Flexible Approach for Novel Transcript Reconstruction from RNA-Seq Data using Maximum Likelihood Integer Programming

Serghei Mangul
Georgia State University
Atlanta, GA, 30303, USA
serghei@cs.gsu.edu

Adrian Caciula
Georgia State University
Atlanta, GA, 30303, USA
acaciula@cs.gsu.edu

Sahar A. Seesi
University of Connecticut
Storrs, CT, 06269, USA
sahar@engr.uconn.edu

Dumitru Brinza
Life Technologies Corporation
Foster City, CA, 94404, USA
Dumitru.Brinza@lifetech.com

Abdul R. Banday
University of Connecticut
Storrs, CT, 06269, USA
abdul.banday@uconn.edu

Rahul Kanadia
University of Connecticut
Storrs, CT, 06269, USA
rahul.kanadia@uconn.edu

Ion Mandoiu
University of Connecticut
Storrs, CT, 06269, USA
ion@engr.uconn.edu

Alex. Zelikovsky
Georgia State University
Atlanta, GA, 30303, USA
alexz@cs.gsu.edu

Abstract

In this paper, we propose a novel, intuitive and flexible approach for transcriptome reconstruction from single RNA-Seq reads, called " Maximum Likelihood Integer Programming " (MLIP) method. MLIP creates a splice graph based on aligned RNA-Seq reads and enumerates all maximal paths corresponding to putative transcripts. The problem of selecting true transcripts is formulated as an integer program which minimizes the number of selected candidate transcripts. Our method purpose is to predict the minimum number of transcripts explaining the set of input reads with the highest quantification accuracy. This is achieved by coupling a integer programming formulation with an expectation maximization model for transcript expression estimation. MLIP has the advantage of offering different levels of stringency that would gear the results towards higher precision or higher sensitivity, according to the user preference. We test MLIP method on simulated and real data, and we show that MLIP outperforms both Cufflinks and IsoLasso.

1 Introduction

Massively parallel whole transcriptome sequencing, commonly referred to as RNA-Seq, and its ability to generate full transcriptome data at the single transcript level, provides a powerful tool for gene and isoform specific expression profiling. As a result, RNA-Seq has become the technology of choice for performing transcriptome analysis, rapidly replacing array-based technologies. Most current research, using RNA-Seq, employs methods that depend on existing transcriptome annotations. Unfortunately, as shown by recent targeted RNA-Seq studies [1], existing transcript libraries still miss large numbers of transcripts. The incompleteness of annotation libraries poses a serious limitation to using this powerful technology since accurate normalization of RNA-Seq data critically requires knowledge of expressed transcript sequences [2, 3, 4, 5]. As a result, transcript discovery from RNA-Seq has been the focus of many research in recent years. The sequences of novel transcripts can be reconstructed from deep RNA-Seq data, but this is computationally challenging due to sequencing errors, uneven coverage of expressed transcripts, and the need to distinguish between highly similar transcripts produced by alternative splicing.

2 Related Work

A number of recent works have addressed the problem of transcriptome reconstruction from RNA-Seq reads. These methods fall into three categories: "genome-guided", "genome-independent" and "annotation-guided" methods [6]. Genome-independent methods such as Trinity [7] or transAbyss [8] directly assemble reads into transcripts. A commonly used approach for such methods is de Brujin graph [9] utilizing "k-mers". The use of genome-independent methods becomes essential when there is no trusted genome reference that can be used to guide reconstruction. On the other end of the spectrum, annotation guided methods [10, 11, 12] make use of available information in existing transcript annotations to aid in the discovery of novel transcripts. RNA-Seq reads can be mapped onto reference genome, reference annotations, exon-exon junction libraries, or combinations thereof, and the resulting alignments are used to reconstruct transcripts.

Many transcriptome reconstruction methods fall in the genome-guided category. They typically start by mapping sequencing reads onto the reference genome,using spliced alignment tools, such as TopHat [13] or SpliceMap [14]. The spliced alignments are used to identify exons and transcripts that explain the alignments. While some methods aim to achieve the highest sensitivity, others work to predict the smallest set of transcripts explaining the given input reads. Furthermore, some methods aim to reconstruct the set of transcripts that would insure the highest quantification accuracy. Scripture [15] construct a splicing graph from the mapped reads and reconstructs isoforms corresponding to all possible paths in this graph. It then uses paired-end information to filter out some transcripts. Although scripture achieves very high sensitivity, it may predict a lot of incorrect isoforms. The method of Trapnell et al. [16, 17], referred to as Cufflinks, constructs a read overlap graph and generates candidate transcripts by finding a minimal size path cover via a reduction to maximum matching in a weighted bipartite graph. TRIP [18] uses an integer programming model where the objective is to select the smallest set of putative transcripts that yields a good statistical fit between the fragment length distribution empirically determined during library preparation and fragment lengths implied by mapping read pairs to selected transcripts. Cufflinks, Scripture, and TRIP do not target the quantification accuracy. IsoLasso [19] uses the LASSO [20] algorithm, and it aims to achieve a balance between quantification accuracy and predicting the minimum number of isoforms. It formulates the problem as a quadratic programming one, with additional constraints to ensure that all exons and junctions supported by the reads are included in the predicted isoforms. CLIIQ [21] uses an integer linear programming solution that minimizes the number of predicted isoforms explaining the RNA-Seq reads while minimizing the difference between estimated and observed expression levels of exons and junctions within the predicted isoforms.

In this paper, we present a genome guided method for transcriptome reconstruction from RNA-Seq reads. Our method aims to predict the minimum number of transcripts explaining the set of input reads with the highest quantification accuracy. This is achieved by coupling a integer programming formulation with an expectation maximization model for isoform expression estimation.

Recent advances in Next Generation Sequencing (NGS) technologies made it possible to produce longer single-end reads with the length comparable to length of fragment for paired-end technology[22] . The primary goal of our study is to developed a method for longer single-end reads. In future we plan to extend out method to support paired-end reads. We test our method on simulated and real data and compare with two of the available genome guided methods, Cufflinks and IsoLasso.

3 Methods

The maximum likelihood integer programming (MLIP) method starts from a set of putative transcripts and selects the subset of this transcripts with the highest support from the RNA-Seq reads. We formulate this problem as an integer program. The objective is to select the smallest set of putative transcripts that sufficiently explain the RNA-Seq data. Further, maximum likelihood estimator is applied to all possible combinations of putative transcripts of minimum size determined by integer program. Our method offers different level of stringency from low to high. Low stringency gives priority to sensitivity of reconstruction over precision of reconstruction, high stringency gives priority to precision over sensitivity. The default parameter of the MLIP method is medium stringency that achieves balance between sensitivity and precision of reconstruction

3.1 Model description

We use a splice graph (SG) to represent alternatively spliced isoforms for every gene in a sample. A SG is a directed acyclic graph where each vertex in the graph represents a segment of a gene. Two segments are connected by an edge if they are adjacent in at least one transcript. To partition a gene into a set of non-overlapping segments, information about alternative variants is used. Typically, alternative variants occurs due alternative transcriptional events and alternative splicing events [23] . Transcriptional events include: alternative first exon, alternative last exon. Splicing events include: exon skipping, intron retention, alternative 5’ splice site (A5SS), and alternative 3’ splice site (A3SS). Transcriptional events may consist only of non-overlapping exons. If exons partially overlap and they serve as a first or last exons we will refer to such event as A5SS or A3SS respectively.

Figure 1-A shows an example of a gene with 4 different exons, and 3 transcripts produced by alternative splicing. To represent such alternative variants we suggest to process the gene as a set of so called “pseudo-exons” based on alternative variants obtained from aligned RNA-seq reads. A *pseudo-exon* is a region of a gene between consecutive transcriptional or splicing events, i.e. starting or ending of an exon, as shown in figure 1-B. Hence every gene has a set of non-overlapping pseudo-exons, from which it is possible to reconstruct a set of putative transcripts.

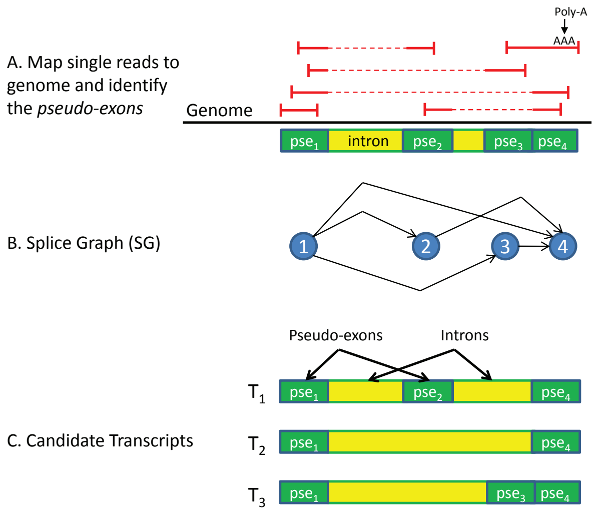


Figure 1: Model Description. A - Pseudo-exons. Pseudo-exons(green boxes) : regions of a gene between consecutive transcriptional or splicing events; B - Splice graph. The red horizontal lines represent single-end reads. Reads interrupted by dashed lines are spliced reads. Each vertex of the splice graph corresponds to a pseudo-exon and each directed edge corresponds to a (spliced) junction between two pseudo-exons; C - Candidate Transcripts. Candidate transcripts corresponds to maximal paths in the splice graph, which are enumerated using a depth-first-search algorithm.

SG is a directed acyclic graph (see figure 1-B), whose vertices represent pseudo-exons and edges represent pairs of pseudo-exons immediately following one another in at least one transcript (which is witnessed by at least one spliced read, as depicted in figure 1-B with red lines).

First we infer exon-exon junction from mapped reads, this information is used to build the SG. Then we enumerate all maximal paths in the SG using a depth-first-search algorithm. These paths correspond to putative transcripts and are the input for the MLIP algorithm. A gene with n pseudo-exons may have up to $2^n - 1$ possible candidate transcripts, each composed of a subset of the n pseudo-exons. Actual number of candidate transcripts depends on number of exons, this way splitting exons into pseudo-exons has no effect on number of candidate transcripts.

Information about poly-A site (PAS) can be integrated in the SG which improves accuracy of candidate transcript set. The PAS represents transcription end site of the transcript. Theoretically, any vertex in the splicing graph can serve as PAS, which will lead to increased number of false candidates transcripts. For this reason we computationally infer PAS from the data. Alternatively, one can use existing annotation for PAS or specialized protocols such as the PolyA-Seq protocol [24].

3.2 Maximum Likelihood Integer Programming Solution

Here we introduce 2-step approach for novel transcript reconstruction from single-end RNA-Seq reads. First, we introduce the integer program (IP) formulation, which has an objective to minimize number of transcripts sufficiently well covering observed reads. Since such formulation can lead to many identical optimal solutions we will use the additional step to select maximum likelihood solution based on deviation between observed and expected read frequencies. As with many RNA-Seq analyses, the preliminary step of our approach is to map the reads. We map reads onto the genome reference using any of the available splice alignment tools (we use TopHat[13] with default parameters in our experiments).

3.2.1 1st step : Integer Program Formulation

We will use the following notations in our IP formulation:

N	total number of candidate ;
R	total number of reads ;
J_l	l-th spliced junction;
P_l	l-th poly-A site(PAS);
r	single-read, $1 \leq j \leq R$;
t	candidate transcript , $1 \leq k \leq K$;
T	set of candidate transcripts
$T(r)$	set of candidate transcripts where read r can be mapped

For a given instance of the transcriptome reconstruction problem, we formulate the IP. The boolean variables used in IP formulation are:

$x(r \to t)$	1 iff read r is mapped into transcript t and 0 otherwise;
$y(t)$	1 if candidate transcript t is selected, and 0 otherwise;
$x(r)$	1 if the read r is mapped , and 0 otherwise;

The IP objective is to minimize the number of candidate transcripts subject to the constraints (1)-(5):

$$\sum_{t \in T} y(t) \to min$$

Subject to:
(1) For any r, at least one transcript t is selected: $y(t) \geq x(r \to t), \forall r, \forall t$

(2) Read r can be mapped only to one transcript: $\sum_{t \in T(r)} x(r \to t) = x(r), \forall r$

(3) Selected transcripts cover almost all reads: $\sum\limits_{r \in R} x(r) \geq N(1-\epsilon)$

(4) Each junction is covered by at least one selected transcript: $\sum\limits_{t \in J_l} y(t_k) \geq 1, \forall J_l$

(5) Each PAS is covered by at least one selected transcript: $\sum\limits_{t_k \in P_l} y(t_k) \geq 1, \forall P_l$

We use CPLEX [25] to solve the IP, the rest of implementation is done using Boost C++ Libraries and bash scripting language.

3.2.2 2nd step : Maximum Likelihood Solution

In the second step we enumerate all possible subsets of candidate transcripts of size N, where N is determined by solving transcriptome reconstruction IP, that satisfy the following condition: every spliced junction and PAS to be present in the subset of transcripts at least once. Further, for every such subset we estimate the most likely transcript frequencies and corresponding expected read frequencies. The algorithm chooses subset with the smallest deviation between observed and expected read frequencies.

The model is represented by bipartite graph $G = \{T \bigcup R, E\}$ in which each transcript is represented as a vertex $t \in T$, and each read is represented as a vertex $r \in R$. With each vertex $t \in T$, we associate frequency f of the transcript. And with each vertex $r \in R$, we associate observed read frequency o_r. Then for each pair t, r, we add an edge (t, r) weighted by probability of transcript t to emit read r.

Given the model we will estimate maximum likelihood frequencies of the transcripts using our previous approach, refer as IsoEM [4]. Regardless of initial conditions IsoEM algorithm always converge to maximum likelihood solution (see [26]).The algorithm starts with the set of T transcripts. After uniform initialization of frequencies $f_t, t \in T$, the algorithm repeatedly performs the next two steps until convergence:

- E-step: Compute the expected number $n(t_k)$ of reads that come from transcript t_k under the assumption that transcript frequencies $f(t)$ are correct, based on weights h_{t_k,r_j}
- M-step: For each t_k, set the new value of f_t to the portion of reads being originated by transcript t among all observed reads in the sample

We suggest to measure the model quality, i.e. how well the model explains the reads, by the deviation between expected and observed read frequencies as follows:

$$D = \frac{\sum_j |o_j - e_j|}{|R|}, \tag{1}$$

where $|R|$ is number of reads, o_j is the observed read frequency of the read r_j and e_j is the expected read frequencies of the read r_j calculated as follows:

$$e_j = \sum_{r_j} \frac{h_{t_k,r_j}}{\sum_{r_j} h_{t_k,r_j}} f_t^{ML} \tag{2}$$

where h_{t_k,r_j} is weighted match based on mapping of read r_j to the transcript t_k and f_t^{ML} is the maximum-likelihood frequency of the transcript t_k.

The flowchart of MLIP is depicted in figure 2.

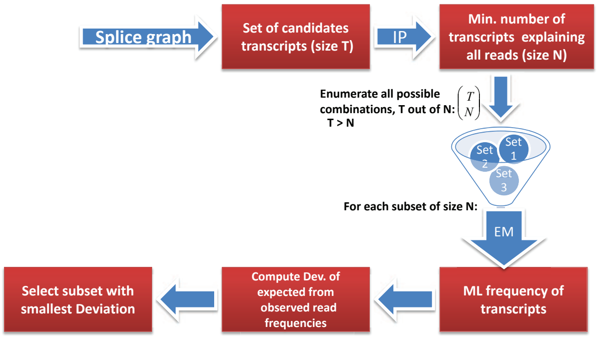

Splice graph
Set of candidates transcripts (size T)
IP
Min. number of transcripts explaining all reads (size N)
Enumerate all possible combinations, T out of N: $\binom{T}{N}$ T > N
Set 1
Set 2
Set 3
For each subset of size N:
EM
ML frequency of transcripts
Compute Dev. of expected from observed read frequencies
Select subset with smallest Deviation

Figure 2: Flowchart for MLIP method. Input : Splice graph. Output : subset of candidate transcripts with the smallest deviation between observed and expected read frequencies.

Figure 3 illustrates how MLIP works on a given synthetic gene with 3 transcripts and 7 different exons (see figure 3-A). First we use mapped reads to construct the splice graph from which we generate T possible candidate transcripts, as shown in figure 3-B. Further we run our IP approach to obtain N minimum number of transcripts that explain all reads. We enumerate all feasible subsets of candidate transcripts, having cardinality N. The subsets which doesn't cover all junctions and PAS will be excluded from consideration. The subset with the smallest deviation between expected and observed read frequencies is selected by the PAP algorithm.

3.2.3 Stringency of Reconstruction

Different level of stringency corresponds to different strategies of transcriptome reconstruction. High stringency has the goal to optimize precision of reconstruction, with some loss in sensitivity. On the other hand, low stringency corresponds to increase in sensitivity and some decrease in prediction. Medium stringency strikes balance between sensitivity and precision of reconstruction. The medium stringency is chosen as a default setting for the proposed MLIP method.

Below, we will describe how different stringency levels are computed. For the default medium level we will

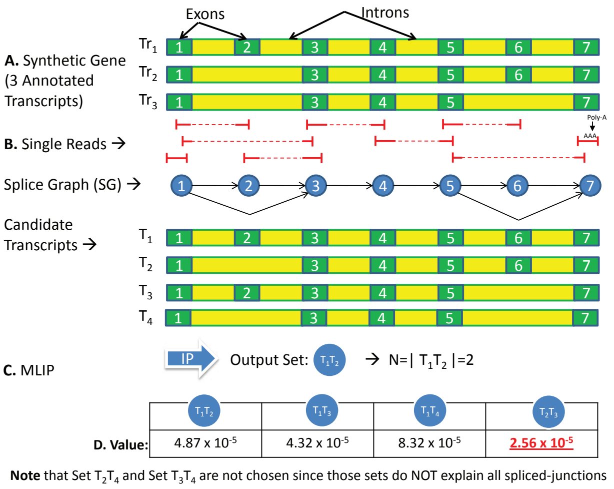


Figure 3: A. Synthetic gene with 3 transcripts and 7 different exons. B. Mapped reads are used to construct the splice graph from which we generate T possible candidate transcripts. C. MLIP. Run IP approach to obtain N minimum number of transcripts that explain all reads. We enumerate N feasible subsets of candidate transcripts.The subsets which doesn't cover all junctions and PAS will be excluded from consideration. The subset with the smallest deviation between expected and observed read frequencies is selected by the PAP algorithm.

use the subset of candidate transcripts selected based on the smallest deviation between observed and expected read frequency. For the low stringency level, our method selects the subset of transcripts that will correspond to the union of the solution obtained by solving the IP and the solution supported by the smallest deviation. High stringency level will correspond to the intersection of above solutions.

4 Experimental Results

4.1 Simulation Setup and Matching Criteria

Simulation Setup. We first evaluated performance of our MLIP solution on simulated human RNA-Seq data. The human genome sequence (hg18, $NCBI$ build 36) was downloaded from $UCSC$ together with the the KnownGenes transcripts annotation table. Genes were defined as clusters of known transcripts defined by the GNFAtlas2 table.

In our simulation experiment, we simulate reads together with splice read alignment to the genome, splice read alignment is provided for all methods. We varied the length of single-end reads, which were randomly generated per gene by sampling fragments from known transcripts maintaining $100x$ coverage per transcript. In order to compare different next generation sequencing (NGS) platforms, including the most recent one able to produce longer reads, all the methods were run on datasets with various read length, i.e. 50bp, 100bp, 200bp, and 400bp. Expression levels of transcripts inside gene cluster follows uniform and geometric distribution. To address library preparation process for RNA-Seq experiment we simulate fragment lengths from a normal probability distribution with different mean and 10% standard deviation.

Matching Criteria. All reconstructed transcripts are matched against annotated transcripts. Two transcripts match iff internal pseudo-exon boundaries coordinates (i.e., all pseudo-exons coordinates except the beginning of the first pseudo-exon and the end of the last pseudo-exon) are identical. Similar matching criteria is suggested in [16] and [27].

We use $Sensitivity$, $Precision$ and *F-Score* to evaluate the performance of different methods. Sensitivity is defined as the proportion of reconstructed sequences that match annotated transcript sequences, i.e.,

$$Sens = \frac{TP}{TP + FN}$$

where TP (True Positive) represents the number of correctly reconstructed transcripts and FN (False Negative) is the number of incorrectly reconstructed transcripts.

Precision is defined the proportion of annotated transcript sequences among reconstructed sequences, i.e.,

$$Prec = \frac{TP}{TP + FP}$$

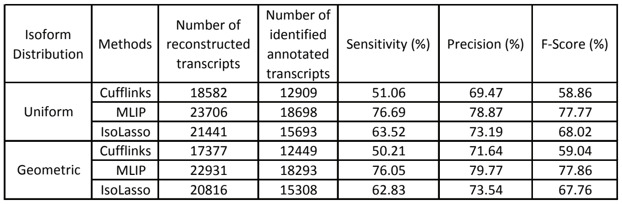
and the F-Score is defined as the harmonic mean of $Sensitivity$ and $Precision$, i.e.,

$$\textit{F-Score} = 2 \times \frac{Prec \times Sens}{Prec + Sens}$$

4.2 Comparison of Methods on Simulated Data

In this section, we use sensitivity, precision, and F-score defined above to compare the MLIP method to the other genome guided transcriptome reconstruction tools. The most recent versions of Cufflinks (version 2.0.0) from [16] and IsoLasso (v 2.6.0) from [27] are used for comparison with the default parameters. We explore the influence of read length, fragment length, and coverage on reconstruction accuracy.

Figure 4 reports the transcriptome reconstruction accuracy for reads of length 400bp, simulated assuming both uniform and geometric distribution for transcript expression levels. MLIP significantly overperforms the other methods, achieving an F-score over 79% for all datasets. For all methods the accuracy difference between datasets generated assuming uniform and geometric distribution of transcript expression levels is small, with the latter one typically having a slightly worse accuracy. Thus, in the interest of space we present remaining results for datasets generated using uniform distribution.



Isoform Distribution	Methods	Number of reconstructed transcripts	Number of identified annotated transcripts	Sensitivity (%)	Precision (%)	F-Score (%)
Uniform	Cufflinks	18582	12909	51.06	69.47	58.86
	MLIP	23706	18698	76.69	78.87	77.77
	IsoLasso	21441	15693	63.52	73.19	68.02
Geometric	Cufflinks	17377	12449	50.21	71.64	59.04
	MLIP	22931	18293	76.05	79.77	77.86
	IsoLasso	20816	15308	62.83	73.54	67.76

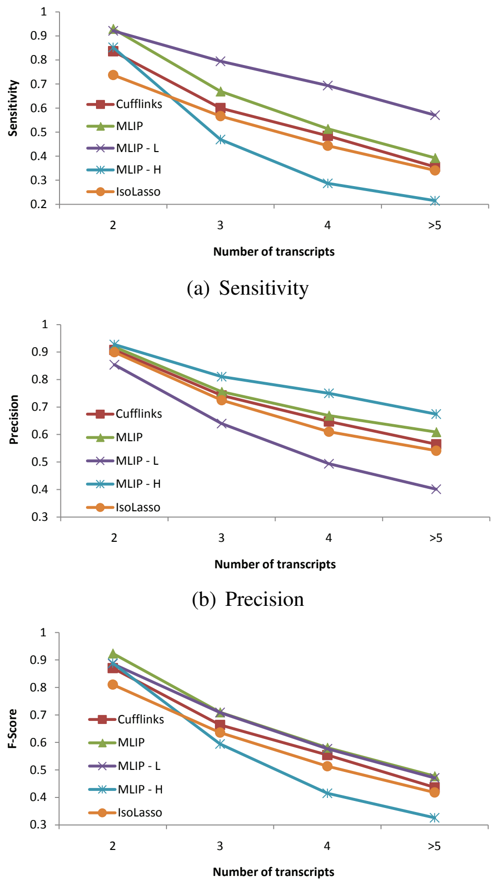
Figure 4: Transcriptome reconstruction results for uniform and geometric fragment length distribution. Sensitivity, precision and F-Score for transcriptome reconstruction from reads of length 400bp, mean fragment length 450bp and standard deviation 45bp simulated assuming uniform, respectively geometric expression of transcripts.

Intuitively, it seems more difficult to reconstruct the alternative splicing transcripts in genes with higher number of alternative variants. There is a strong correlation between number of alternative variants and number of annotated transcripts. Also high number of alternative variants leads to high number of candidate transcripts, which make difficult the selection process. To explore the behavior of the methods depending on number of annotated transcripts we divided all genes into categories according to the number of annotated transcripts and calculated the sensitivity, precision and F-Score of the methods for every such category.

Figures 5(a)-5(c) compare the performance of 5 methods (Cufflinks, IsoLasso, MLIP - medium stringency settings, $MLIP-L$ - low stringency settings, $MLIP-H$ - high stringency settings) for read length 100bp and fragment length 250bp. Genes are divided into 4 categories according to number of annotated transcripts per gene. In this experiment, we present results for the three different stringency settings for MLIP i.e. low, medium, and high. For the medium stringency (default settings), MLIP achieves better results in both sensitivity and precision. As for F-score, the best results are produced by low and medium stringency versions of MLIP, with different trade-off between sensitivity and precision.

Figure 6 compares sensitivity, precision and F-score of Cufflinks, IsoLasso, and MLIP for different combinations of read and fragment lengths: (50bp,250bp), (100bp,250bp), (100bp,500bp), (200bp,250bp), (400bp,450bp). The results show that MLIP provide 5-15% improvement in sensitivity and 1-10% improvement in precision.

In order to explore influence of coverage on precision and sensitivity of reconstruction we simulated 2 datasets with $100X$ and $20X$ coverage. Figure 7 shows how accuracy of transcriptome reconstruction depends on the coverage. For all methods higher coverage ($100X$ vs. $20X$) doesn't provide significant improvement in precision and sensitivity.



(a) Sensitivity

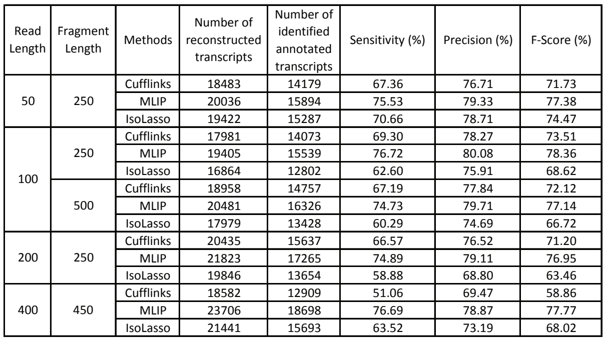
(b) Precision

(c) F-Score

Figure 5: Transcriptome reconstruction results with respect to number of transcripts per gene. Comparison between 5 methods (Cufflinks, IsoLasso, MLIP - medium stringency settings, $MLIP-L$ - low stringency settings, $MLIP-H$ - high stringency settings) for groups of genes with n transcripts(n=1,..., $\geq$ 5) on simulated dataset with mean fragment length 250bp, standard deviation 25bp and read length of 100bp.

4.3 Comparison of Methods on real RNA-Seq dataset

We tested MLIP on real RNA-Seq data that we sequenced from a CD1 mouse retina RNA samples. We selected a specific gene that has 33 annotated transcripts in Ensembl. The dataset used consists of 46906 alignments for 44692 single reads of length 68 bp. The read alignments falling within the genomic locus of the selected gene were used to construct a splicing graph; then MLIP with default



Read Length	Fragment Length	Methods	Number of reconstructed transcripts	Number of identified annotated transcripts	Sensitivity (%)	Precision (%)	F-Score (%)
50	250	Cufflinks	18483	14179	67.36	76.71	71.73
		MLIP	20036	15894	75.53	79.33	77.38
		IsoLasso	19422	15287	70.66	78.71	74.47
100	250	Cufflinks	17981	14073	69.30	78.27	73.51
		MLIP	19405	15539	76.72	80.08	78.36
		IsoLasso	16864	12802	62.60	75.91	68.62
	500	Cufflinks	18958	14757	67.19	77.84	72.12
		MLIP	20481	16326	74.73	79.71	77.14
		IsoLasso	17979	13428	60.29	74.69	66.72
200	250	Cufflinks	20435	15637	66.57	76.52	71.20
		MLIP	21823	17265	74.89	79.11	76.95
		IsoLasso	19846	13654	58.88	68.80	63.46
400	450	Cufflinks	18582	12909	51.06	69.47	58.86
		MLIP	23706	18698	76.69	78.87	77.77
		IsoLasso	21441	15693	63.52	73.19	68.02

Figure 6: Transcriptome reconstruction results for various read and fragment lengths. Sensitivity, precision and F-score for different combinations of read and fragment lengths: (50bp,250bp), (100bp,250bp), (100bp,500bp), (200bp,250bp), (400bp,450bp).

settings(medium stringency) was used to select candidate transcripts. MLIP method was able to infer 5 out of 10 transcripts confirmed by qPCR while cufflinks reconstructed 3 out of 10 and IsoLasso 1 out of 10 transcripts.

5 Future Work

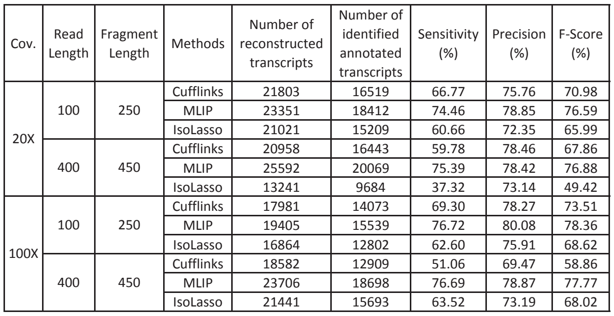
As a part of our future work, we plan to extend MLIP to support paired-end reads. Also we plan to compute the deviation not only for minimum number of transcripts N reported by IP, but also for $N + 1$. This will allow us to address cases when set of transcripts in a gene cannot be explained by a set of minimum transcripts.

6 Conclusions

In this paper we propose and implemented MLIP, a novel, intuitive and flexible method for transcriptome reconstruction from single RNA-Seq reads. Our method has the advantage of offering different levels of stringency that would gear the results towards higher precision or higher sensitivity, according to the user preference. Preliminary experimental results on both real and synthetic datasets generated with various sequencing parameters and distribution assumptions show that our MLIP approach is scalable and has increased transcriptome reconstruction accuracy compared to previous approaches.

Acknowledgements

This work has been partially supported by Life Technologies Collaborative Research Grant AG110891 on "Software for Robust Transcript Discovery and Quantification", NSF award IIS-0916401,NSF award IIS-0916948, Agriculture and Food Research Initiative Competitive Grant no. 201167016-30331 from the USDA National Institute of Food and Agriculture and Second Century Initiative Bioinformatics University Doctoral Fellowship.



Cov.	Read Length	Fragment Length	Methods	Number of reconstructed transcripts	Number of identified annotated transcripts	Sensitivity (%)	Precision (%)	F-Score (%)
20X	100	250	Cufflinks	21803	16519	66.77	75.76	70.98
			MLIP	23351	18412	74.46	78.85	76.59
			IsoLasso	21021	15209	60.66	72.35	65.99
	400	450	Cufflinks	20958	16443	59.78	78.46	67.86
			MLIP	25592	20069	75.39	78.42	76.88
			IsoLasso	13241	9684	37.32	73.14	49.42
100X	100	250	Cufflinks	17981	14073	69.30	78.27	73.51
			MLIP	19405	15539	76.72	80.08	78.36
			IsoLasso	16864	12802	62.60	75.91	68.62
	400	450	Cufflinks	18582	12909	51.06	69.47	58.86
			MLIP	23706	18698	76.69	78.87	77.77
			IsoLasso	21441	15693	63.52	73.19	68.02

Figure 7: Transcriptome reconstruction results with respect to different coverages. Sensitivity, precision and F-Score for transcriptome reconstruction from reads of length 100bp and 400bp simulated assuming $20X$ coverage, respectively $100X$ coverage per transcript. For read length 100bp fragment length of 250 with 10% standard deviation was used. For read length 400bp fragment length of 450 with 10% standard deviation was used.

References

[1] T. R. Mercer, D. J. Gerhardt, M. E. Dinger, J. Crawford, C. Trapnell, J. A. Jeddeloh, J. S. Mattick, and J. L. Rinn, "Targeted RNA sequencing reveals the deep complexity of the human transcriptome." *Nature Biotechnology*, vol. 30, no. 1, pp. 99–104, 2012.

[2] A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nature methods*, 2008. [Online]. Available: http://dx.doi.org/10.1038/nmeth.1226

[3] E. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. Kingsmore, G. Schroth, and C. Burge, "Alternative isoform regulation in human tissue transcriptomes." *Nature*, vol. 456, no. 7221, pp. 470–476, 2008. [Online]. Available: http://dx.doi.org/10.1038/nature07509

[4] M. Nicolae, S. Mangul, I. Mandoiu, and A. Zelikovsky, "Estimation of alternative splicing isoform frequencies from rna-seq data," *Algorithms for Molecular Biology*, vol. 6:9, 2011. [Online]. Available: http://www.almob.org/content/6/1/9

[5] B. Li, V. Ruotti, R. Stewart, J. Thomson, and C. Dewey, "RNA-Seq gene expression estimation with read mapping uncertainty," *Bioinformatics*, vol. 26, no. 4, pp. 493–500, 2010. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btp692

[6] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," *Nature Methods*, vol. 8, no. 6, pp. 469–477, May 2011. [Online]. Available: http://dx.doi.org/10.1038/nmeth.1613

[7] M. Grabherr, "Full-length transcriptome assembly from rna-seq data without a reference genome." *Nature biotechnology*, vol. 29, no. 7, pp. 644–652, 2011. [Online]. Available: http://dx.doi.org/10.1038/nbt.1883

[8] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, and et al., "De novo assembly and analysis of rna-seq data." *Nature Methods*, vol. 7, no. 11, pp. 909–912, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20935650

[9] P. A. Pevzner, "1-Tuple DNA sequencing: computer analysis." *J Biomol Struct Dyn*, vol. 7, no. 1, pp. 63–73, Aug. 1989.

[10] A. Roberts, H. Pimentel, C. Trapnell, and L. Pachter, "Identification of novel transcripts in annotated genomes using rna-seq," *Bioinformatics*, 2011. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/early/2011/06/21/bioinformatics.btr355.abstract

[11] S. Mangul, A. Caciula, I. Mandoiu, and A. Zelikovsky, "Rna-seq based discovery and reconstruction of unannotated transcripts in partially annotated genomes," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, nov. 2011, pp. 118 –123.

[12] J. Feng, W. Li, and T. Jiang, "Inference of isoforms from short sequence reads," in *Proc. RECOMB*, 2010, pp. 138–157.

[13] C. Trapnell, L. Pachter, and S. Salzberg, "TopHat: discovering splice junctions with RNA-Seq." *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btp120

[14] K. F. Au, H. Jiang, L. Lin, Y. Xing, and W. H. Wong, "Detection of splice junctions from paired-end rna-seq data by splicemap," *Nucleic Acids Research*, 2010. [Online]. Available: http://nar.oxfordjournals.org/content/early/2010/04/05/nar.gkq211.abstract

[15] M. Guttman, M. Garber, J. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. Koziol, A. Gnirke, C. Nusbaum, J. Rinn, E. Lander, and A. Regev, "*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. [Online]. Available: http://dx.doi.org/10.1038/nbt.1633

[16] C. Trapnell, B. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. van Baren, S. Salzberg, B. Wold, and L. Pachter, "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. [Online]. Available: http://dx.doi.org/10.1038/nbt.1621

[17] A. Roberts, C. Trapnell, J. Donaghey, J. Rinn, and L. Pachter, "Improving rna-seq expression estimates by correcting for fragment bias," *Genome Biology*, vol. 12, no. 3, p. R22, 2011.

[18] S. Mangul, A. Caciula, S. Al Seesi, D. Brinza, A. R. Banday, R. Kanadia, I. Mandoiu, and A. Zelikovsky, "An integer programming approach to novel transcript reconstruction from paired-end rna-seq reads," *ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 2012.

[19] W. Li, J. Feng, and T. Jiang, "IsoLasso: A LASSO Regression Approach to RNA-Seq Based Transcriptome Assembly," *Journal of Computational Biology*, vol. 18, no. 11, pp. 1693–707, 2011. [Online]. Available: http://online.liebertpub.com/doi/full/10.1089/cmb.2011.0171

[20] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of Royal Statistical Society*, vol. 58, pp. 267–288, 1996.

[21] Y. Y. Lin, P. Dao, F. Hach, M. Bakhshi, F. Mo, A. Lapuk, C. Collins, and S. C. Sahinalp, "Cliiq: Accurate comparative detection and quantification of expressed isoforms in a population," *Proc. 12th Workshop on Algorithms in Bioinformatics*, 2012.

[22] J. M. Rothberg, W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, and et al., "An integrated semiconductor device enabling non-optical genome sequencing." *Nature*, vol. 475, no. 7356, pp. 348–352, 2011. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature10242

[23] S. Pal, R. Gupta, H. Kim, P. Wickramasinghe, V. Baubet, L. C. Showe, N. Dahmane, and R. V.

Davuluri, “Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development,” *Genome Research*, 2011. [Online]. Available: http://genome.cshlp.org/content/early/2011/06/28/gr.120535.111.abstract

[24] A. Derti, P. Garrett-Engele, K. D. MacIsaac, R. C. Stevens, S. Sriram, R. Chen, C. A. Rohl, J. M. Johnson, and T. Babak, “A quantitative atlas of polyadenylation in five mammals,” *Genome Research*, vol. 22, no. 6, pp. 1173–1183, 2012.

[25] IBM, “Inc: IBM ILOG CPLEX 12.1.” http://www.ibm.com/software/integration/optimization/cplex/, 2009.

[26] B. Paşaniuc, N. Zaitlen, and E. Halperin, “Accurate estimation of expression levels of homologous genes in RNA-seq experiments,” in *Proc. 14th Annual Intl. Conf. on Research in Computational Molecular Biology (RECOMB)*, ser. Lecture Notes in Computer Science, B. Berger, Ed., vol. 6044. Springer Berlin / Heidelberg, 2010, pp. 397–409.

[27] W. Li, J. Feng, and T. Jiang, “IsoLasso: A LASSO Regression Approach to RNA-Seq Based Transcriptome Assembly,” *Lecture Notes in Computer Science*, vol. 6577, pp. 168–+, 2011.