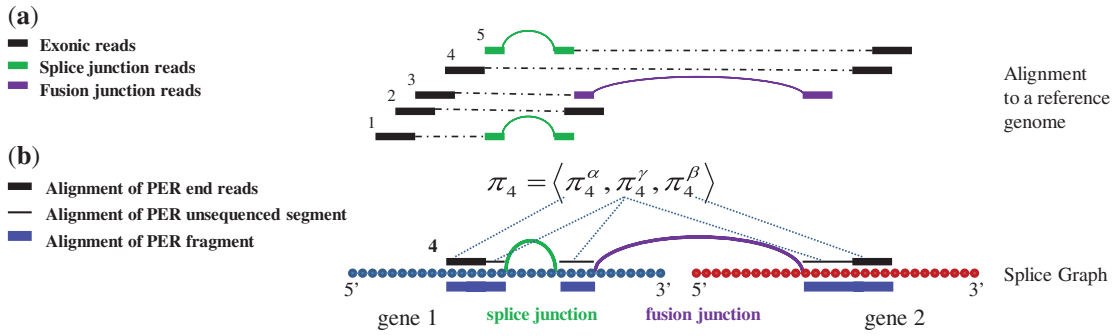


Fig. 3. An illustration of the framework proposed in Section 3 applied to the example in Figure 1. The input is a set of RNA-seq PERs that have both ends aligned to the reference genome (**a**). A splice graph can be constructed by taking each base as a node and connecting adjacent bases in the same chromosome as well as bases that constitute a potential splice junction or fusion junction (**b**). A candidate alignment of a PER is a path in the splice graph from its start position to end position with the proper orientation.

where $P(\hat{\pi}_x^\gamma = \pi_x^\gamma | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})$ is the probability of an alignment π_x^γ given the end read alignment $\pi_x^{\alpha,\beta}$. We will determine $P(\hat{\pi}_x^\gamma = \pi_x^\gamma | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})$ during the maximization step of EM algorithm described in Section 3.3.

The probability of the sequenced end reads, $P(x_\alpha | \hat{\pi}_x^\alpha = \pi_x^\alpha)$ and $P(x_\beta | \hat{\pi}_x^\beta = \pi_x^\beta)$ should be evaluated first based on their alignments, i.e. the probability of an alignment is not erroneous given their sequence similarity to the reference genome and their base call quality score (Li *et al.*, 2008), denoted as $P_q(x_\alpha | \pi_x^\alpha)$. In case a read spans one or more splice junctions or fusion junctions, the probability of a read is also dependent upon the joint probability of these junctions. Let $\Lambda(\pi_x^\alpha)$ be the set of junctions spanned by the end read alignment π_x^α. Considering both the spliced alignment and the matching quality, the probability of an accurate end read alignment can be calculated as

$$P(x_\alpha | \hat{\pi}_x^\alpha = \pi_x^\alpha) = \begin{cases} P_q(x_\alpha | \pi_x^\alpha), & \Lambda(\pi_x^\alpha) = \emptyset; \\ P_q(x_\alpha | \pi_x^\alpha) \cdot \prod_{\lambda \in \Lambda(\pi_x^\alpha)} P(\lambda), & \text{otherwise.} \end{cases}$$

The probability $P(x_\beta | \hat{\pi}_x^\beta = \pi_x^\beta)$ can be calculated similarly.

Splice junction probability: splice junctions are derived from the spliced alignment of end reads to the reference genomes without relying on existing annotations. Such approach enables us to discover novel junctions but some of these junctions might be false positives. For example, if a junction has few and/or low probability PER supports, it may be spurious. On the other hand, a junction is likely to be true if it is crossed by at least one PER alignment with high probability. Therefore, we may evaluate the probability of a junction based on the set of PERs crossing it.

Mathematically, let $\Pi(\lambda)$ be the set of PER alignments going through the junction λ,

$$\Pi(\lambda) = \{\pi_x | \lambda \text{ is crossed by } \pi_x\}.$$

For each alignment $\pi_x = \pi_x^{\alpha,\gamma,\beta}$ in $\Pi(\lambda)$, the junction λ may be crossed in a spliced alignment of either π_x^α and π_x^β or be part of the splicing path of π_x^γ.

The *probability of the junction* λ can be expressed as the probability that there is at least one PER alignment π_x in $\Pi(\lambda)$ supporting the junction, i.e.,

$$\begin{aligned} P(\lambda) &= 1 - \prod_{\pi_x \in \Pi(\lambda)} (1 - P(x, \hat{\pi}_x = \pi_x)) \\ &= 1 - \prod_{\pi_x \in \Pi(\lambda)} (1 - P(x | \hat{\pi}_x = \pi_x) \\ &\quad \cdot P(\hat{\pi}_x = \pi_x)) \end{aligned}$$

where

$$P(x | \hat{\pi}_x = \pi_x) = P(x_\alpha | \hat{\pi}_x^\alpha = \pi_x^\alpha) P(x_\beta | \hat{\pi}_x^\beta = \pi_x^\beta) P(x_\gamma | \hat{\pi}_x^{\gamma|\alpha,\beta} = \pi_x^{\gamma|\alpha,\beta}),$$

i.e., the probability that x is true at the alignment $\pi_x = \pi_x^{\alpha,\gamma,\beta}$.

In the next section, we will discuss an expectation maximization approach that determines the alignment for each PER maximizing the probability of all PERs as in Equation 1.

3.3 Probability estimation

In this section, we apply the EM algorithm (Dempster *et al.*, 1977; Wu , 1983) to maximize the log likelihood of all the sampled PERs. The dependency relationships of all the variables are summarized in Figure 4.

3.3.1 Initialization The probability of a PER is dependent upon the joint probability of the junctions within the span of the PER end read alignment. And the probability of a junction is calculated based on the probabilities of the PERs supporting the junction. In order to start the maximization, we initiate the probability of each junction as 1, and calculate the probability of each PER alignment.

At the alignment $\pi_x^{\alpha,\beta}$, the probability that the PER x takes π_x^γ as the unsequenced segment alignment is initiated with the expectation

$$P(\hat{\pi}_x^\gamma = \pi_x^\gamma | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}) = \frac{P(x_\gamma, \hat{\pi}_x^\gamma = \pi_x^\gamma | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})}{\sum_{\tilde{\pi}_x^\gamma \in \Pi_x^{\gamma|\alpha,\beta}} P(x_\gamma, \hat{\pi}_x^\gamma = \tilde{\pi}_x^\gamma | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})}.$$

Meanwhile the expected probability that $\pi_x^{\alpha,\beta}$ is true is estimated by

$$P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}) = \frac{P(x | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})}{\sum_{\tilde{\pi}_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} P(x | \hat{\pi}_x^{\alpha,\beta} = \tilde{\pi}_x^{\alpha,\beta})}. \quad (2)$$

Then the probability $P(x)$ of every PER x and the probability $P(\lambda)$ of every junction λ can be computed based on the initial estimation.

3.3.2 Maximization and Expectation The likelihood of the data is based on the probability $P(\hat{\pi}_x^\gamma = \pi_x^\gamma | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})$. We define the function $Q(D, P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}))$,

$$\begin{aligned} &Q(D, P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})) \\ &= \sum_{x \in D} \sum_{\pi_x^{\alpha,\beta} \in \Pi_x^{\alpha,\beta}} P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}) log \frac{P(x, \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})}{P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})}. \end{aligned}$$

The EM algorithm performs maximization and expectation iteratively. At each iteration, hill climbing algorithm is applied to estimate $P(\hat{\pi}_x^\gamma = \pi_x^\gamma | \hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta})$ for every PER x such that $Q(D, P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}))$ is maximized. The proof that the maximization of $Q(D, P(\hat{\pi}_x^{\alpha,\beta} = \pi_x^{\alpha,\beta}))$ will lead to the maximization of $l(D)$ is included in the supplemental materials. At the end of each iteration, the probability that PER x is mapped to alignment

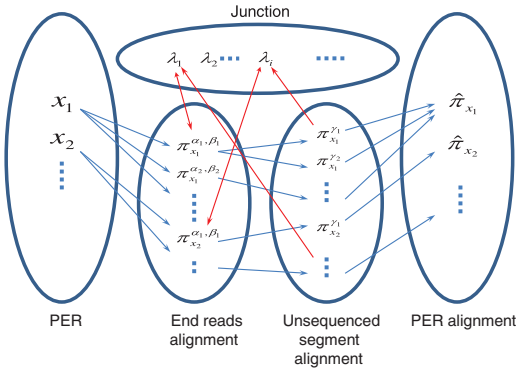


Fig. 4. An illustration of the dependency relationship among the alignments of end reads, the alignments of unsequenced segments and junctions during the inference of the PER alignments. Within this probabilistic model, the probability of a junction is dependent on the PERs that support the junction, and the probability of a read alignment is dependent on the joint probability of the junctions spanned by the read alignment. Taking PERs as input, our method aims at identifying the most probable alignments for every mate-pair x.

$\pi_x^{\alpha,\beta}$ is updated by taking the expectation, as calculated in Equation 2. The proof of correctness for the EM approach is included in the supplemental materials.

4 IMPLEMENTATION DETAILS

Applying EM algorithm to millions of PERs to evaluate all their candidate alignments is computationally intensive. We have developed the following two strategies to speed up the computation.

4.1 Maximal exonic blocks

One of the most time-consuming steps is the search for all possible splicing paths at each PER end read alignment. In the naive method, one may compute these paths one by one for each PER based on the splice graph G. However, such an implementation is not feasible for large RNA-seq data. To improve on it, our method first identifies the clusters of PER alignments sharing the same set of splicing paths. Therefore, instead of computing the paths for one PER alignment at a time, the search can be conducted for the entire cluster.

The clusters were identified by partitioning the genome into maximum blocks in which no junction starts and/or ends. We define them as *m*aximum exonic blocks. To identify these blocks given a set of splice junctions, we start with the whole genome as one block. Each junction will be examined next. For each junction, if it falls into a block, the block will be split into two smaller blocks. For paired read alignment x, suppose its start read x_α is mapped to position $[a_1, b_1]$ and its end read x_β is mapped to position $[a_2, b_2]$. Then x_α can be mapped to a start block $B_\alpha = [B_\alpha^l, B_\alpha^r]$ and x_β can be mapped to an end block $B_\beta = [B_\beta^l, B_\beta^r]$, such that $B_\alpha^l \le b_1 \le B_\alpha^r$ and $B_\beta^l \le a_2 \le B_\beta^r$. After all the junctions are examined, the resulted blocks are all maximum blocks containing no junctions.

We then map all the PER alignments onto these blocks. If two paired read alignments π_{x_1} and π_{x_2} belong to the same start block and the same end block, they cover the same set of junctions and hence have the same set of possible paths. In this case, we group them into one *c*luster of PERs. For every cluster, we only need to compute the possible paths once. Then the particular set of possible distances for every alignment of this cluster can be calculated by adding the particular distance on the start block and the end block to the shared distance from the start block to the end block.

Table 1. Summary of the experimental datasets

	Same chromosome		Cross chromosome	
#PERs	Input	Mapped	Input	Mapped
MCF-7	12.7 M	11.5 M	541 K	79 K
SUM-102	13.6 M	12.5 M	527 K	61 K

4.2 Independent set of PERs

Performing iterative EM on all PERs is both memory and time consuming. Since most of the alternative splicing events occur locally within a gene and are independent among different genes, we adopt a divide and conquer approach by dividing the set of PERs into a number of *minimum independent sets*. Two sets of PERs are called independent if they do not share junctions. A set of PERs is a minimum independent set if it cannot be divided into two subsets of PERs that are independent. The probability of a PER is dependent on the junctions only if they overlap in their genomic span. This procedure helps to speed up the program significantly by confining EM procedures within each independent set, which is much smaller than the whole data.

5 EXPERIMENTAL RESULTS

5.1 Datasets and parameters

We applied our methods on two 2×35 bp paired-end RNA-seq datasets sampling two well-studied breast cancer cell lines, MCF-7 and SUM-102. The RNA-seq data were generated by the Illumina Genome Analyzer II.

Both datasets were first mapped by MapSplice by aligning all 35 bp end reads individually. The error tolerance was set to 5%, allowing up to two mismatches in the alignment for each 35 bp read. For spliced alignment, the minimum anchor size was 6 bp beyond the splice junction.

To understand how PERs might affect the sensitivity and specificity of junction detection, no further filtering was performed on the alignments. Next, PER fragment alignment was computed using the methods proposed in this article. The mate-pair distance distribution was fit to a Gaussian model with a mean of 112 bp and SD of 40 bp.

The software was implemented in C++. The results presented here were run on an Intel(R) Xeon(R) E5540 (2.53 GHz) CPU running Linux. The program is single-threaded,[2] and finished within 5 h on each dataset, using <10 G memory. The software requires alignments of the individual end reads following the standard SAM format. In our case this was produced using MapSplice, but it can also be produced by TopHat or other RNA-Seq aligners producing read alignments in the SAM format. The output is the predicted alignment of the PER fragments also following a simplified SAM format.

5.2 Resolving ambiguous alignments

For each dataset, the number of input PERs and the number of successfully mapped PERs are summarized in Table 1. About 91% of the PERs with both end reads mapped to the same chromosomes have fragment alignments with high probability. In contrast, <15% of the PERs have a highly probable fragment alignment if their end

[2]We are currently working on a multi-threaded implementation of the software.

reads are mapped to different chromosomes. This might reflect the susceptibility of multiple alignment for short reads as a result of repeats or homologous genes across the genome.

Among the 11.5 million mapped PERs in the MCF-7 sample, about 7 million PERs have unique fragment alignments. Most of these PERs map onto exonic regions and, therefore, contain only 49% of the junctions found among the single end reads. The rest of the mapped PERs either have ambiguous end alignments or ambiguous splicing paths. In these cases, expectation maximization has assigned the most likely alignment. Without these alignments, it would be difficult to evaluate the quality of the majority of splice junctions. Restricting PER fragment alignments to unique alignments would also decrease junction coverage. The average support of the splice junctions covered by unambiguous alignments is 14.1 reads, whereas the average support from all PER alignments is 37.7 reads. Therefore, the expectation maximization method improves splice junction discovery as well as providing more accurate quantification of junction coverage, as shown in Section 5.3.

5.3 Splice junction discovery

5.3.1 Sensitivity and specificity for splice junction detection The alignment of the individual 35 bp end reads yields a set J of putative splice junctions. During PER fragment alignment, the probability of each junction in J is evaluated according to the PER fragment alignments that incorporate it, and some putative junctions will be eliminated if there do not exist reliable PER alignments supporting them. We denote the set of junctions remaining following PER fragment alignment as J_{PER}. In Figure 5a and b, we compare the sensitivity and specificity of junctions detected in end reads (J) with those remaining following PER fragment alignment (J_{PER}). In both datasets, ~79% of the junctions in J_{PER} were confirmed by transcripts in the Genbank database, while only 66% of junctions in J could be confirmed in this fashion. On the other hand, 93% of the total confirmed junctions in J were also present in J_{PER}. The small loss of sensitivity might be due to junctions present in one end of a PER whose other end failed to be aligned.

Among the unconfirmed junctions in J_{PER}, in both datasets nearly 50% were found to be either splice junctions connecting known exon boundaries or coordinates close to known exon boundaries. The majority of the unconfirmed junctions were highly supported and had coverage profiles resembling true junctions. In summary, splice junction discovery through PER fragment alignment mostly preserves the sensitivity of the discovery via individual end reads while significantly improving specificity.

5.3.2 Increased junction coverage with PER We next look at how PER fragment alignment may change the coverage of junctions. The coverage of each junction $j \in J$ is the number of alignments of end reads that include j. Each $j \in J_{\mathrm{PER}}$ is covered by the number of PER fragments in which the junction is part of the most probable alignment. Since each PER fragment length is significantly longer than a single end read, we expect the coverage of junctions in J_{PER} to be significantly higher than the same junctions in J.

On both datasets, the average coverage of confirmed junctions is 37.7 using PER alignment, compared to only 11.1 using end read alignment. The scatter plots shown in Figure 5c and d illustrate the PER support versus single end support for all confirmed junctions.

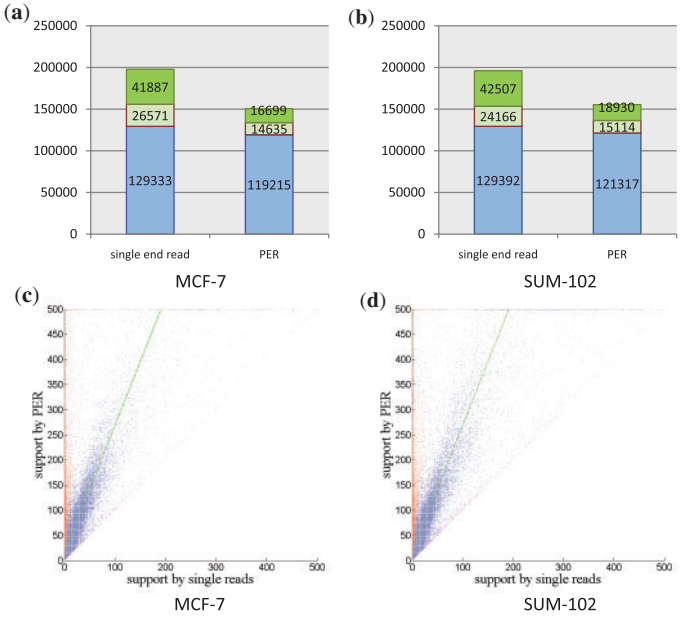
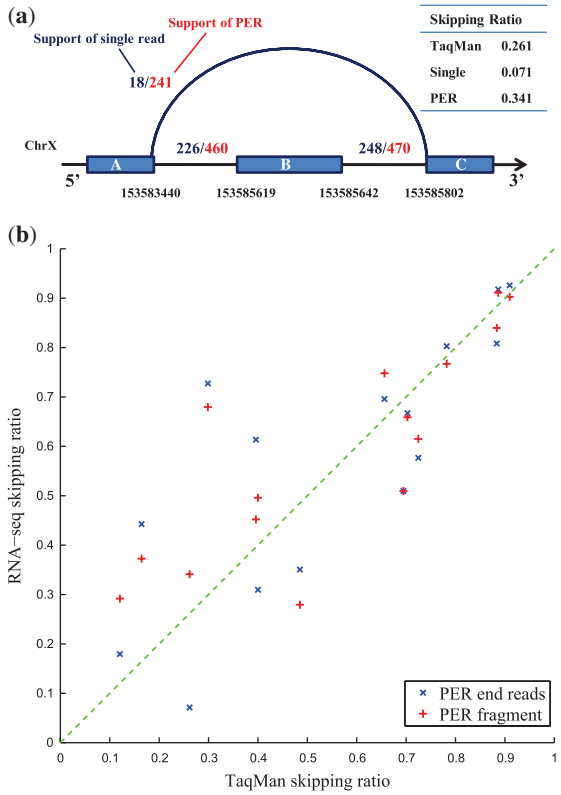


Fig. 5. (**a**) and (**b**) *Comparison of sensitivity and specificity of splice junction discovery*. In each chart, the left bar represents junctions found by (spliced) alignment of PER end reads, and the right bar represents junctions found by alignment of the whole PER fragment. Each bar counts junctions in three categories: the bottom block is the number of junctions confirmed by GenBank; the middle block is the number of junctions whose 5′ and 3′ ends connect known exon boundaries or are close to such boundaries; the top block corresponds to the number of junctions that cannot be confirmed either way. (**c**)and (**d**) *Comparison of junction coverage*. For each confirmed junction, the x-coordinate is the junction coverage among end read alignments, and the y-coordinate is the junction coverage among PER fragment alignments. Points close to y-axis, are junctions primarily supported by PER fragment alignments, while points close to the diagonal, are junctions primarily supported by end read alignments.

Around 25% of junctions are primarily supported by PER fragments, while only around 7% of junctions gain substantial support from single end reads. Furthermore, the majority of the junctions (>67%), corresponding to points, have PER support 3-fold higher than single end reads.

To evaluate the accuracy of the junction coverage in the presence of alternative splicing, we selected eight known skipped-exon alternative splicing events. We used quantitative RT–PCR to measure, in both of our datasets, the exon skipping ratio of the event, i.e. the fraction of transcript isoforms that include the preceding exon and the successor exon, but not the skipped exon. We compared these experimental values with exon skipping ratios calculated using the ratio of splice junction counts determined using individual end read alignments and using PER fragments alignments (Fig. 6a). With a Pearson's correlation of 0.83 across all 16 measurements, the PER fragment alignments achieved high agreement with experimental values, as shown in Figure 6b. The accuracy is higher than the exon skipping ratio derived using counts from single end reads, which has a correlation of 0.78.

In summary, PER fragment alignment yields higher coverage of junctions than obtained from alignment of the end-reads only. The agreement with experimental measurements suggests that PER fragment alignment yields accurate coverage and assigns the correct



comparison	correlation
Taqman - PER end reads	0.778955
Taqman - PER fragment	**0.828681**

Fig. 6. (**a**) An example of an exon skipping event in gene FLNA with junction counts determined from the SUM102 RNA-seq data via end read alignments and PER fragment alignments, respectively. The skipping ratio is computed as count(AC)/(count(AC)+$\frac{1}{2}$(count(AB)+count(BC). (**b**) Correlation of eight exon skipping ratios derived from qRT–PCR in each dataset and those computed using PER end read alignments and PER fragment alignments, respectively.

splicing alternative to the individual PER fragment alignments that have splice graphs with alternative edges.

5.4 Fusion junctions

Finally, we apply the methods of this article to the problem of gene fusion detection. Generally, 35 bp reads are too short to identify long-range fusion junctions with any confidence, since genome-wide spliced alignment of a 35 bp sequence will yield multiple occurrences due to chance as well as repeats and homologous genes.

We obtained a candidate set of fusion junctions from two sets of 75 bp single read RNA-seq datasets from the same cell lines. The 75 bp dataset was aligned genome-wide using MapSplice without filtering (to maximize sensitivity) and candidate fusion junctions were selected by a spliced 75 bp alignment whose prefix and suffix (of length at least 25 bp) were mapped to different genes.

Even by limiting the 75bp alignment to be unique, we obtained 13 513 candidate fusion junctions in the MCF-7 dataset and 11 665 putative junctions on SUM-102 dataset. Taking these fusion

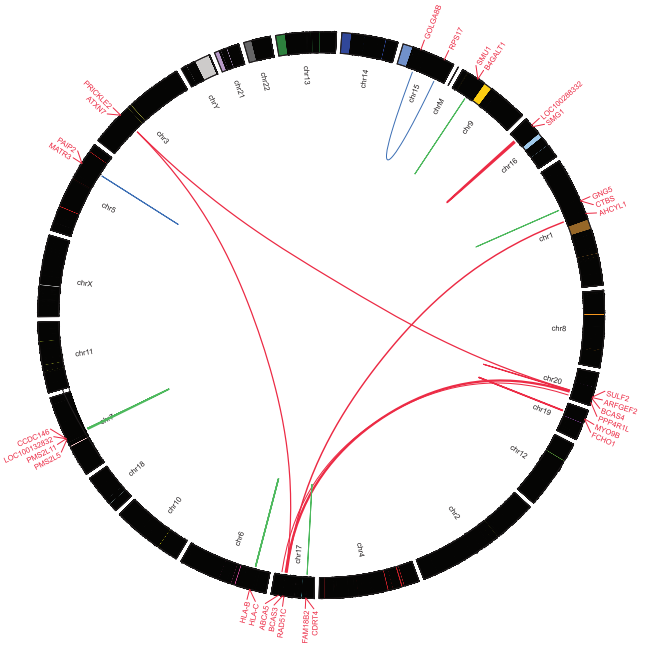


Fig. 7. A set of gene fusion events confirmed by PER data, plotted with Circos (Krzywinski, 2009).

Table 2. A list of rediscovered gene fusions specific to MCF-7 reported by (Maher *et al.*, 2009)

Donor		Acceptor		Similarity	#PERs
BCAS4	chr20	BCAS3	chr17	24.3%	731
ARFGEF2	chr20	SULF2	chr20	27.6%	3
SULF2	chr20	PRICKLE2	chr3	29.7%	4
AHCYL1	chr1	RAD51C	chr17	22.4%	9
ATXN7	chr3	BCAS3	chr17	30.3%	4
LOC100288332	chr16	SMG1	chr16	3.3%	29
PPP4R1L	chr20	ABCA5	chr17	11.4%	3
MYO9B	chr19	FCHO1	chr19	28.7%	15

junctions as putative edges in the splice graph, our PER alignment using 2×35 bp greatly reduced the possible fusion candidates. About 2904 junctions in MCF-7 and 2990 junctions in SUM-102 remained supported. This set of fusion junctions was further filtered by eliminating pairs of genes with high sequence similarity to avoid false positive predictions due to homologous genes. Figure 7 shows a final set of 18 fusion events where the genes connected by the junctions have <35% identity similarity evaluated by the Align program from Emboss. This includes 10 fusion events in MCF-7 and 8 fusion events in SUM102. Eight out of 10 MCF-7 fusion events were previously reported by (Maher *et al.*, 2009), where they were confirmed by experimental qRT–PCR validation. The detailed information of these gene fusion events are listed in Table 2.

6 DISCUSSION

RNA sequencing using the paired-end protocol is a cost-efficient way to sample transcript fragments longer than the sequencing

capability by sequencing only the ends. We propose a probabilistic framework to predict the alignment of each transcript fragment to a reference genome. The alignment chosen is determined by maximizing the likelihood of all PER alignments through an expectation maximization method.

PER transcript fragment alignment offers a number of advantages over the alignment of just the end reads. First, the fragment alignments significantly increase coverage of the transcriptome, providing a more robust measure of transcriptome expression profiles. Second, the splice junctions in the transcript fragments have higher specificity than the junctions in the individual end reads because the PER fragment alignments maximize information from the entire set of end read alignments. Third, the splice graph accurately captures alternative paths between two end reads and the expected mate-pair distance of end reads can effectively disambiguate them, as shown by the high correlation with experimental measurement of alternative splicing events.

The recently published SpliceMap method (Au *et al.*, 2010) also mentions the use of PERs to filter splice junctions.[3] SpliceMap examines the PER support of a junction within some neighborhood. However, lacking a splice graph model of the connection between the end reads, the method may miss true support and include spurious support especially in genes that are alternative spliced or are not highly expressed. In comparison, our likelihood-based method finds the accurate and complete set of PER supports without relying on an arbitrary threshold.

A major impetus for our work is the detection of novel gene fusion events that result from genomic rearrangement in cancer cells. However, identifying long-range fusion junctions is particularly challenging due to the increased frequency of repeats and homologous genes at the genome wide scale. Our PER alignment approach is capable of detecting *trans*-chromosome and *trans*-strand gene fusion events. And the length of the aligned transcript fragments make more likely the detection of such an event with highly significant long anchors on each side of the fusion. We have demonstrated the application of our method using 2×35 bp PER reads together with single 75 bp reads from MCF-7 and SUM-102 breast cancer cell lines. Our result detected 10 events, 8 of which are gene fusion events identified by (Maher *et al.*, 2009), demonstrating high specificity of the proposed method. If longer PERs are used, such as 2×75 bp, no additional single reads would be necessary for the initial fusion detection.

[3]The comparison between SpliceMap and our method is provided in the Supplementary Material.

ACKNOWLEDGMENTS

We wish to thank Matthew S. Hestand and Peter Huggins for their critical comments on the manuscript.

Funding: National Science Foundation (grant number 0850237 to J.L. and J.F.P.); National Institutes of Health (grant number P20RR016481 to J.L.); Alfred P. Sloan Foundation (to D.Y.C.). Funding for open access charge: National Institutes of Health (grant number CA143848).

Conflict of Interest: none declared.

REFERENCES

Au,K.F. *et al.* (2010) Detection of splice junctions from paired-end RNA-seq data by splicemap. *Nucleic Acids Res.*, [Epub ahead of print, doi:10.1093/nar/gkq211].

Berger,M.F. *et al.* (2010) Integrative analysis of the melanoma transcriptome. *Genome Res.*, [Epub ahead of print, doi:10.1101/gr.103697.109].

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.*, **39**, 1–38.

Krzywinski,M. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.

Li,H. *et al.* (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1821–1858.

Maher,C.A. *et al.* (2009). Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA*, **106**, 12353–12358.

Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-seq. *Bioinformatics*, **25**, 1105–1111.

Wang,K. *et al.* (2010) Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acid Res.* [Epub ahead of print, doi: 10.1093/nar/gkq622].

Wu,C.F.J. (1983) On the convergence properties of the EM algorithm. *Ann. Stat.*, **11**, 95–103.