IPSJ Transactions on Bioinformatics Vol.5 27–33 (Apr. 2012)

[DOI: 10.2197/ipsjtbio.5.27]

Original Paper

A Method for Isoform Prediction from RNA-Seq Data by Iterative Mapping

Tomoshige Ohno[1,a)] Shigeto Seno[1] Yoichi Takenaka[1] Hideo Matsuda[1]

Received: December 8, 2011, Accepted: January 20, 2012, Released: April 19, 2012

Abstract: Alternative splicing plays an important role in eukaryotic gene expression by producing diverse proteins from a single gene. Predicting how genes are transcribed is of great biological interest. To this end, massively parallel whole transcriptome sequencing, often referred to as RNA-Seq, is becoming widely used and is revolutionizing the cataloging isoforms using a vast number of short mRNA fragments called reads. Conventional RNA-Seq analysis methods typically align reads onto a reference genome (mapping) in order to capture the form of isoforms that each gene yields and how much of every isoform is expressed from an RNA-Seq dataset. However, a considerable number of reads cannot be mapped uniquely. Those so-called multireads that are mapped onto multiple locations due to short read length and analogous sequences inflate the uncertainty as to how genes are transcribed. This causes inaccurate gene expression estimations and leads to incorrect isoform prediction. To cope with this problem, we propose a method for isoform prediction by iterative mapping. The positions from which multireads originate can be estimated based on the information of expression levels, whereas quantification of isoform-level expression requires accurate mapping. These procedures are mutually dependent, and therefore remapping reads is essential. By iterating this cycle, our method estimates gene expression levels more precisely and hence improves predictions of alternative splicing. Our method simultaneously estimates isoform-level expressions by computing how many reads originate from each candidate isoform using an EM algorithm within a gene. To validate the effectiveness of the proposed method, we compared its performance with conventional methods using an RNA-Seq dataset derived from a human brain. The proposed method had a precision of 66.7% and outperformed conventional methods in terms of the isoform detection rate.

Keywords: RNA-Seq, alternative splicing, isoform, mapping

1. Introduction

There have recently been tremendous strides in transcriptomics. Revealing the alternative splicing by which multiple isoforms are produced from a single gene in eukaryotes is great biological interest since it contributes to the elucidation of specific biological functions. An isoform is defined as an alternatively spliced transcript. A survey has estimated that the average number of exons within a human gene is eight [1], and up to 92–94% of all human genes yield multiple isoforms [2]. Other work reveals that the vast majority of alternative splicing in the human genome results in changes in encoded proteins [3]. Together with the high frequency of alternative exonic events, the diversity in proteins is brought by the abundance of exonic combinations. Meanwhile, splicing errors are relevant to many diseases, including cancers [4], [5]. It has been reported that 15–60% of known disease-causing mutations affect splicing [6]. A comprehensive understanding of alternative splicing is therefore essential for medical and pharmaceutical studies.

Applying massively parallel sequencing technology to transcriptomic analyses has been widely used due to its high throughput and cost-effectiveness [7], [8], [9], [10], [11]. So-called next-generation sequencers provide a large number of short mRNA sequence tags (reads), ranging from several dozens to hundreds of bases in sequence length depending on the platform [12], [13], [14], [15], [16]. Computational predictions of alternative isoforms is highly desired since it can contribute to the advancement of transcriptomics [10].

Transcriptome assembly approaches using RNA-Seq data are categorized three ways: reference-based strategies, *de novo* strategies, and combined strategies [17]. Reference-based strategies generally take a mapping-first approach in which genome-guided transcript reconstruction is performed. Reads are aligned on a reference genome and piled up into transcripts. Cufflinks [18] and Scripture [19] are representative of this class. Another way to reconstruct a transcriptome is *de novo* assembly as represented in Trans-ABySS [20] and Trinity [21]. This approach aims to find overlaps between the reads and assemble them into longer contigs, followed by traversing the de Bruijn graphs in order to reconstruct transcripts. Although many assemblers have been developed, *de novo* short read assembly still remains challenging due to insufficient read length and combinatorial explosion, and such methods may eventually incur wrongly assembled contigs [22], [23]. The *de novo* assembly of higher eukaryotic transcriptomes is considerably more complicated than revealing bacterial, archaeal, and lower eukaryotic transcriptomes, not only because of the larger data set sizes, but also because of the difficulties involved in identifying alter-

[1] Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565–0871, Japan

[a)] tm-ohno@ist.osaka-u.ac.jp

natively spliced variants. For a more comprehensive transcriptome, there is an alternative way that combines the two strategies above. The combined assembly strategy can be carried out by either an align-then-assemble or an assemble-then-align approach. Reference-based assembly and *de novo* assembly need to be brought together in a sophisticated manner so that they compensate for the disadvantages of each other, but the respective approaches incorporate challenges that should be independently resolved. At this moment, no automated combined assemblers have yet been implemented [17]. To take full advantage of information concerning coverage and depth, we focused on a reference-based strategy.

In reference-based transcriptome assembly, as mentioned above, reads are first mapped onto a reference genome to examine where each read originated in order to reconstruct a transcriptome from an RNA-Seq dataset and the reference genome of the sample species. A considerable number of short reads are aligned to a reference genome [24], [25]. Several implementations have been proposed to detect splice junctions by mapping divided reads [26], [27]. Once mapping is completed, transcript graphs are generated based on the spliced reads mapped onto respective junctions, and thus isoforms are predicted. Isoform-level expressions are obtained by estimating the quantity of reads originating from each variant. Therefore, the accuracy of mapping critically affects transcriptomic analyses because inaccurate mapping may lead to errors in the prediction of isoforms. However, the existence of multireads is an obstacle to accurate estimation of expression levels [28]. Multireads are reads that are mapped onto multiple locations due to the short length of the reads, the presence of repeated sequences, individual differences between the reference and the sample genomes, and possibly sequencing errors. Discarding multireads critically affects the subsequent quantification of expression levels, and diminishes the power to detect differential gene expression [29].

Given accurate mapping of multireads, inferring isoform-level expression still remains challenging [30]. In order to estimate the expression levels of isoforms using RNA-Seq data, many methods, including those of Li et al. [31], Nicolae et al. [32], and Richard et al. [33] have been proposed to distribute reads onto each isoform within a gene location using EM algorithms. While these methods are heuristic in nature, Pasaniuc et al. [34] they use a generative model and take into account gene variation between the reference genome sequence and the sequence of the studied sample in order to improve accuracy in estimating expression levels. The methods mentioned above assume that complete lists of isoforms are already known. This assumption suggests that novel isoforms cannot be detected. Compared to these methods, Trapnell et al. [18] proposed Cufflinks, which analyzes transcriptomes by applying Bayesian Network modeling to an RNA-Seq dataset.

In this paper, we propose a method to predict isoforms using RNA-Seq data by iterative mapping. Our method potentially possesses the capability to predict novel isoforms by creating lists of isoforms from the results of mapping and gene locations referred to from a database. An important feature of our work is iterative mapping in which multireads are repeatedly allocated based on estimated expression levels to improve the accuracy in predicting isoforms and estimating isoform-level expression by an EM algorithm.

2. Method

We propose a method to predict isoforms from an RNA-Seq dataset. Additionally, isoform-level expressions can be obtained by this method. This section provides an overview of the proposed method followed by details.

2.1 Overview

The existence of multireads is an obstacle to estimating accurate expression levels. The differences between actual expression levels and estimations must be reduced in order to improve the performance of isoform prediction. Our approach to resolving this problem is iterative mapping. The abundance of uniquely mapped reads at a certain position can be a clue to reducing estimation errors of the proportion. The differences are expected to diminish by remapping multireads based on the abundance. Since the remapping can result in changes in the predicted isoforms, quantification of expressions must be repeated. Once post-remapping expressions are obtained, remapping should be done again based on the modified proportion. By repeating this cycle, isoforms can be precisely predicted. Simultaneously, the estimated expression levels of the isoforms converge to the actual values.

The conceptual flow of the proposed method is depicted in **Fig. 1**. It consists of five steps. The arrows in Fig. 1 indicate the respective operations. Reads are mapped over a genome in Step 1 and splicing junctions are detected in Step 2. Step 3 constructs tentative gene models and generates candidate isoforms on each locus. Expression levels of the candidate isoforms are estimated in Step 4. Multireads are remapped in Step 5 to reduce the differences between actual expression levels and estimated ones. By iterating Steps 3, 4, and 5, final candidates are output as predicted isoforms when this loop terminates.

2.2 Details

The details of the steps are described as follows.

2.2.1 Step 1: Mapping

Map reads onto a reference genome sequence under a certain condition, e.g., up to two mismatches. Here, we use the Bowtie program that enables fast genome-wide mapping [24]. In order to reduce computational time, Bowtie employs the Burrows-Wheeler Transform instead of naive BLAST-like alignment algorithms [25]. In the resultant alignment, reads can be classified into three categories: uniquely mapped reads, multireads, and unmapped reads. Multireads are set aside for later steps.

2.2.2 Step 2: Junction Detection

Detect splicing junctions as well using unmapped reads (**Fig. 2**). Unmapped reads potentially originate from splicing junctions. To specify where splicing junctions are located, several methods, including TopHat [26], MapSplice [27], and HMMSplicer [35] have been proposed. We use the SpliceMap aligner [36]. This software first performs half-read mapping to take advantage of reasonably long reads offered by the newest

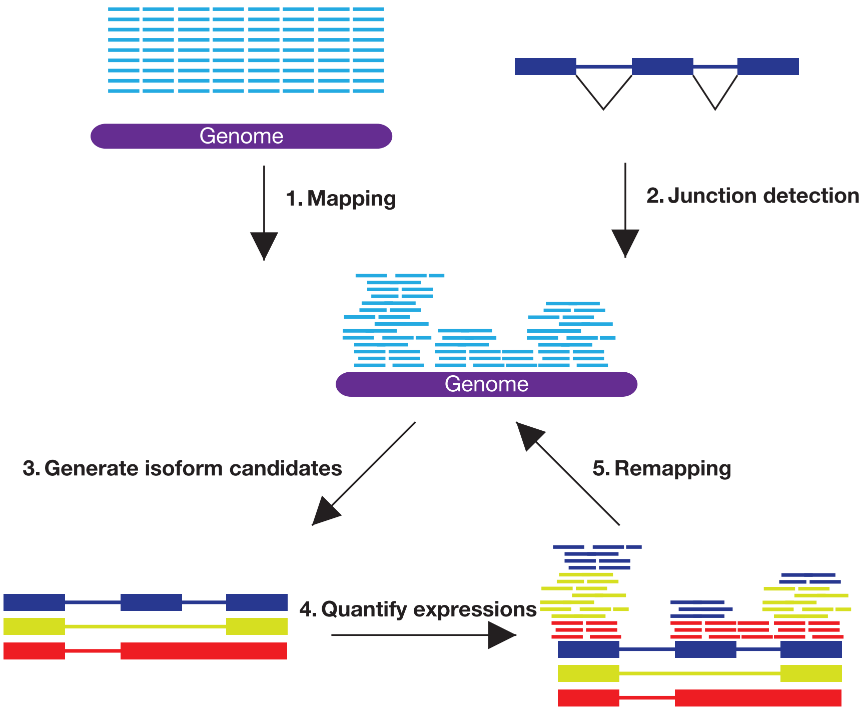

Fig. 1 Flow of the proposed method.

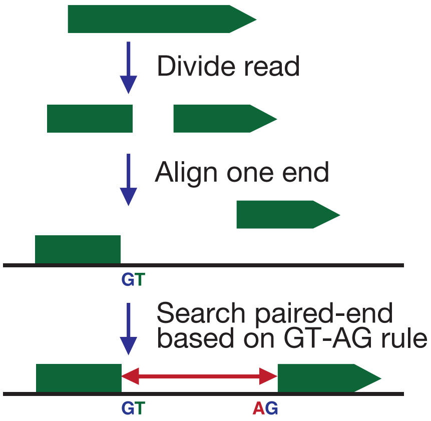

Fig. 2 Detection of splicing junctions.

models of second-generation sequencers. Reads are halved and one end is aligned against the reference genome. The mapped hits of the half-reads are extended base by base to find the splicing points until the sequences GT or AG are detected. These sequences are well-known as a significant splicing signal that appears in 98% of splicing sites of human transcriptomes [37]. This principle, that almost all introns begin with GT and end with AG, is often referred to as the GT-AG rule [38], [39]. The partner half-reads are subsequently aligned within a specific distance. Several mismatches are allowed so that homologous sequences can be detected. The resultant regions are reported as candidate junctions.

2.2.3 Step 3: Generation of Isoform Candidates

Once the result of mapping is obtained, isoform candidates are generated using a reference database for transcriptional sites. For a simpler explanation of notation, the approach is described for a certain gene g, but this model can be applied to all genes so that the whole transcriptome of the target sample can be analyzed. The first task in this step is the construction of a tentative gene model as depicted in **Fig. 3**.

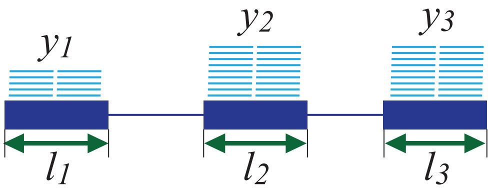

Fig. 3 Estimated gene model.

Using the information of junctions obtained in Step 2, g is separated into a set of (possible) exons. Provided the number of the potential exons within g is N_g, we denote the expression value of each exon as $e_{g,k}$ where k is an integer from 1 to N_g. Suppose y_k is the number of reads mapped onto exon k, and let $e^{(n)}_{g,k}$ denote the normalized expression of exon k in the n-th iteration. Note that y_k may change after remapping in Step 5. At the moment when the initial mapping in Step 1 and the junction-aware alignment in Step 2 are completed, y_k represents the number of uniquely mapped reads. Initialization and normalization is computed as y_k/l_k. This expression level is a temporary estimate and will be updated by means of iterative mapping.

The subsequent operation in this step is to examine whether or not alternative splicing occurs within g. To assess the possibility, we employed Richard et al.'s model [33] for the detection of spliced exons, along with an estimation of expression levels (Step 4). Note that in contrast to Richard et al.'s method, which assumes that a complete list of exons and introns is known, our method constructs a set of candidate isoforms without known variant information. Due to the hypothesis that reads are positioned randomly along every transcript, and suppose g is not alternatively spliced, read counts are considered to obey a multinomial distribution $M((p_k)_{k=1..n}, T)$, i.e.:

$$e_{g,k} = \frac{T}{l_k} \cdot \frac{(\sum l_i)!}{l_1! \cdots l_n!} p_1^{l_1} \cdots p_n^{l_n} \tag{1}$$

where T is the total number of reads across g and p_k is the probability that a read is mapped on exon k, which can be accounted for by the following equation:

$$p_k = \frac{l_k}{\sum_{i=1}^{n} l_i} \tag{2}$$

A chi-square test with freedom of degree $(n - 1)$ is performed under the null hypothesis that alternative splicing does not occur in this gene. A small p-value implies that there are multiple isoforms. If the alternative hypothesis is adopted, Z-scores z_k are calculated on each exon by Eq. (3) in order to specify which exons are involved in alternative splicing.

$$z_k = \frac{R_k - \mathrm{median}(R)}{\mathrm{MAD}(R)} \tag{3}$$

where

$$R_k = \log \frac{y_k}{l_k} = \log e_{g,k} \tag{4}$$

and MAD represents the maximum absolute deviation which is defined as $\max |R_k - \mathrm{median}(R)|$ [33]. For robust estimates, we used the median and MAD instead of mean and standard deviation in order to avoid a bias for genes with few exons. Isoform candidates are all the possible combination of exons whose Z-scores exceed a significance level α. Thus, the most plausible set of isoforms is chosen.

2.2.4 Step 4: Quantification of Expression Levels

This step estimates isoform-level expressions by evaluating the proportions of each isoform. Given a binary matrix I, where $I_{k,j} = 1$ if exon k is expressed in the isoform j, 0 otherwise, the following equation describes the relationship:

$$y_k = \sum_{j \in \mathrm{isoforms}} \frac{p_k}{\sum_i p_i \cdot I_{i,j}} \cdot I_{k,j} \cdot T_j \tag{5}$$

where T_j is the number of reads that originate from isoform j, and y_k is the read count of exon k. In general, T_j cannot be uniquely determined. The expectation maximization (EM) algorithm is deployed to optimize T_j so that it coincides with the number of observed reads mapped on the same exon as shown in Eq. (5). This operation is equivalent to distributing mapped reads onto each isoform within a genetic location.

2.2.5 Step 5: Remapping

Estimated expressions may be incorrect due to numerous multireads. Conceptually, the number of multireads depends on the expression level at a position. To reduce the differences between the estimated and actual values, multireads are remapped so that the number of multireads is in proportion to R_k. This can be done by replacing each column of Y into appropriate values so that the number of multireads is in proportion of the estimated expression level calculated in Step 4. After Step 5, Steps 3 and 4 are repeated. If changes in the result of remapping across the entire genome are less than a threshold, i.e.

$$\sum_{g,k} \left| e_{g,k}^{(n)} - e_{g,k}^{(n-1)} \right| < \epsilon \tag{6}$$

where ϵ is also a given parameter, this routine is considered to converge and is terminated. In case this loop may not converge within a reasonable computational time, we set an upper limit for the number of iterations. This loop is repeated n times, where n is an integer given as a parameter upon execution.

3. Results and Discussions

We conducted an experiment to evaluate the effectiveness of the proposed method.

3.1 Experimental Conditions

Our method was applied to an RNA-Seq dataset, Clontech 636530 [40], which is derived from an adult human brain. The read length is 75 bases and the number of reads is 16,748,521. The UCSC human genome sequence (hg19) was used as the reference genome sequence. We applied the proposed method and validated its accuracy. We used the default parameters in mapping reads with Bowtie except for the *-p* option. In searching junctions, the maximum distance between exons was set to 400,000 bases by default considering the intron length distribution conducted by previous studies, but can be accordingly modified. We also compared our method with Cufflinks [18]. The TopHat aligner was used before predicting isoforms with Cufflinks. In both programs parameters were set to default values.

Ensembl GRCh37.59 human genome annotations were used to validate our method and Cufflinks [41]. Since Cufflinks does not refer to any database except for a reference sequence to determine transcriptional locations, comparisons were limited to the overlaps with Ensembl. Non-genetic regions (non-overlaps with Ensembl) were eliminated from the Cufflinks output. We compared the precision of these two methods. As the number of known annotations is rapidly increasing and real negatives cannot be identified, false negatives were ignored. Although next generation sequencing has the capability to analyze sequences in single-nucleotide resolution, a very small variation in exon boundaries was allowed using the Cuffcompare program, which yields still useful information. In comparing isoforms that consist of multiple exons, candidate isoforms were considered to be true positives only if all the exons matched Ensembl entries.

3.2 Results

Table 1 shows a comparison of genome-wide analyses.

While Cufflinks obtained 40.7% in its precision, our method outperformed it with 66.7%. The precision of exon and intron prediction is shown in **Table 2** and **Table 3**, respectively. Ensembl GRCh37.59 includes 151,225 isoforms, which indicates that the recall of our method and cufflinks were 9.43% and 33.5%, respectively. Of the 124,411 isoforms predicted by Cufflinks, 98,018 were single-exon transcripts, while the output of our method contained 14,485 single exons. Simultaneously, our method surpassed Cufflinks in accuracy, which is calculated as the percentage of exons and introns that match Ensembl over

Table 1 Genome-wide isoform prediction.

	TP+FP	TP	Precision (%)
Proposed method	21,393	14,269	66.7
Cufflinks	124,411	50,635	40.7

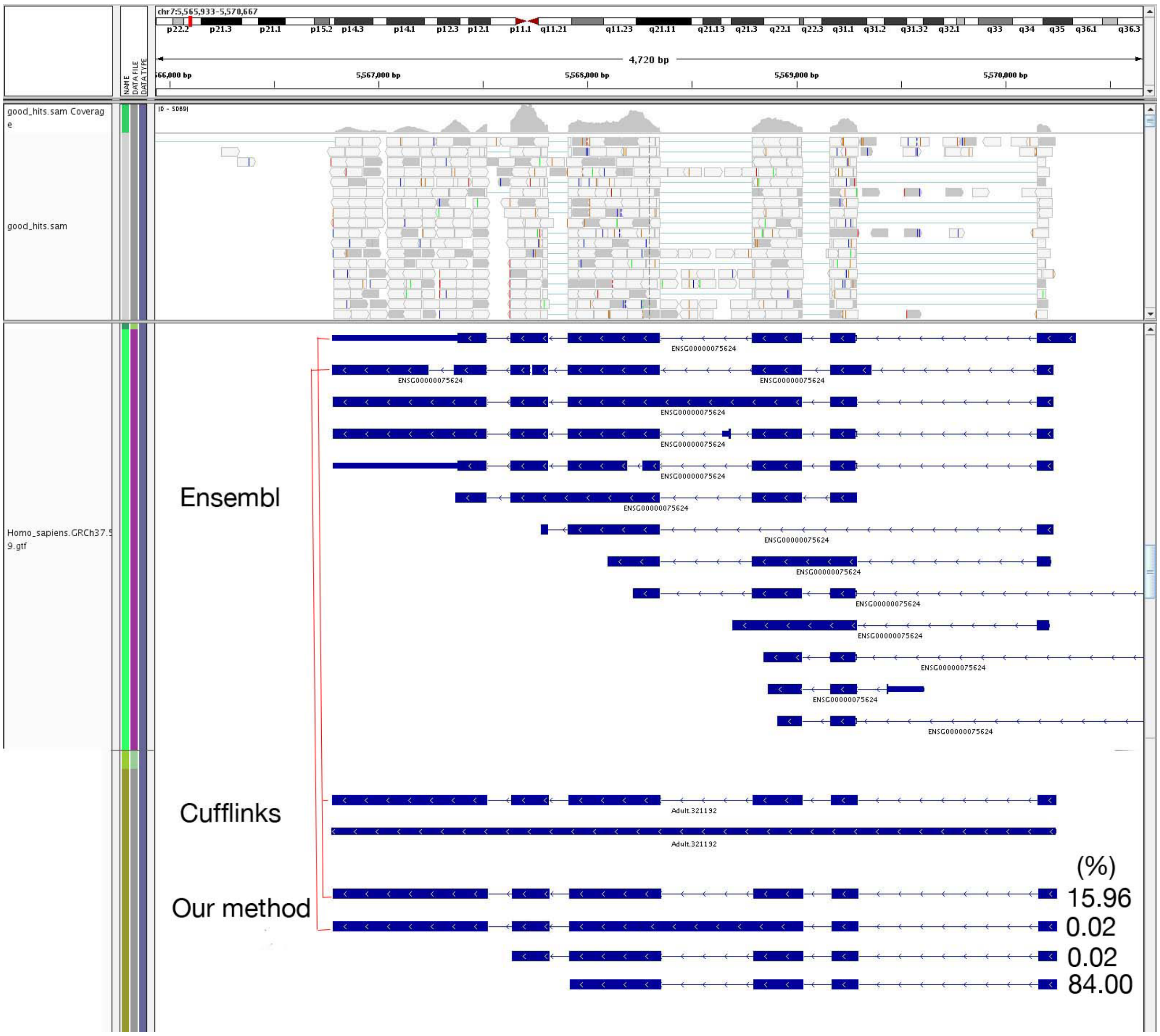

Fig. 4 Results in ACTB.

Table 2 Precision of exon prediction.

(%)	TP+FP	TP	Precision (%)
Proposed method	27,158	25,152	96.4
Cufflinks	320,788	141,133	43.9

Table 3 Precision of intron prediction.

(%)	TP+FP	TP	Precision (%)
Proposed method	25,152	27,158	96.4
Cufflinks	81,087	77,068	95.0

the whole prediction.

Figure 4 shows an example of the result in a specific gene called Actin, Beta (ACTB). There are 13 isoforms registered in Ensembl. Cufflinks detected essentially only one isoform that matches Ensembl. The other isoform predicted by Cufflinks was the entire gene region. In other words, Cufflinks detected essentially one isoform. On the other hand, our method successfully predicted four isoforms, two of which are registered on Ensembl. Another advantage of our method is that it is capable of estimating the relative expression levels of each isoform. Given the total number of reads mapped on this gene, our method can calculate the number of reads that originate from each isoform.

Table 4 Change of accuracy in each loop.

Loop	1st	2nd	3rd
%	38.7	62.3	66.7

Our method was successful in detecting isoforms in a human transcriptome. We also investigated how effective our remapping strategy was. **Table 4** shows the change of performance in each loop. Each figure represents the precision of isoforms. There was a significant improvement in the accuracy from the first loop to the second.

3.3 Discussions

From the comparison of isoform prediction between our method and Cufflinks, our method showed a better precision rate. Table 4 suggests that iterative mapping is effective in enhancing precision since the performance dramatically improved as remapping was carried out. However, there is a significant difference in the number of isoforms predicted. This is potentially caused by the abundant presence of single-exon transcripts in the Cufflinks output. Cufflinks presented many false positive single-exons, while in our method those false positives were eliminated by remapping. As a consequence our method represented remarkably better precisions in exon and intron predictions, suggesting that our method may have the potential to detect novel isoforms.

Unlike most other iterative methods, our method does not use random sampling. For initial values the result of read mapping is used, and the iteration is done deterministically based on the number of reads mapped. Therefore the proposed method is not stochastic. This suggests that our method potentially has the capability of predicting isoforms from a variety of datasets.

4. Conclusions

We propose an analysis method for isoform prediction using RNA-Seq data by the iterative mapping of reads. The proposed method demonstrated improved performance.

Future work includes improvement on the quantification of expression. The prediction of transcriptional locations from the results of mapping is another challenge. This method can enhance the detection of novel isoforms and hence enhance the applications of RNA-Seq analyses.

Acknowledgments We thank the Dragongenomics Center, Takara Bio Inc. for providing the RNA-Seq dataset.

This work was partially supported by Grants-in-Aid for Scientific Research (22680023 and 22310125) from JSPS, and by SPIRE Supercomputational Life Science from MEXT.

Reference

[1] Lander, E.S. et al.: Initial sequencing and analysis of the human genome, *Nature*, Vol.409, No.6822, pp.860–921 (2001).

[2] Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. and Burge, C.B.: Alternative isoform regulation in human tissue transcriptomes, *Nature*, Vol.456, No.7221, pp.470–476 (2008).

[3] Modrek, B. and Lee, C.: A genomic view of alternative splicing, *Nature Genetics*, Vol.30, No.1, pp.13–19 (2002).

[4] Faustino, N.A. and Cooper, T.A.: Pre-mRNA splicing and human disease, *Genes & Development*, Vol.17, No.4, pp.419–437 (2003).

[5] Garcia-Blanco, M.A., Baraniak, A.P. and Lasda, E.L.: Alternative splicing in disease and therapy, *Nature Biotechnology*, Vol.29, No.5, pp.535–546 (2004).

[6] Wang, G.S. and Cooper, T.A.: Splicing in disease: disruption of the splicing code and the decoding machinery, *Nature Reviews Genetics*, Vol.8, No.10, pp.749–761 (2007).

[7] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M.: The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing, *Science*, Vol.320, No.5881, pp.1344–1349 (2008).

[8] Wang, Z., Gerstein, M. and Snyder, M.: RNA-Seq: A revolutionary tool for transcriptomics, *Nature Reviews Genetics*, Vol.10, No.1, pp.57–63 (2009).

[9] Marguerat, S. and Bähler, J.: RNA-seq: from technology to biology, *Nature Reviews Genetics*, Vol.67, No.4, pp.569–579 (2010).

[10] Ozsolak, F. and Milos, P.M.: RNA sequencing: advances, challenges and opportunities, *Nature Reviews Genetics*, Vol.12, No.2, pp.87–98 (2010).

[11] Zhou, X., Ren, L., Meng, Q., Li, Y., Li, Y. and Yu, J.: The next-generation sequencing technology and application, *Protein & Cell*, Vol.1, No.6, pp.520–536 (2010).

[12] Chan, E.Y.: Advances in sequencing technology, *Mutation Research*, Vol.573, No.1-2, pp.13–40 (2005).

[13] Morozova, O. and Marra, M.A.: Applications of next-generation sequencing technologies in functional genomics, *Genomics*, Vol.92, No.5, pp.255–264 (2008).

[14] Shendure, J. and Ji, H.: Next-Generation DNA sequencing, *Nature Biotechnology*, Vol.26, No.10, pp.1135–1145 (2008).

[15] Metzker, M.L.: Sequencing technologies — the next generation, *Nature Reviews Genetics*, Vol.11, No.1, pp.31–46 (2009).

[16] Pareek, C.S., Smoezynski, R. and Tretyn, A.: Sequencing technologies and genome sequencing, *Journal of Applied Genetics*, Vol.52, No.4, pp.413–435 (2011).

[17] Martin, J.A. and Wang, Z.: Next-generation transcriptome assembly, *Nature Reviews Genetics*, Vol.12, No.10, pp.671–682 (2011).

[18] Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L.: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, *Nature Biotechnology*, Vol.28, No.5, pp.511–515 (2010).

[19] Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S. and Regev, A.: Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs, *Nature Biotechnology*, Vol.28, No.5, pp.503–510 (2010).

[20] Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y.S., Newsome, R., Chan, S.K., She, R., Varhol, R., Kamoh, B., Prabhu, A.-L., Tam, A., Zhao, Y., Moore, R.A., Hirst, M., Marra, M.A., Jones, S.J.M., Hoodless, P.A. and Birol, I.: De novo assembly and analysis of RNA-seq data, *Nature Methods*, Vol.7, No.11, pp.909–912 (2010).

[21] Garber, M., Grabherr, M.G., Guttmann, M. and Trapnell, C.: Computational methods for transcriptome annotation and quantification using RNA-seq, *Nature Methods*, Vol.8, No.6, pp.469–477 (2011).

[22] Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C.J. and Deng, H.-W.: Comparative Studies of de novo Assembly Tools for Next-generation Sequencing Technologies, *Bioinformatics*, Vol.27, No.15, pp.2031–2037 (2011).

[23] Narzisi, G. and Mishra, B.: Comparing De Novo Genome Assembly: The Long and Short of It, *PLoS ONE*, Vol.6, No.4, p.e19175 (2011).

[24] Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L.: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biology*, Vol.10, p.R25 (2009).

[25] Li, H. and Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, Vol.25, No.14, pp.1754–1760 (2009).

[26] Trapnell, C., Pachter, L. and Salzberg, S.L.: TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*, Vol.25, No.9, pp.1105–1111 (2009).

[27] 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*, Vol.38, No.18, p.e178 (2010).

[28] Pepke, S., Wold, B. and Mortazavi, A.: Computation for ChIP-seq and RNA-seq studies, *Nature Methods*, Vol.6, No.11, pp.522–532 (2009).

[29] Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B.: Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature Methods*, Vol.5, No.7, pp.621–628 (2008).

[30] Wilhelm, B.T. and Landry, J.-R.: RNA-Seq — quantitative measurement of expression through massively parallel RNA-sequencing, *Methods*, Vol.48, No.3, pp.249–257 (2009).

[31] Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. and Dewey, C.N.: RNA-Seq gene expression estimation with read mapping uncertainty, *Bioinformatics*, Vol.26, No.4, pp.493–500 (2010).

[32] Nicolae, M., Mangul, S., Mandoiu, I.I. and Zelikovsky, A.: Estimation of alternative splicing isoform frequencies from RNA-Seq data, *Algorithms for Molecular Biology*, Vol.6, No.1, p.9 (2011).

[33] Richard, H., Schulz, M.H., Sultan, M., Nürnberger, A., Schrinner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., Haas, S.A. and Yaspo, M.-L.: Prediction of alternative isoforms from expression levels in RNA-Seq experiments, *Nucleic Acids Research*, Vol.38, No.10, p.e112 (2010).

[34] Pasaniuc, B., Zaitlen, N. and Halperin, E.: Accurate Estimation of Expression Levels of Homologous Genes in RNA-seq Experiments, *Journal of Computational Biology*, Vol.18, No.3, pp.459–468 (2011).

[35] Dimon, M.T., Sorber, K. and DeRisi, J.L.: HMMSplicer: A Tool for Efficient and Sensitive Discovery of Known and Novel Splice Junctions in RNA-Seq Data, *PLoS One*, Vol.5, No.11, p.e13875 (2010).

[36] Au, K.F., Jiang, H., Lin, L., Xing, Y. and Wong, W.H.: Detection of splice junctions from paired-end RNA-Seq data by SpliceMap, *Nucleic Acids Research*, Vol.38, No.14, pp.4570–4578 (2010).

[37] Burset, M., Seledtsov, I.A. and Solovyev, V.V.: Analysis of canonical and non-canonical splice sites in mammalian genomes, *Nucleic Acids Research*, Vol.28, No.21, pp.4364–4375 (2000).

[38] Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P.: Ovalbumin gene: evidence for a leader sequencing in mRNA and DNA sequence at the exon-intron boundaries, *Proc. National Academy of Sciences of the United States of America*, Vol.75, No.10, pp.4853–4857 (1978).

[39] Breathnach, R. and Chambon, P.: Organization and Expression of Eucaryotic Split Genes Coding for Proteins, *Annual Review of Biochem-*

istry, Vol.50, pp.394–383 (1981).

[40] Kraev, A., Quednau, B.D., Leach, S., Li, X.-F., Dong, H., Winkfein, R., Perizzolo, M., Cai, X., Yang, R., Philipson, K.D. and Lytton, J.: Molecular Cloning of a Third Member of the Potassium-dependent Sodium-Calcium Exchanger Gene Family, NCKX3, *The Journal of Biological Chemistry*, Vol.276, No.25, pp.23161–23172 (2001).

[41] Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kuesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H.S., Rios, D., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y.A., Trevanion, S., Vandrovcova, J., Vilella, A.J., White, S., Wilder, S.P., Zadissa, A., Zamora, J., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X.M., Herrero, J., Hubbard, T.J.P., Parker, A., Proctor, G., Vogel, J. and Searle, S.M.J.: Ensembl 2011, *Nucleic Acids Research*, Vol.39, No.Suppl 1, pp.D800–D806 (2011).

Tomoshige Ohno received his B.E. and M.E. degrees from Osaka University in 2009 and 2011, respectively. He is currently a Ph.D. student in the Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Japan.

Shigeto Seno is an assistant professor of the Graduate School of Information Science and Technology, Osaka University. He received his B.E., M.E., and Ph.D. degrees from Osaka University in 2001, 2003 and 2006 respectively. He is a member of IPSJ, ISCB, JSBi and MBSJ.

Yoichi Takenaka received his M.E. and Ph.D. in 1997, and 2000 from Osaka University, respectively. He worked for Osaka Universitiy from 2000 to 2002 as an assistant professor, and now he is an associate professor at Graduate School of Information Science and Technology, Osaka University. His research intersts include Bioinformatics, DNA computing, and Neural Networks.

Hideo Matsuda is Professor of the Department of Bioinformatic Engineering, the Graduate School of Information Science and Technology, Osaka University. He received his B.S., M.Eng., and Ph.D. degrees from Kobe University in 1982, 1984 and 1987, respectively. His research interests include computational analysis of genomic sequences and integrated biological databases. He is a member of JSBi, ISCB, IEEE CS and ACM.

(Communicated by *Kenji Sato*)