BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 5 2014, pages 644–651
doi:10.1093/bioinformatics/btt591

Sequence analysis

Advance Access publication October 15, 2013

ORMAN: Optimal resolution of ambiguous RNA-Seq multimappings in the presence of novel isoforms

Phuong Dao[1,†], Ibrahim Numanagić[1,†], Yen-Yi Lin[1,†], Faraz Hach[1,†], Emre Karakoc[2], Nilgun Donmez[1,3], Colin Collins[3], Evan E. Eichler[2] and S. Cenk Sahinalp[1,4,*]

[1]School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, [2]Department of Genome Sciences, University of Washington, Seattle, WA, USA, [3]Vancouver Prostate Centre & Department of Urologic Sciences, University of British Columbia, Vancouver, BC, Canada and [4]Division of Computer Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: RNA-Seq technology is promising to uncover many novel alternative splicing events, gene fusions and other variations in RNA transcripts. For an accurate detection and quantification of transcripts, it is important to resolve the mapping ambiguity for those RNA-Seq reads that can be mapped to multiple loci: >17% of the reads from mouse RNA-Seq data and 50% of the reads from some plant RNA-Seq data have multiple mapping loci.

In this study, we show how to resolve the mapping ambiguity in the presence of novel transcriptomic events such as exon skipping and novel indels towards accurate downstream analysis. We introduce ORMAN (**O**ptimal **R**esolution of **M**ultimapping **A**mbiguity of R**N**A-Seq Reads), which aims to compute the minimum number of potential transcript products for each gene and to assign each multimapping read to one of these transcripts based on the estimated distribution of the region covering the read. ORMAN achieves this objective through a combinatorial optimization formulation, which is solved through well-known approximation algorithms, integer linear programs and heuristics.

Results: On a simulated RNA-Seq dataset including a random subset of transcripts from the UCSC database, the performance of several state-of-the-art methods for identifying and quantifying novel transcripts, such as Cufflinks, IsoLasso and CLIIQ, is significantly improved through the use of ORMAN. Furthermore, in an experiment using real RNA-Seq reads, we show that ORMAN is able to resolve multimapping to produce coverage values that are similar to the original distribution, even in genes with highly non-uniform coverage.

Availability: ORMAN is available at http://orman.sf.net

Contact: cenk@cs.sfu.ca

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

Received on March 2, 2013; revised on September 30, 2013; accepted on October 10, 2013

1 INTRODUCTION

Massively parallel RNA sequencing (RNA-Seq) technologies are replacing microarrays in determining the structure and dynamics of the transcriptome. Analysis of RNA-Seq data helps to uncover many novel alternative splicing events, gene fusions and other variations in RNA transcripts. Unfortunately, there are many RNA-Seq reads that can be mapped to several loci equally well, and it is of key importance to resolve their mapping ambiguity in order to perform a comprehensive and accurate analysis of the whole RNA-Seq data.

There are several mappers that can align RNA-Seq reads to a reference genome or previously known transcript sequences (Au *et al*., 2010; Trapnell *et al*., 2009; Wang *et al*., 2010; Yorukoglu *et al*., 2012). Owing to the presence of paralogs and homologous regions within a gene, RNA-Seq mappers typically report a fraction of multireads, i.e. reads that map to multiple loci on a reference genome. Based on TopHat mappings on the human reference genome, ~10% of human RNA-Seq reads are multireads. Similarly, ~17% of mouse and 50% of some plant RNA-Seq reads are multireads (Li and Dewey, 2011). The presence of multireads complicates the downstream analysis such as determining alternative splicing patterns, gene fusions and other variations.

The common practice for handling multireads is ignoring them in the downstream analysis. This leads to inaccurate estimation of the abundance of expressed transcripts (Li and Dewey, 2011; Nicolae *et al*., 2011). A simple approach for determining the exact genomic location of a multiread is 'RESCUE' (Mortazavi *et al*., 2008). Here, the initial gene expression values are calculated based on the unique reads that map to them. Each multiread is then assigned to the gene with the fraction equal to the ratio between the gene's initial expression value and the total expression value of all genes that the multiread maps to. A more complex approach based on expected maximization (EM) (Pasaniuc *et al*., 2011) is designed to handle mapping ambiguity of a read to two or more homologous genes—for the purpose of determining the expression value of each of these genes and not to determine or quantify isoforms. Finally, RSEM (Li and Dewey, 2011), IsoEM (Nicolae *et al*., 2011) and iReckon (Mezlini *et al*., 2013) are EM methods based on statistical generative models for sequencing processes to resolve mapping ambiguity.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

Many of the above approaches are designed specially for estimating expression values of *known/annotated isoforms*. Their performance is highly dependent on the completeness of the isoform database in use. Furthermore, they cannot handle alternative splicing events such as novel exon skipping, alternative 5′ donor and 3′ acceptor sites, intron retention and other structural differences such as insertions or deletions. Some recent computational approaches, in particular IsoLasso (Li *et al*., 2011), CLIIQ (Lin *et al*., 2012) and Cufflinks (Trapnell *et al*., 2010), can identify and quantify unknown isoforms and certain types of transcriptomic variations. Unfortunately, neither IsoLasso nor CLIIQ takes into account multireads, and Cufflinks handles multireads through a simple RESCUE-based approach.

In this article, we show how to resolve the multimapping ambiguity in the presence of novel isoforms involving exon skipping, intron retention and small indels towards accurate downstream analysis. To be mathematically precise, we introduce the notion of a partial transcript—a substring of a potential transcript product of a gene, which satisfies certain conditions (a formal definition is provided in the next section).

The objective of our multiread resolution approach, ORMAN (**O**ptimal **R**esolution of **M**ultimapping **A**mbiguity of RNA-Seq Reads), is (i) to compute the minimum number of partial transcripts that cover all the multireads and (ii) to assign each multiread to one of these partial transcripts such that each partial transcript is covered according to the estimated local distribution. We achieve the first objective approximately through a reduction to the standard set cover problem. We achieve the second objective through an integer linear programming formulation, which we handle using available integer linear program (ILP) solvers such as CPLEX, or through greedy heuristics we describe in this article.

We evaluate ORMAN on both simulated and real human RNA-Seq datasets. For the first experiment, we generate paired-end RNA-Seq reads from a random subset of transcripts from the University of California, Santa Cruz (UCSC) database with the expression distribution modelled after a real human dataset. On this simulated data, we show that the performance of state-of-the-art methods for identifying and quantifying transcripts such as CLIIQ, Cufflinks and IsoLasso is typically improved through the use of our multiread resolution approach. Notably, when combined with IsoLasso or CLIIQ, ORMAN gives the most accurate and comprehensive novel isoform detection and quantification pipeline available.

To evaluate ORMAN in a more 'real world' setting, we also design an experiment using real RNA-Seq data from a cancer patient (Lapuk *et al*., 2012). For this experiment, we implant artificial genomic repeats into several genes and compare the performance of ORMAN with that of RESCUE in resolving the multireads mapping to these regions. We show that on this dataset, the multiread assignment by ORMAN approximates the original distributions quite well with a maximum relative error of ≤ 0.3.

2 METHODS

Online databases such as the UCSC browser provide known transcripts from specific gene regions. Let $T = \{T_1, T_2, \ldots, T_p\}$ be the set of known transcripts from a gene region GR. Each transcript T_i is a string that can be partitioned into 'exonic segments' $E(T_i) = \{E_1, E_2, \ldots, E_{|E(T_i)|}\}$. We define the 'gene model' GM implied by the set of transcripts T as an ordered set of alternating substrings of the gene region called canonical exons and canonical introns. Each exonic segment is a maximal substring of a canonical exon, which is either completely present in a transcript or excluded by that transcript. We refer the readers to Figure 1 for an illustration of a gene model derived from known transcripts.

Given a read R mapping to a gene region GR, the partial transcript PT supported by R is the shortest substring of the gene model that completely covers R, which starts and ends with an exonic segment (or a canonical intron in the case of intron retention). If there is a small insertion or deletion (our method limits the size of each indel to 15 nt) between the read and the reference sequence, we introduce a modified partial transcript with the corresponding indel. Figure 2 illustrates several examples of partial transcripts derived from read mappings to a gene model.

2.1 Combinatorial optimization formulation

The set of partial transcripts present in a sample can be derived from the mappings of RNA-Seq reads to a reference genome with the supply of transcript annotation or to a reference transcriptome. Depending on the given mapping to the reference genome or transcriptome, our objective is

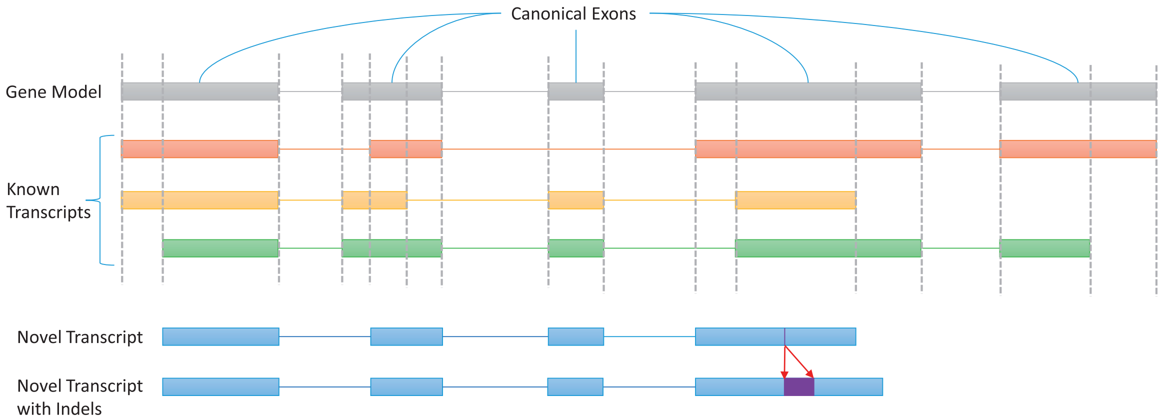

Fig. 1. A gene model, known transcripts (KT) of the gene model, a novel transcript (NT) derived from known transcripts and a novel transcript with indels (NTID). Note that the latter may also be derived from known transcripts

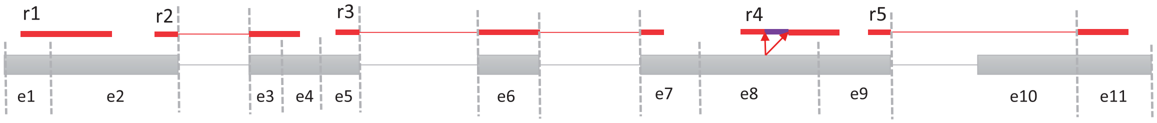

Fig. 2. Example reads mapping to the gene model of Figure 1. The partial transcripts derived from these reads are as follows: r1:{e1,e2}; r2:{e2,e3,e4}; r3:{e5,e6,e7}; r4:{e8^{ins},e9} and r5:{e9,e11}. Above, e8^{ins} denotes exon 8 with the implied insertion

to assign each multiread to a single locus on the genome or a transcript. We also need to determine the partial transcript that the multiread should map to. This is done in two phases. In the first phase, we are interested in the minimum number of partial transcripts that could cover all the (multi)reads. In the second phase, we try to distribute the (multi)reads to the set of partial transcripts from the first phase such that the distribution of mappings for each partial transcript follows the most likely distribution.

2.1.1 First phase Let $C = \{PT_1, PT_2, .., PT_p\}$ be the collection of all the partial transcripts derived from mapping results. We also denote by PT_i $(1 \leq i \leq p)$ the set of all reads that support the same partial transcript PT_i. In addition, each PT_i is assigned a positive weight that is proportional to the number of splicing events, i.e. exon skipping and intron retention events with respect to the known transcript it is associated with. For the partial transcripts without any variations with respect to their associated known transcripts, the weight is 1. Each variation adds a user-defined fixed value to this weight. Our default weight contribution of exon skipping is 100 and of indel is 10 000. Indel events have high weight due to their significantly low relative frequency (Karakoc *et al.*, 2012). Note that, in the case of paired-end reads, each end of a fragment may be assigned to a different partial transcript. In that case, we assign such a pair to the partial transcript formed by taking the union of the exons from the partial transcripts on both ends. The weight of this new partial transcript PT_i is assigned as the sum of the weights of the two partial transcripts it is composed of. We aim to determine the minimum-weighted set of partial transcripts that can cover all the reads. This problem can be defined as an instance of the minimum-weighted set cover problem, where sets are represented by the partial transcripts, and reads represent set elements. Because the minimum-weighted set cover problem is NP-hard, we use the standard greedy algorithm, which provides a logarithmic factor approximation guarantee (Chvatal, 1979) to solve this problem and obtain the set of partial transcripts used for the smoothing step.

2.1.2 Second phase First, we give the formulation of the problem in this phase in terms of an ILP below. We then show the computational complexity of the problem. Finally, we show how to solve the problem in practice.

Let $C' = \{PT_1, PT_2, .., PT_{p'}\}$ be a set of partial transcripts returned from the first phase. For the partial transcript $PT_j \in C'$, we aim to distribute multireads across the partial transcript such that the coverage function of the reads in each partial transcript resembles the most likely distribution. In the case of paired-end reads, we use both ends for the coverage determination. For a read R, let $SPT(R)$ be the set of partial transcripts that R could map to.

Now let len_R be the read length and $len(PT_j)$ be the length of the partial transcript PT_j. Let R_{ij} $(1 \leq i \leq |R|$ and $1 \leq j \leq |SPT(R_i)|)$ be indicator variables, where $R_{ij} = 1$ means that we assign R_i to the partial transcript PT_j; otherwise $R_{ij} = 0$. We enforce that R_i can only be assigned to one partial transcript:

$$\sum_{\{j|PT_j \in SPT(R_i)\}} R_{ij} = 1 \qquad (1)$$

Let NR_{jk} $(1 \leq j \leq p'$ and $1 \leq k \leq len(PT_j))$ be the number of reads that cover position k in PT_j. Let $Multi(PT_j, k)$ be the set of the multireads that cover the position k in PT_j. In similar manner, we define $Unique(PT_j, k)$ to be the number of reads that are uniquely mapped and cover the position k in PT_j. NR_{jk} could be written as the summation of the number of uniquely mapped reads and multireads that cover the location k:

$$NR_{jk} = Unique(PT_j, k) + \sum_{\{i|R_i \in Multi(PT_j, k)\}} R_{ij} \qquad (2)$$

Let AV_{jk} be the desired number of reads covering the position k in the partial transcript PT_j. Because we do not know the original distribution of the reads, we approximate AV_{jk} as follows. First, we find the multimapping region M_k of PT_j, which encompasses position k. Next, we calculate the average coverage in the left and right neighbourhoods of M_k (the size of each neighbourhood is set to $h \times len_R$ base pairs, where h is a user-defined parameter). We use the calculated average values as defining points for a line l, which approximates the desired function AV. Then, AV_{jk} is calculated as a value on the line l at the position k. The rationale of this approach lies in the observation that coverage level of a small region is often similar to the level of the immediately neighbouring regions, even when the coverage varies significantly along the entire gene (see Fig. 3).

Let $d_j \geq 0$ denote the maximum difference between the desired number of reads AV_{jk} per position of partial transcript PT_j and the observed number of reads NR_{jk} at any position k. We enforce the following constraints:

$$-d_j \leq AV_{jk} - NR_{jk} \leq d_j \qquad (3)$$

Our objective is to minimize the total difference:

$$\sum_{1 \leq j \leq p'} d_j \qquad (4)$$

The problem of smoothing of the distribution of reads along partial transcripts, named SMOOTH, is provably hard. In addition, it is unlikely to have a constant factor approximation algorithm for the SMOOTH problem. The proofs are described in Supplementary Materials.

2.1.3 Practical implementation The ILP formulation of ORMAN is solved by IBM ILOG CPLEX. In practice, the running time of the proposed ILP depends on the number of integer and non-integer variables and the number of constraints. The number of integer variables of the provided ILP is proportional to the number of mappings of multireads, which can be in the order of millions. Here we propose a strategy to decompose the original problem into smaller subproblems such that the solution of each smaller one is independent from each other. We create a graph $G_{PT} = (V_{PT}, E_{PT})$ among the partial transcripts returned from the first phase, i.e. $V_{PT} = C'$. There is an edge between two partial transcripts if there is a multiread r mapping to both of them. It is easy to see that the solution of the ILP corresponding to each connected component in G_{PT} is independent from the solutions of other components. Thus, we can obtain the solution for each component separately using CPLEX. There may still exist some components that could not be solved using CPLEX. In these cases, we propose a heuristic strategy as described in the Supplementary Materials.

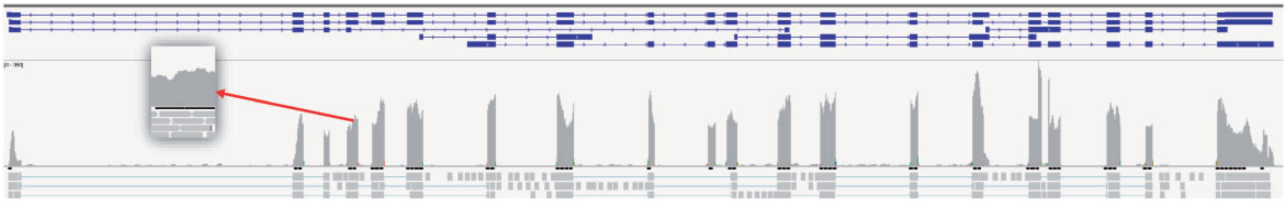

Fig. 3. The read distribution of gene USP5 taken from a real RNA-Seq dataset (see Section 3.2 for details). Although the overall sequence coverage varies significantly along the gene, a small region often coincides well with its neighbourhood

3 EXPERIMENTAL RESULTS

We evaluate the performance of ORMAN on both simulated and real datasets. On both types of data, we show that ORMAN resolves mapping ambiguity of multireads accurately and improves the performance of the leading transcript identification and quantification tools.

3.1 Transcript identification and expression quantification in simulated data

First, we focus on quantifying how much ORMAN improves downstream analysis tools. We compare the performance of the leading transcript identification and quantification tools by (i) first running each tool without any pre- or postprocessing (ORIGINAL), and (ii) then running each tool after preprocessing the mappings by ORMAN.

Because there are no real world benchmark datasets that provide comprehensive and accurate information on all transcripts and their abundance levels validated by wetlab techniques, we use simulation data for this evaluation. [Even though the MAQC project (Shi *et al*., 2006) used RNA-Seq technologies to quantify the expression of a limited number of genes, a significant number of these genes have a single isoform and have unique sequence composition (Li and Dewey, 2011)].

3.1.1 Simulation data We generated RNA-Seq reads of human transcripts with expression distribution similar to one derived from a real dataset from the GEO database (accession number GSM759513). This dataset comprises paired-end 50-bp RNA-Seq reads of a prostate tissue from Illumina Human BodyMap 2.0 project (Shen *et al*., 2012). The reference transcriptome has 76 969 transcripts based on the UCSC database. We used TopHat version 2.0.7—with the number of mismatches at most 2—to obtain the mappings of the RNA-Seq reads to the reference sequence (version hg19). We ran IsoEM to quantify the expression profile of the UCSC reference transcriptome and determined that 39 388 of them are highly expressed.

For the simulations, we assigned one random transcript out of all 76 969 transcripts to each one of the expressed transcripts of the prostate dataset. These randomly assigned transcripts represented the expression of 17 956 genes. We then set the expression value of each random transcript to that of the prostate dataset transcript it is associated with. We finally selected 10% of this randomly selected set of the simulated transcripts for the production of novel transcripts; for each such transcript, we randomly skip an exon.

To ensure this transcript is novel, we check whether it is highly similar to other known transcripts. We consider a novel transcript to be highly similar to a known transcript if they have the same number of exons and their percentage sequence similarity is >90%. The novel transcript is then assigned the same abundance level as the original transcript.

We generated 80 million paired-end RNA-Seq reads of 75-bp length from the chosen transcripts. The fragment length is determined based on the normal distribution with a mean of 250 bp and a standard deviation of 25 bp. Each transcript received a number of reads proportional to its predetermined expression level, and each read was picked uniformly at random over all possible starting positions of the transcript. We then randomly introduced sequencing errors in the generated reads according to sequencing error model described in Dohm *et al*. (2008). This model places the majority of mismatch errors towards the 3′-end of the reads. The error percentage per base was set to be 1%. We used TopHat with the above settings to map the generated reads to the reference genome. Approximately 4% of the generated reads had multiple mapping loci.

3.1.2 Performance evaluation Our performance evaluation is based on three tools: Cufflinks (version 2.0.2), IsoLasso (version 2.6.0) and CLIIQ (version 0.1.0.2). Cufflinks uses a modified rescue strategy to resolve multireads, whereas the latter two are not capable of resolving multimappings. We run CLIIQ in both its standard mode, where it selects the minimum possible number of isoforms, which minimizes quantification errors, and preference mode, where it prefers known isoforms when there are multiple candidate solutions (abbreviated as CLIIQ_pref below). To measure the relative performance of these tools, we provided the complete UCSC gene annotations and disallowed any novel splice sites while allowing novel exon skipping and intron retention events.

The expression values of transcripts are measured in fragments per kilobase per million mapped reads. For each transcript, we define the 'relative quantification error' produced by a given tool as follows. (i) If the known expression value of the transcript is e and the expression value of the transcript reported by the tool is $\hat{e}$, then the relative quantification error is $|e-\hat{e}|/e$. (ii) If the tool reports a transcript that is not among the simulated expressed transcripts, the relative quantification error is $+\infty$. (iii) If the tool misses a known expressed transcript, the relative quantification error is 1. Following (Li and Dewey, 2011; Nicolae *et al*., 2011), we first investigate the proportion of transcripts whose relative quantification error is above a threshold.

For each tool, we also compare how ORMAN affects its performance on detecting novel isoforms. The novel isoforms in our simulation generate reads that are incompatible to any known gene annotations. For existing mapping ambiguity resolving tools that require the full list of known transcripts, these reads might be discarded; hence, novel isoforms with multireads may not be detected. On the other hand, ORMAN allows such reads to be used in the solution. In our experiments, all three tools detect more novel isoforms based on ORMAN mappings as can be seen from Table 1.

In Figure 4, we see that ORMAN improves the performance of IsoLasso and CLIIQ significantly in both modes, which, in comparison with Cufflinks, return fewer incorrectly quantified isoforms for smaller error thresholds. Overall, Figure 4 demonstrates that the combination of ORMAN and CLIIQ_pref provides the best results.

We also report the performance of tools on genes that produce a high proportion of multimapping reads separately. Here, we focus on 3784 genes (expressing 7275 transcripts) to which TopHat mapped reads have the top 20% highest mapping multiplicity (see later).

Figure 5 shows the proportion of transcripts whose relative quantification error is above a threshold on this subset. As before, ORMAN improves the performance of IsoLasso and CLIIQ significantly in both modes, which are better compared with that of Cufflinks.

Next, we consider the performance of each tool in novel isoform detection for those genes that produce multimapping reads. First, we sort all expressed genes according to their mapping multiplicity (i.e. the proportion of the reads that can be mapped to such a gene, which can also be mapped to other genes). Then for genes ranked in the top 10, 20, 30, 40 and 50%, we examine how each tool performs in detecting novel isoforms. Figure 5 demonstrates that, in the case of novel isoforms, all tools benefit from ORMAN mappings. In addition, for those genes whose multiplicity is in top 10% in the sample, ORMAN performs particularly well.

Table 1. Number of novel isoforms correctly identified by each tool with and without ORMAN

	Cufflinks2	IsoLasso	CLIIQ	CLIIQ_pref
ORIGINAL	1043	1292	1513	1325
ORMAN	1055	1308	1533	1334

3.2 Multimapping resolution in real RNA-Seq data

It has been known that real RNA-Seq experiments often suffer from various biases resulting in a rather non-uniform coverage across a gene model (Roberts *et al*., 2011; Wu *et al*., 2011). Unfortunately, modelling of such complex biases in simulations would be cumbersome. To overcome this problem, we design a controlled experiment with real RNA-Seq reads. For this experiment, we use a previously published RNA-Seq dataset with 51-bp Illumina paired-end reads sampled from a human prostate cancer patient (Lapuk *et al*., 2012). On this dataset, we introduce artificial repeats in 10 genes based on sequences of other genes. By modifying the sequences of the original reads mapping to the artificial repeats, yet keeping everything else intact, we essentially create a multimapping dataset for which the true coverage distribution is known. We then evaluate ORMAN's performance in resolving these multireads.

The following section explains the experimental setup in detail. In the next section, we elaborate on the experiment results.

3.2.1 Experimental setup First, we map the reads using TopHat (version 1.3.2) to the reference sequence (hg19) and Ensembl annotations (GRCh37.62). Next, we randomly select 10 'decoy' genes according to the following rules:

(1) The gene is annotated to have a single transcript based on the Ensembl annotations.

(2) The total gene length (i.e. the sum of all canonical exons) is at least 2000 bp.

(3) The gene is sufficiently expressed in the sample, having an average coverage >100.

(4) The gene is uniquely mappable (i.e. there are no multireads mapping to the gene model).

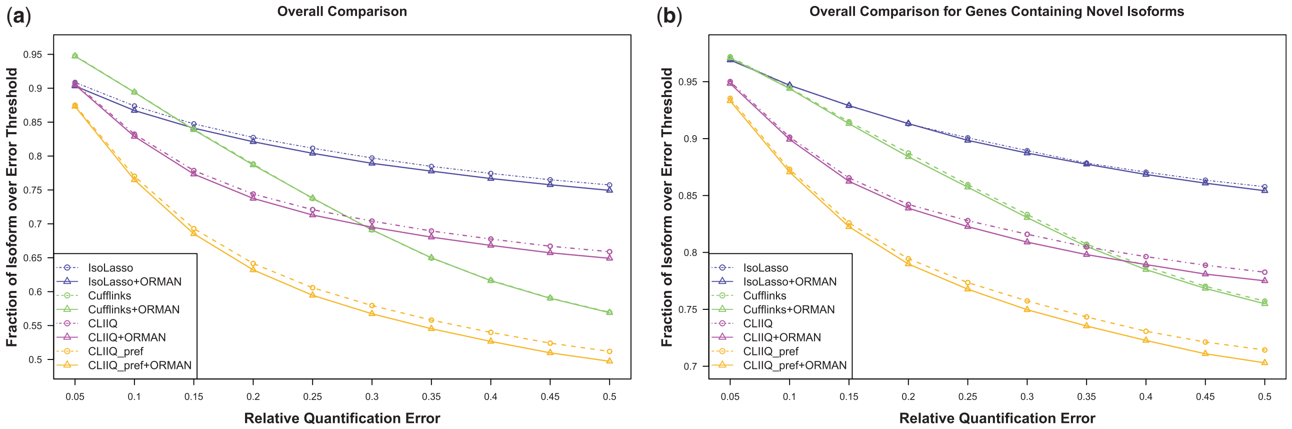

Fig. 4. Comparative performance of each tool and its enhanced version with ORMAN measured as the proportion of transcripts whose relative quantification error is above a threshold, as a function of the threshold. We show results of three tools (ORIGINAL) as well as their ORMAN enhanced versions of (**a**) all 17 956 expressed genes (left) and (**b**) 3148 genes containing novel transcripts (right)

Similarly, we randomly select 10 ‘replacement’ genes according to the rules 2, 3 and 4 above. Within each replacement gene, we select a 400-bp region to serve as an artificial repeat. This 400-bp sequence is then used to replace the sequence of a region of the same length in the decoy gene. In other words, in each decoy gene, we create an artificial repeat for which the sequence is taken from a randomly chosen replacement gene. The selected genes and the repeat regions are given in Table 2.

In the next step, we identify the reads mapping to the coordinates coinciding with the artificial repeat region in each decoy gene. The sequences of these reads are changed according to the new sequence of the decoy gene. All other reads are kept the same. The entire set of reads is then mapped to the new genome reference and the original Ensembl annotations using TopHat with the same parameters.

3.2.2 Evaluation In this experiment, we compare ORMAN with the modified version of RESCUE as used in Cufflinks (Mortazavi *et al.*, 2008; Trapnell *et al.*, 2010). This modified version calculates the initial gene/transcript abundances first by equally distributing the multireads to each gene they map to. In the second phase, each multiread is distributed in proportion to the relative abundance of each gene as computed in the first phase.

Figure 6 shows the relative error of coverage in the artificial repeat regions after resolution with ORMAN and RESCUE. This measure is calculated as:

$$\frac{|c_{original} - c_{orman}|}{c_{tophat}} \qquad (5)$$

where $c_{original}$, c_{tophat} and c_{orman} are the original coverage, raw coverage after the second TopHat mappings and coverage after multiread resolution with ORMAN, respectively. The relative error for RESCUE is defined similarly.

On genes APPBP2, CD164, PPM1H, RCOR1, RYBP, SERPINB6, SSR2, TXNDC16, UQCRC2 and ZBTB42, we see that ORMAN produces lower error values than RESCUE, whereas in the rest of the genes, it produces a higher relative

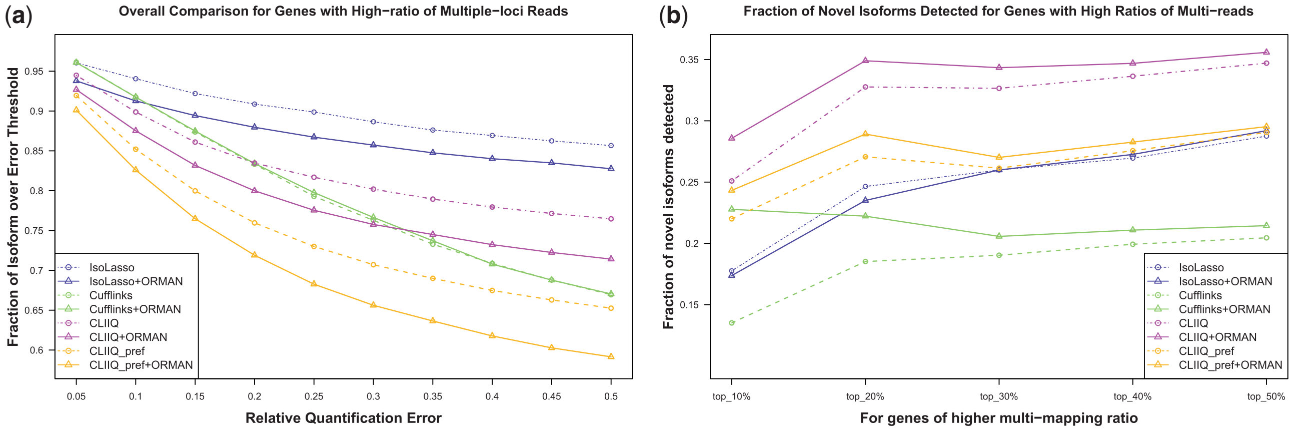

Fig. 5. Comparative performance of each tool and its enhanced version with ORMAN on selected genes that produce multireads, measured as the proportion of transcripts whose relative quantification error is above a threshold, as a function of the threshold. We show results of three tools (ORIGINAL) as well as their ORMAN enhanced versions for 3784 genes containing high ratio of multi-loci reads (left). We also examine the performance of novel isoform detections for gene whose multiread ratio ranked as top 10–50% in the whole sample (right)

Table 2. The genes and the artificial repeat locations used in the experiments

	Replacement						Decoy				
Gene	Chromosome	Strand	Start	End	# of reads	Gene	Chromosome	Strand	Start	End	# of reads
ZBTB42	14	+	105269519	105269918	1026	PPM1H	12	–	63041681	63042080	1839
NFE2L1	17	+	46128078	46128477	3697	UBL3	13	–	30339161	30339560	1949
USP5	12	+	6975253	6975652	931	BCL2L2	14	+	23776979	23777378	708
CD164	6	–	109689719	109690118	9461	TXNDC16	14	–	52898046	52898445	2072
APPBP2	17	–	58522733	58523132	733	RCOR1	14	+	103193777	103194176	902
SERPINB6	6	–	2948403	2948802	6203	UQCRC2	16	+	21994419	21994818	1196
SCAMP2	15	–	75136401	75136800	2386	USP43	17	+	9632438	9632837	838
UBE2K	4	+	39780509	39780908	1724	MUL1	1	–	20827015	20827414	1108
SSR2	1	–	155978849	155979248	7319	RYBP	3	–	72426808	72427207	289
COPG	3	+	128996147	128996546	6110	STK38	6	–	36462615	36463014	935

Note: ‘# of reads’ denotes the initial number of reads mapping to the 400-bp region that is used to introduce the artificial repeats.

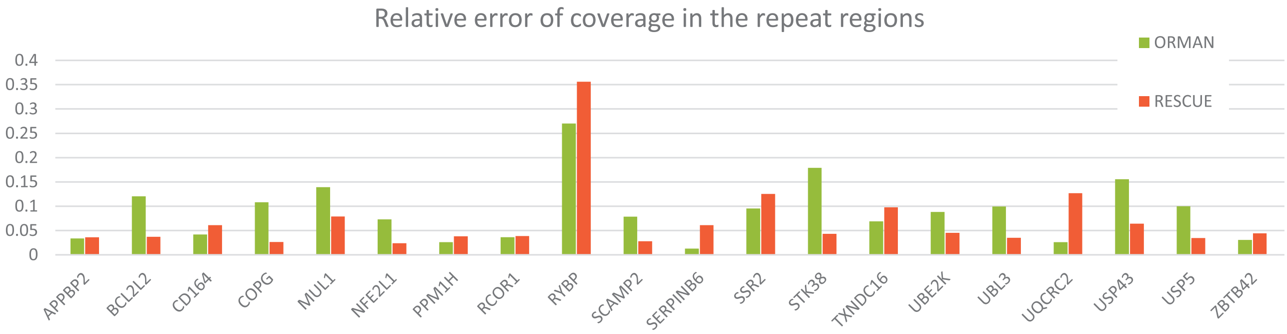

Fig. 6. The relative error of coverage in the repeat regions after multimapping resolution by ORMAN and RESCUE in the decoy and replacement genes used in the experiments

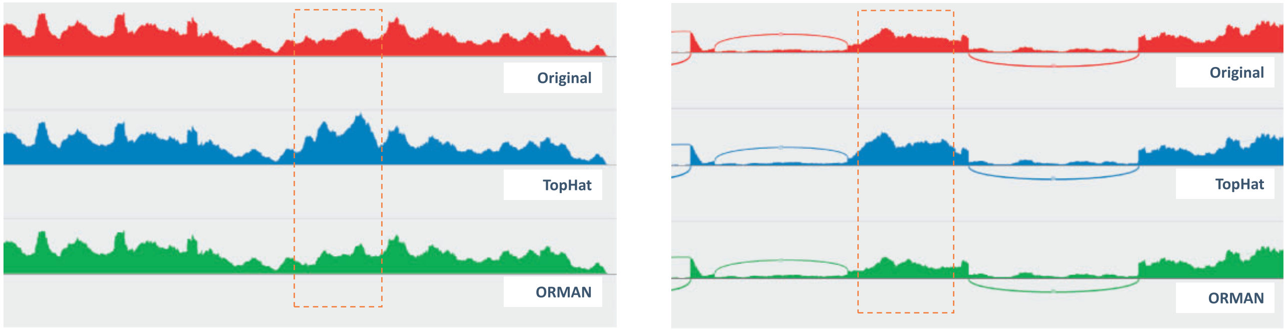

Fig. 7. Coverage plots for two genes (left: ZBTB42, right: BCL2L2) before and after processing with ORMAN compared with the original mappings. Top track (red) shows the mappings in the unaltered dataset; middle track (blue) shows the TopHat mappings after artificial repeats are introduced and the bottom track (green) shows the mappings after multiread resolution with ORMAN. The boxes outlined with dashed orange lines depict the artificial repeat region

error. On the other hand, the relative error of ORMAN never exceeds 0.3. Furthermore, a closer look on some of the genes suggests that ORMAN is still able to reproduce the look of the original distribution quite well despite the fact that RESCUE has a lower relative error. Figure 7 illustrates two such genes. Note that although both genes have a high variation in coverage, the coverage distribution in the repeat region is close to the original distribution after processing with ORMAN.

4 DISCUSSION

In this article, we introduce a combinatorial optimization formulation for resolving mapping ambiguity of RNA-Seq reads. Using a simulated RNA-Seq dataset on humans, we have shown that ORMAN improves the performance of popular computational tools in transcript identification and quantification, especially for genes with novel isoforms. Furthermore, our experiments based on real RNA-Seq reads suggest that the localized approach of ORMAN is able to approximate the original read distribution of the multimapping regions even in genes with highly variable coverage. Although ORMAN's performance was similar to that of RESCUE in our small-scale experiment, we suspect that datasets that suffer from elevated level of sequencing biases such as severe RNA degradation could benefit even more from our approach.

Funding: This research was supported by NSERC Discovery and Discovery Accelerate Grant to S.C.S., Canadian Cancer Society Innovation Grant to C.C. and S.C.S., Vanier Canada Graduate Scholarship to I.N. and Ebco/Epic Visiting Chairs and ICGC Grant to N.D.

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*., **38**, 4570–4578.

Chvatal,V. (1979) A greedy heuristic for the set-covering problem. *Math. Oper. Res*., **4**, 233–235.

Dohm,J.C. *et al*. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*., **36**, e105.

Karakoc,E. *et al*. (2012) Detection of structural variants and indels within exome data. *Nat. Methods*, **9**, 176–178.

Lapuk,A.V. *et al*. (2012) From sequence to molecular pathology, and a mechanism driving the neuroendocrine phenotype in prostate cancer. *J. Pathol*., **227**, 286–297.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li,W. *et al*. (2011) Isolasso: a lasso regression approach to RNA-seq based transcriptome assembly. *J. Comput. Biol*., **18**, 1693–1707.

Lin,Y.-Y. *et al*. (2012) CLIIQ: accurate comparative detection and quantification of expressed isoforms in a population. *Algorithms Bioinformatics*, **7534**, 178–189.

Mezlini,A.M. *et al*. (2013) iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res*., **23**, 519–529.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Nicolae,M. *et al.* (2011) Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol. Biol.*, **6**, 9.

Pasaniuc,B. *et al.* (2011) Accurate estimation of expression levels of homologous genes in RNA-seq experiments. *J. Comput. Biol.*, **18**, 459–468.

Roberts,A. *et al.* (2011) Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biol.*, **12**, 1–14.

Shen,Y. *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–120.

Shi,L. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.

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

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Wang,K. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.

Wu,Z. *et al.* (2011) Using non-uniform read distribution models to improve isoform expression inference in RNA-seq. *Bioinformatics*, **27**, 502–508.

Yorukoglu,D. *et al.* (2012) Dissect: detection and characterization of novel structural alterations in transcribed sequences. *Bioinformatics*, **28**, i179–i187.