Nariai et al. BMC Genomics 2014, 15(Suppl 10):S5
http://www.biomedcentral.com/1471-2164/15/S10/S5

BMC Genomics

RESEARCH Open Access

TIGAR2: sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads

Naoki Nariai*, Kaname Kojima, Takahiro Mimori, Yukuto Sato, Yosuke Kawai, Yumi Yamaguchi-Kabata, Masao Nagasaki*

From The 25th International Conference on Genome Informatics (GIW/ISCB-Asia)
Tokyo, Japan. 15-17 December 2014

Abstract

Background: High-throughput RNA sequencing (RNA-Seq) enables quantification and identification of transcripts at single-base resolution. Recently, longer sequence reads become available thanks to the development of new types of sequencing technologies as well as improvements in chemical reagents for the Next Generation Sequencers. Although several computational methods have been proposed for quantifying gene expression levels from RNA-Seq data, they are not sufficiently optimized for longer reads (e.g. > 250 bp).

Results: We propose TIGAR2, a statistical method for quantifying transcript isoforms from fixed and variable length RNA-Seq data. Our method models substitution, deletion, and insertion errors of sequencers based on gapped-alignments of reads to the reference cDNA sequences so that sensitive read-aligners such as Bowtie2 and BWA-MEM are effectively incorporated in our pipeline. Also, a heuristic algorithm is implemented in variational Bayesian inference for faster computation. We apply TIGAR2 to both simulation data and real data of human samples and evaluate performance of transcript quantification with TIGAR2 in comparison to existing methods.

Conclusions: TIGAR2 is a sensitive and accurate tool for quantifying transcript isoform abundances from RNA-Seq data. Our method performs better than existing methods for the fixed-length reads (100 bp, 250 bp, 500 bp, and 1000 bp of both single-end and paired-end) and variable-length reads, especially for reads longer than 250 bp.

Background

Massively parallel sequencing of cDNA libraries constructed from RNA samples (RNA-Seq) has become a popular choice for quantifying gene expression levels of transcript isoforms [1]. Advantages of RNA-Seq over conventional microarray technologies include its larger dynamic range for quantification and capacity of identifying novel isoforms at one nucleotide resolution without the need for designing cDNA probes. A typical RNA-Seq data analysis workflow consists of two components: aligning sequenced reads to the reference cDNA sequences, and quantifying transcript isoform abundances based on the number of mapped reads on the reference sequences. In measuring gene expression levels, FPKM (Fragments Per Kilobase of transcript per Million mapped reads) is calculated under the assumption that a relative expression level of an isoform is proportional to the number of cDNA fragments that originate from it [2].

Since reads are typically 50-300 bp paired-end for Illumina sequencers, in many cases, they can be aligned to more than one isoform and/or locations on the reference sequences. One of challenges for accurate estimation of gene expression is to handle such multi-mapped reads [3]. Several approaches have been proposed to model uncertainty of read mappings in a probabilistic framework, and it has been shown that the statistical inference of read mapping is effective for more accurate estimation of gene

* Correspondence: nariai@megabank.tohoku.ac.jp; nagasaki@megabank.tohoku.ac.jp
Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, 2-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi, 980-8573, Japan

expression levels [4,5]. Although rigorous simulation analyses with various conditions (such as 35 bp vs. 70 bp, and single-end vs. paired-end data) have been performed with several tools in the literature [6], cases for longer reads, such as 250 bp or longer that can be produced from the latest Illumina MiSeq sequencer, have not been extensively studied so far. Moreover, there are few methods suitable for RNA-Seq data produced from new types of sequencers, such as the Ion Torrent PGM sequencer, which generate variable-length reads with relatively higher error rate of substitutions, deletions, and insertions [7,8].

In this paper, we present a statistical method, TIGAR2, which implements new features for improving sensitivity and accuracy of quantification of isoform expression levels from RNA-Seq data by extending the originally developed method [5]. First, for achieving maximum sensitivity for mapping longer reads to reference sequences, TIGAR2 can handle aligned reads from BWA-MEM [9], as well as other widely used alignment tools such as Bowtie2 [10]. Sequencing errors (substitutions, deletions and insertions) within reads that can be inferred from the gapped alignments of reads to reference sequences are modelled under a probabilistic framework in TIGAR2. Second, in order to speed up the variational Bayesian inference in TIGAR2, a new algorithm is implemented so that only reads that can influence isoform abundance parameters in the next iteration are detected and considered in the following update equations.

In order to evaluate quantification performance with TIGAR2, we prepare simulation data that emulates Illumina fixed-length reads (both single-end and paired-end) and Ion Torrent variable-length reads data. For simulating the variable-length reads, a variable read length distribution is empirically estimated from the actual RNA-Seq data by non-parametric regression with Gaussian kernels as basis functions in our analysis. We also apply TIGAR2 to real data of human cell line samples and evaluate consistency of estimated gene expression levels among technical replicates.

Methods

A pipeline of running TIGAR2 consists of two steps: alignment of reads to reference sequences, and estimation of transcript isoform abundances based on the alignment result (Figure 1). Since the first part of the pipeline uses external alignment tools for aligning reads to the reference sequences, it is recommended to run the whole pipeline in the UNIX environment. Details of each step are described in the following sections.

Alignment of reads to reference sequences

Reference cDNA sequences in the FASTA format of model organisms are either available from the RefSeq database [11], or can be generated from the whole genome reference sequence and a gene annotation file (GTF format) with a tool called "gffread", which is included in the Cufflinks package [2]. For cases of non-model organisms, de novo transcriptome assembly might be considered [12], and then the resulting contigs can be used as reference sequences. Given a set of cDNA sequences in FASTA format, the FM-index for the following alignment step is constructed with the corresponding alignment tool. Then, gapped-alignments of reads to the reference sequences are generated with Bowtie2 or BWA-MEM with allowing multiple mappings of reads to the reference cDNA sequences.

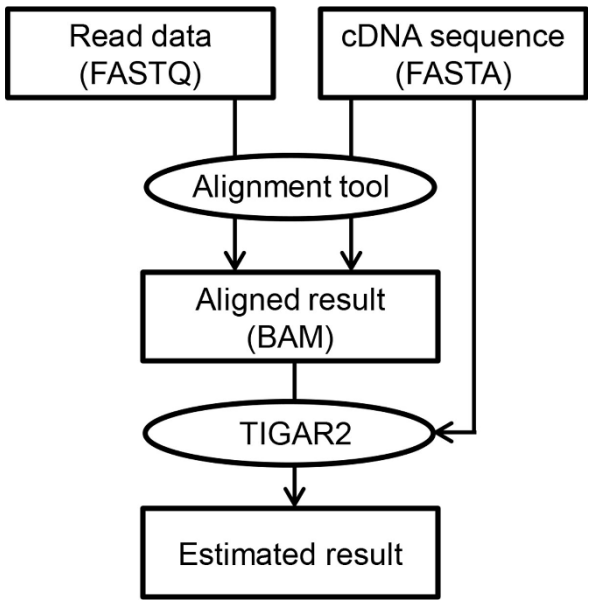

Figure 1 The TIGAR2 pipeline. The schematic diagram shows a typical workflow of running TIGAR2 software. Alignment tools such as Bowtie2 take two input files, read data in FASTQ format and cDNA reference sequences in FASTA format. After reads are aligned to the cDNA sequence, the generated BAM file and the reference FASTA file are used in TIGAR2 for estimating transcript isoform abundances and calculating FPKMs.

Generative model of RNA-Seq data

After the alignment is complete, TIGAR2 takes the resulting SAM/BAM and the FASTA files as input for transcript isoform abundance estimation. We use a generative model for RNA-Seq data as described in Figure 2, which is an extended version of the original model [5]. Here, $\boldsymbol{\theta}$ is a model parameter that represents transcript isoform abundances, and Z_{nt} is an indicator variable and it takes one if read n is generated from transcript isoform n, and zero otherwise. R_n^1 and R_n^2 are the nucleotide sequence of the first and second pair of read n, respectively. Then, the joint probability of the model is decomposed as the product of conditional probabilities as follows:

$$P(\theta, Z_{nt}, R_n^1, R_n^2) = P(\theta)P(Z_{nt}|\theta)P(R_n^1, R_n^2|Z_{nt}).$$

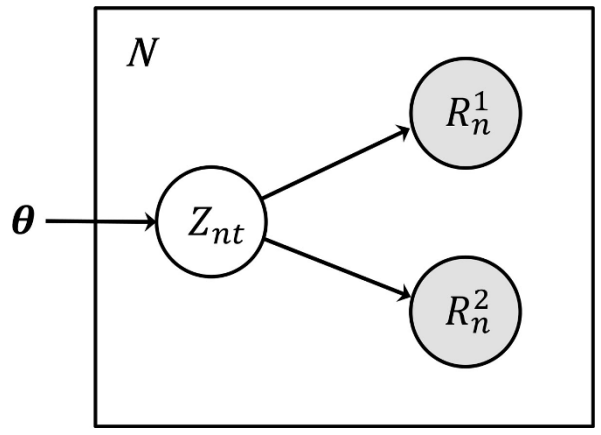

Figure 2 The generative model for RNA-Seq data in TIGAR2. The transcript isoform abundance parameter, indicator variable for transcript isoform choice, nucleotide sequence of the first and second pair read n are represented by $\boldsymbol{\theta}$, Z_{nt}, R_n^1, and R_n^2, respectively.

$P(\boldsymbol{\theta})$ is the prior distribution of the parameter and we assume the Dirichlet distribution:

$$P(\theta) = \frac{1}{C}\prod_{t=0}^{T}\theta_t^{\alpha_t - 1},$$

where $\alpha_t > 0$ is a hyperparameter, C is a constant, T is the number of transcript isoforms, and $\Sigma_{t=0}^{T}\theta_t = 1$. Here, θ_0 represents the noise isoform abundance (reads that are not generated from any known isoform are assigned).

$P(Z_{nt}|\theta)$ is the conditional probability of Z_{nt} given $\boldsymbol{\theta}$ and we further decompose as follows:

$$P(Z_{nt}|\theta) = P(T_n|\theta)P(F_n|T_n)P(S_n|T_n, F_n)P(O_n|T_n)P(A_n^1, A_n^2|T_n, F_n, S_n, O_n),$$

where T_n, F_n, S_n, O_n, A_n^1, and A_n^2 respectively represent the transcript isoform choice, fragment size, read start position, orientation, and alignment state of the first pair and second pair of read n. $P(T_n|\boldsymbol{\theta})$ represents the probability of read n generated from transcript isoform T_n given a parameter vector, and we compute $P(T_n = t|\boldsymbol{\theta}) = \theta_t$. Compared to the original model of TIGAR [5], a fragment size variable is now included in the model. The conditional probability of observing $F_n = f_n$ given $T_n = t_n$ is calculated by truncated and normalized distribution [6,13,14]:

$$P(F_n = f_n|T_n = t) = \frac{d_F(f_n)}{\sum_{x=1}^{l_t} d_F(x)},$$

Where l_t is the length of transcript isoform n, and $d_F(x)$ is the global fragment size distribution. We construct $d_F(x)$ based on the normal distribution with mean μ_F and standard deviation σ_F, which can be either specified according to experimental protocols, or can be estimated from the primary alignments of reads for the case of paired-end data. $P(S_n|T_n, F_n)$ represents the probability of the start position of the first pair of read n given the transcript isoform choice and fragment size, and calculate $P(S_n = s|T_n = t) = 1/f_t$ if mRNAs have poly (A) tails, and $P(S_n = s|T_n = t) = 1/(f_t - L + 1)$ if mRNAs do not have poly(A) tails. $P(O_n|T_n)$ represents the probability of the orientation of read n given the transcript isoform choice. For a strand specific protocol, it can be set as $P(O_n = 0|T_n = t) = 1$ and $P(O_n = 0|T_n = t) = 0$. Otherwise, it can be automatically estimated from the primary alignment of reads from the RNA-Seq data. $P(A_n^1, A_n^2|T_n, F_n, S_n, O_n)$ represents the probability of the alignment state of read n given the transcript isoform choice, fragment size, start position, and orientation of read n. The transition probability of the alignment state is calculated as described previously [5].

Finally, $P(R_n^1, R_n^2|Z_{nt} = 1)$ is the conditional probability of sequence of the first and second pair of read n given $Z_{nt} = 1$. We calculate this probability considering the observed read length as

$$P(R_n^1, R_n^2|Z_{nt} = 1)$$

$$= \prod_{x=1}^{X^1} emit(r^1[x], q^1[x], c^1[x], a^1[x]) \prod_{x=1}^{X^2} emit(r^2[x], q^2[x], c^2[x], a^2[x]),$$

where $emit\ (r^1[x], q^1[x], c^1[x], a^1[x])$ is the emission probability of nucleotide characters of the first pair of read n, $r^1[x]$ is the nucleotide character, $q^1[x]$ is the base call quality score, $c^1[x]$ is the nucleotide character of the corresponding reference sequence, $a^1[x]$ is the alignment state of the first pair of read n at position x. $emit\ (r^2[x], q^2[x], c^2[x], a^2[x])$ is similarly calculated as for the first pair of the read.

Modelling of variable read length distribution

Some sequencers, such as Ion Torrent PGM, produce reads whose lengths are variable. In order to simulate such variable read length, we model the conditional probability of the read length given the fragment size, which is also calculated by the truncated distribution [4]

$$P(L_n = length\ (R_n)|F_n = f_n) = \frac{d_R(length(R_n))}{\sum_{x=1}^{f_n} d_R(x)},$$

where $length\ (R_n)$ is the observed length of read n, and is the global read length distribution. Here, $d_R(x)$ can be constructed based on a linear combination of the smooth functions by fitting it to the data in a non-parametric manner with M equally spaced Gaussian kernels as basis functions. Let

$$g(x) = \sum_{i=1}^{M} a_i m_i(x),$$

where a_i is the coefficient parameter, and $m_i(x)$ is the normal distribution with mean μ_i and standard deviation σ. From the RNA-Seq data, observations of read lengths and their frequency, (x_n, y_n), are constructed, where x_n is the read length, and is y_n the frequency of x_n, and $\Sigma_{n=1}^{N} y_n = 1$. Then, the least squares estimate (LSE) of the parameter vector $a = (a_1, \ldots, a_M)^T$ is obtained by

$$\hat{a} = \arg\min_{a} \sum_{n=1}^{N} \{y_n - g(x_n)\}^2.$$

Define a real value matrix $\mathbf{B}_{ij} = m_j(x_i)$. Then, the ordinary LSE is calculated by

$$\hat{a} = (B^T B)^{-1} B^T y.$$

Then, the global read length distribution $d_R(x)$ can be constructed from $g(x)$ as:

$$d_R(x) = \frac{g(x)}{\Sigma_{x\prime=1}^{\max(L)} g(x\prime)},$$

where $\max(L)$ is the maximum read length of the read data.

An example of the estimated read length distribution from the real sequencing data of a human cell line (HeLa) sequenced by the Ion Torrent PGM sequencer (http://ioncommunity.lifetechnologies.com) is shown in Figure 3.

Estimation of transcript isoform abundances

In our variational Bayesian inference approach, latent variables (true alignments of reads) as well as model parameters (transcript isoform abundances) are estimated as the posterior distribution. We use the Dirichlet distribution for the prior distribution $\theta \sim D(\alpha_0, \ldots, \alpha_0)$

with a single hyperparameter $\alpha_0 > 0$. For $\alpha_0 < 1$, the prior favors solutions in which some of isoforms have zero abundance. Hence, α_0 controls the complexity of model parameters (the number of possible transcript isoforms). A hyperparameter α_0 is selected as a maximizer of the lower bound of the marginal log likelihood of the observed data. Here, we consider α_0 = 0.001, 0.01, 0.1, or 1.0. Each iteration step of the variational approximation updates posterior distribution until a convergence criterion is satisfied. In the VBE step, the expected number of reads that

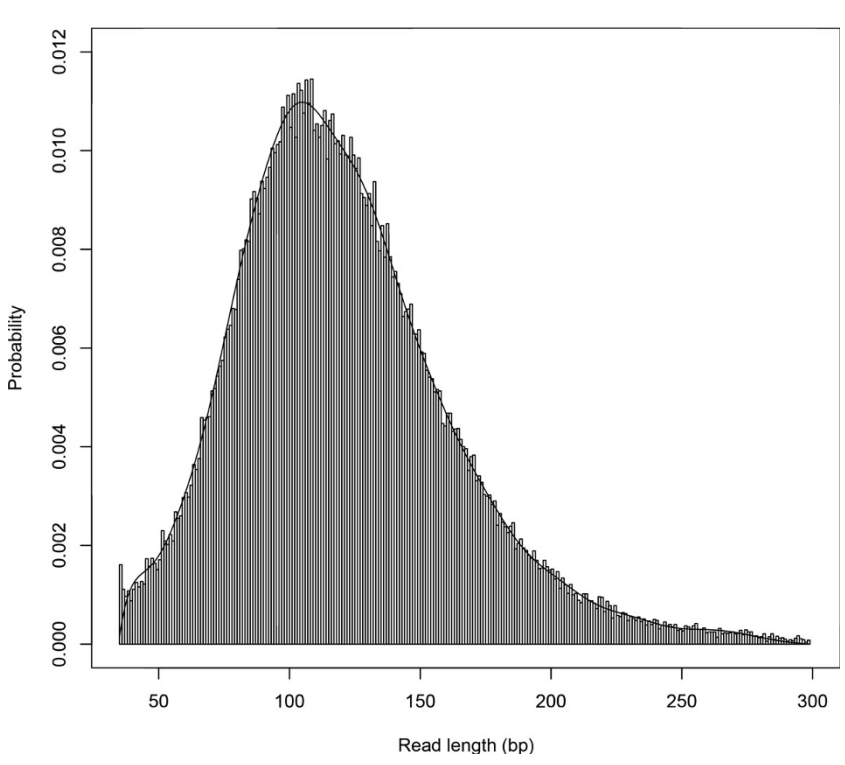

Figure 3 Estimated read length distribution from the Ion Torrent PGM RNA-Seq data. From the histogram of variable read lengths, the smoothly fitted probability distribution is constructed from a linear combination of 20 equally spaced Gaussian distributions(standard deviations are set to 20 bp) that minimizes the least squares error of the estimate.

are mapped to the transcript isoform t is obtained by $\widehat{r_t} = \Sigma_n E_Z[Z_{nt} = 1]$. In the VBM step, the expected abundance of transcript isoform t is obtained by $E_\theta[\hat{\theta}_t] = \widehat{\alpha_t}/(\Sigma_{t'}\widehat{\alpha_{t'}})$, where $\widehat{\alpha_t} = \alpha_0 + \widehat{r_t}$. Details of these update equations and calculation of the lower bound of the marginal likelihood are described in [5]. Recently, it has been shown that the variational inference described here is accurate in estimating the mean of posterior transcript expression, but not the variance [15].

The bottleneck of the computational cost of the inference algorithm is the calculation of $E_z[Z_{nt} = 1]$ for all the possible alignments in the VBE step, which takes $O(M)$ time if the total number of possible alignments is M. This time complexity is upper bounded by $O(NT)$, where N is the number of reads and T is the number of cDNA reference sequences. Suppose some $E_\theta[\widehat{\theta_t}]$ are already converged (unchanged from the previous iteration step) at the current step. We store the information in a Boolean variable *theta_converged* $[t]$, which takes *true* if $E_\theta[\widehat{\theta_t}]$ is converged, and *false* otherwise for each isoform t. Let τ_n be a set of isoforms to which read n is aligned. In the next VBE step, for each read n, $E_z[Z_{nt} = 1]$ will not change if *theata_converged* $[t]$ is *true* for all $t \in \tau_n$. To represent this information, we introduce a Boolean variable *read_movable* $[n]$, which takes *false* if $E_z[Z_{nt} = 1]$ will not change in the next VBE step, and *true* otherwise. The following algorithm computes *read_movable* $[n]$ at the start of each iteration:

1. For each t, set *theta_converged* $[t]$ to *true* if $E_\theta[\widehat{\theta_t}]$ did not change from the previous step, and *false* otherwise.
2. For each n, if *theta_converged* $[t]$ is *true* for all $t \in \tau_n$, then set *read_movable* $[n]$ to *false*, and *true* otherwise.

Then, in the VBE step, $E_z[Z_{nt} = 1]$ is computed where *read_movable* $[n]$ is *true*. The algorithm heuristically eliminates unnecessary calculations of $E_z[Z_{nt} = 1]$ drastically in the later part of iterations, in which most of $E_\theta[\widehat{\theta_t}]$ are already converged and only a fraction of reads should be considered for calculating the update equations.

Results and discussion

Simulation data analysis

We evaluate the performance of quantifying gene expression levels with TIGAR2 compared to existing methods using simulation data. First, 10,000 transcript isoforms in the human RefSeq database [11] are randomly chosen. Second, a set of true gene expression levels is constructed, in which log of isoform abundance is sampled from the standard normal distribution. Then, we generated 20 million, 8 million, 4 million, and 2 million RNA-Seq single-end reads of 100 bp, 250 bp, 500 bp, and 1000 bp, respectively, so that the total throughput of nucleotides remains the same. Similarly, 10 million, 4 million, 2 million, and 1 million paired-end reads of 100 bp, 250 bp, 500 bp, and 1000 bp, respectively, have been generated whose fragment size follows the normal distribution with μ_F = 300, 750, 1250, and 2500, and σ_F = 40, 100, 200, and 400, respectively. In order to simulate sequencing errors, we prepared a set of simulation data with 1% substitution, 1% deletion, and 1% insertion errors. All the simulation data was generated by our in-house software. After aligning reads to the reference cDNA sequences with Bowtie2 (the maximum number of allowed alignments per read is 100), transcript isoform abundances are estimated with TIGAR2. For comparing the performance, TIGAR1 [5], RSEM v1.2.10 [6] and Cufflinks v2.1.1 (with default options except '-u' and '-G' options) [2] are applied to the same simulation data. Although BitSeq [16] is also a relevant method, it is not included in our experiment since performance comparison with TIGAR was already conducted in their analysis [17]. Similarly, variable-length reads are generated according to the estimated read length distribution as shown in Figure 3, and isoform expression levels are estimated with each method. The root mean square errors of the estimated abundances (log of FPKMs) compared to the true gene expression levels are calculated and shown in Figure 4 and 5. For both fixed-length (single-end and paired-end) and variable-length reads, TIGAR2 consistently performed better than others. Especially, when read lengths > 250bp, the prediction accuracies with TIGAR2 over those with RSEM and Cufflinks are markedly better, which can be explained by more sensitive mapping with the latest alignment tools and efficient optimization of multi-mapped reads by the variational Bayesian inference implemented in TIGAR2. Since RSEM uses Bowtie as an aligner in the integrated pipeline, it becomes more difficult to align longer reads to the reference sequences without gapped-alignments of the reads, which potentially loses sensitivity of mappings.

Real data analysis

To evaluate performance with TIGAR2 for real RNA-Seq data analysis, we obtained 4.25 million single-end reads of variable lengths of the human HeLa cell, which is publicly available from the Life Technologies' web site (http://ioncommunity.lifetechnologies.com). The sequencing was performed with the Ion PGM sequencer, which detects the protons released sequentially when one of the four nucleotide bases is introduced in real-time [18]. We divided the RNA-Seq data into two data sets, assuming that they are technical replicates obtained from the same experimental conditions. Gene expression levels were estimated with TIGAR2, RSEM, and Cufflinks and

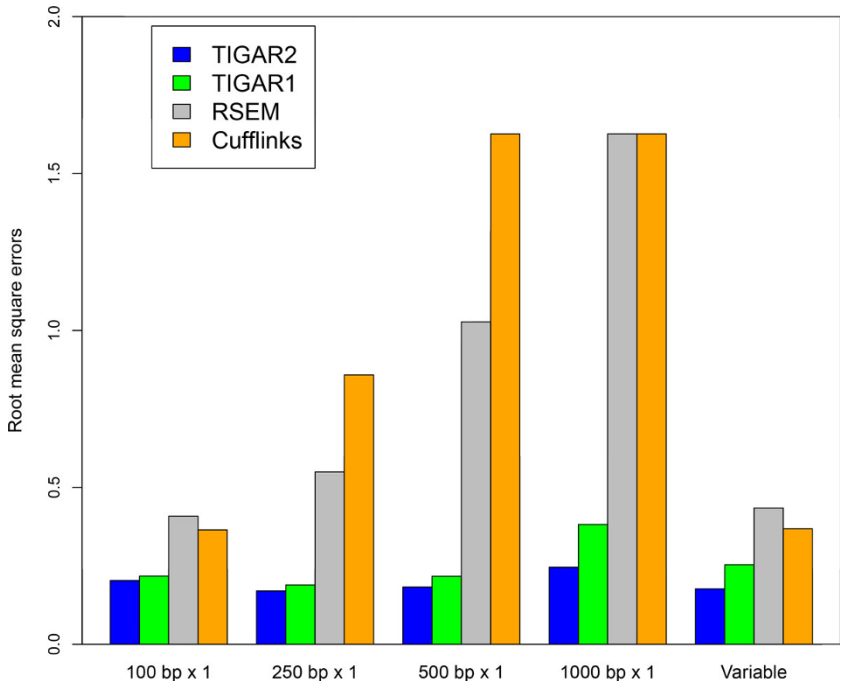

Figure 4 Performance evaluation with TIGAR2, TIGAR1, RSEM, and Cufflinks using single-end and variable lengths simulation data. Root mean square errors of the predicted transcript isoform abundances with each method against the true gene expression levels are shown for 100 bp, 250 bp, 500 bp, and 1,000 bp single-end, and variable-length simulation data. Because RSEM did not produce predictions for 1,000 bp single-end reads, errors were calculated assuming abundances were estimated as zero for all isoforms.

plotted in Figure 6 (the Pearson correlation coefficients of the estimated abundances between the two technical replicates were 0.897, 0.888 and 0.888, respectively). The result shows that the quantification with TIGAR2 was most consistent among the technical replicates, compared to those with RSEM and Cufflinks. TIGAR2 outputs the optimized read alignment on cDNA references in BAM format after inference is done, so that predicted

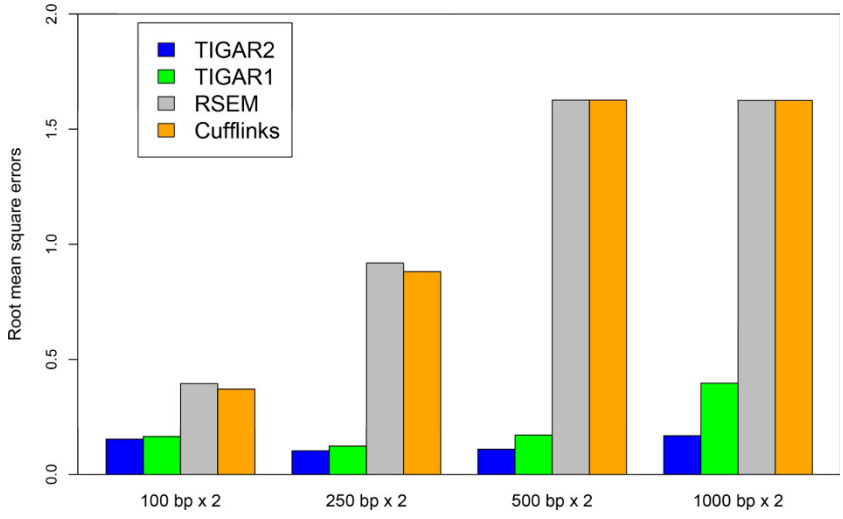

Figure 5 Performance evaluation with TIGAR2, TIGAR1, RSEM, and Cufflinks using paired-end simulation data. Root mean square errors of the predicted transcript isoform abundances with each method against the true gene expression levels are shown for 100 bp, 250 bp, 500 bp, and 1,000 bp paired-end simulation data. Because RSEM and TopHat-Cufflinks did not produce predictions for 500 bp and 1,000 bp paired-end reads, errors were calculated assuming abundances were estimated as zero for all isoforms.

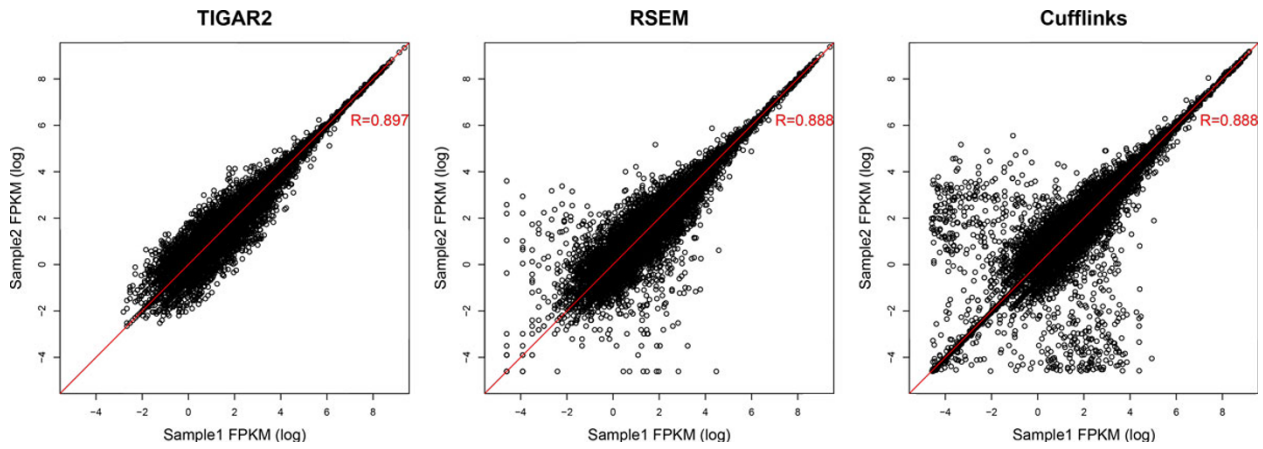

Figure 6 Correlation of gene expression levels estimated from technical replicates. Scatter plots of gene expression levels estimated from technical replicates produced from the Ion Torrent PGM RNA-Seq data. The Pearson correlation coefficients are calculated and shown on each plot. Predictions with TIGAR2 were most consistent among technical replicates.

isoforms can be followed up. The resultant BAM file can be loaded into a genome browser, such as Integrative Genomics Viewer [19]. This function is also a new feature that is not available in the original TIGAR and TopHat-Cufflinks. The bottom track in Figure 7 shows the optimized read alignments estimated with TIGAR2 for NM_001139441, which is an isoform of BAP31 that is known to be expressed in HeLa cells [20]. Compared to the read alignment by Bowtie2 (the top track in Figure 7), not only the amount of reads assigned to the isoform

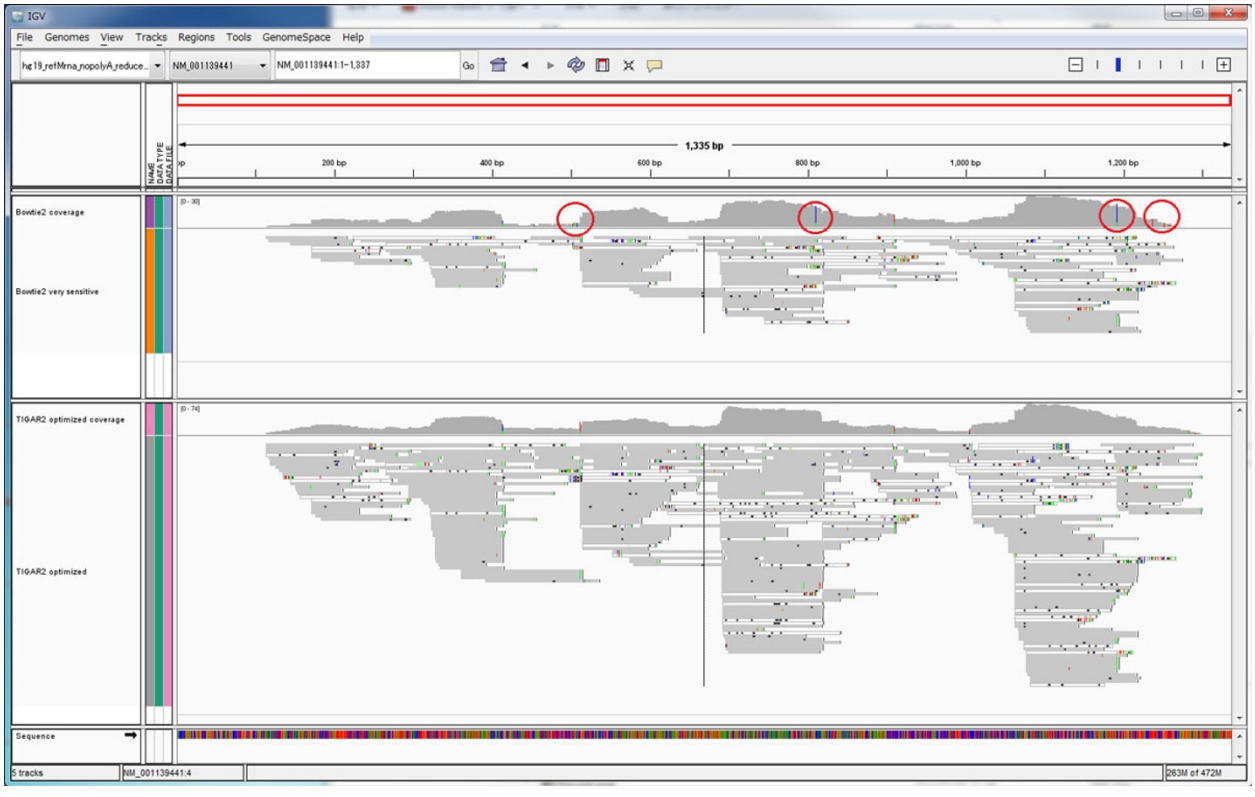

Figure 7 Visualization of the optimized read alignment. Read alignment on NM_001139441, which is an isoform of BAP31, is visualized by IGV. The top track shows the read alignment by Bowtie2, and the bottom track shows the optimized read alignment with TIGAR2. It became easier to identify possible sequencing errors from genetic variants by optimization with TIGAR2 (red circles).

Table 1 Computational resources required for the real data analysis

Tool	Implementation	CPU time in alignment (minutes)	CPU time in estimation (minutes)	Memory (GB)
TIGAR2	Java	18	12	8
TIGAR1	Java	18	53	8
RSEM	C++	10	24	2
TopHat-Cufflinks	C++	1,379	5	4

The CPU time and memory required for the real data analysis with TIGAR2, TIGAR1, RSEM, and TopHat-Cufflinks are summarized.

increased, but also it became easier to identify possible sequencing errors from genomic variants.

Computational resources

CPU time and memory required in the real data analysis are summarized in Table 1. TIGAR2 was the fastest among others, notably more than two times faster than TIGAR1 with practical memory requirement. TopHat-Cufflinks was slower than TIGAR2, TIGAR1 and RSEM, especially in the alignment step.

To see the scalability of TIGAR2 for a large dataset, it is applied to 100 million synthetic reads (100 bp single-end). It required 16 GB memory and 2,621 minutes of CPU time.

All the experiments were performed on an Intel Xeon CPU E5-2670 processor (2.60GHz) with the Red Hat Enterprise Linux Server release 6.2.

Conclusions

We have developed a computational method, named TIGAR2, which is accurate and sensitive in quantifying gene expression levels of transcript isoforms from RNA-Seq data. TIGAR2 outperformed existing methods with simulation data of both single-end and paired-end reads (100 bp, 250 bp, 500 bp and 1000 bp), especially for reads > 250 bp. TIGAR2 will be more effective for accurate detection and quantification of transcript isoforms compared to other existing methods, as new technologies for longer sequencing become available.

Instead of trying to find novel transcript isoforms from RNA-Seq data, reference cDNA sequences of transcript isoforms are assumed to be known in the TIGAR2 pipeline. Although there are a couple of algorithms to predict novel transcript isoforms or fusion genes [2,14,21], TIGAR2 does not provide the novel predictions at the moment. However, once candidates of novel transcript isoforms are predicted by external tools, they can be treated as known and gene expression levels of these novel isoforms can be quantified and assessed with TIGAR2. Another possible extension of TIGAR2 includes modelling of underlying genomic variation for identifying allele-specific gene expression. Because the cost of whole genome-sequencing is dropping sharply, it is becoming feasible to use both genomic information as well as gene expression data. Finally, there should be an optimal balance between the maximum number of allowed alignments per read and the convergence speed. These topics will be investigated as our future works.

Availability of supporting data

The implementation of TIGAR2 and the documentation is available in the GitHub repository, https://github.com/nariai/tigar2.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NN and MN conceived the study, NN, KK, TM, and MN designed the computational experiments, NN performed the analysis, and NN, KK, TM, and MN interpreted the results. YS, YK, and YYK collaborated on data collection and interpretation of the results. NN, KK, TM, YS, YK, YYK and MN wrote the manuscript. All the authors read and approved the final manuscript.

Acknowledgements

This work was supported (in part) by MEXT Tohoku Medical Megabank Project and CREST. All computational resources were provided by the Supercomputing services, Tohoku Medical Megabank Organization, Tohoku University.

Declarations

The publication costs for this article were partly funded by MEXT Tohoku Medical Megabank Project and CREST.
This article has been published as part of *BMC Genomics* Volume 15 Supplement 10, 2014: Proceedings of the 25th International Conference on Genome Informatics (GIW/ISCB-Asia): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/15/S10.

Published: 12 December 2014

References

1. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nature reviews Genetics* 2009, **10(1)**:57-63.
2. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nature protocols* 2012, **7(3)**:562-578.
3. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nature methods* 2008, **5(7)**:621-628.
4. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty.** *Bioinformatics (Oxford, England)* 2010, **26(4)**:493-500.
5. Nariai N, Hirose O, Kojima K, Nagasaki M: **TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference.** *Bioinformatics (Oxford, England)* 2013, **29**:(18):2292-2299.
6. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC bioinformatics* 2011, **12**:323.

7. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: **A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.** *BMC genomics* 2012, **13**:341.
8. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW: **Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data.** *PLoS computational biology* 2013, **9(4)**:e1003031.
9. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics (Oxford, England)* 2009, **25(14)**:1754-1760.
10. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nature methods* 2012, **9(4)**:357-359.
11. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic acids research* 2007, **35(Database)**:D61-65.
12. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nature biotechnology* 2011, **29(7)**:644-652.
13. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, *et al*: **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.** *Nature biotechnology* 2010, **28(5)**:503-510.
14. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nature biotechnology* 2010, **28(5)**:511-515.
15. Papastamoulis P, Hensman J, Glaus P, Rattray M: **Improved variational Bayes inference for transcript expression estimation.** *Statistical applications in genetics and molecular biology* 2014, 1-14.
16. Glaus P, Honkela A, Rattray M: **Identifying differentially expressed transcripts from RNA-seq data with biological variation.** *Bioinformatics (Oxford, England)* 2012, **28(13)**:1721-1728.
17. Hensman J, Glaus P, Honkela A, Rattray M: **Fast Approximate Inference of Transcript Expression Levels from RNA-seq Data.** *ArXiv e-prints* 2013, **1308**:5953.
18. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, *et al*: **An integrated semiconductor device enabling non-optical genome sequencing.** *Nature* 2011, **475(7356)**:348-352.
19. Thorvaldsdottir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** *Briefings in bioinformatics* 2013, **14(2)**:178-192.
20. Wang B, Heath-Engel H, Zhang D, Nguyen N, Thomas DY, Hanrahan JW, Shore GC: **BAP31 interacts with Sec61 translocons and promotes retrotranslocation of CFTRDeltaF508 via the derlin-1 complex.** *Cell* 2008, **133(6)**:1080-1092.
21. Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V: **Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs.** *Bioinformatics (Oxford, England)* 2011, **27(8)**:1068-1075.

doi:10.1186/1471-2164-15-S10-S5
Cite this article as: Nariai *et al.*: TIGAR2: sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads. *BMC Genomics* 2014 **15**(Suppl 10):S5.