An Integer Programming Approach to Novel Transcript Reconstruction from Paired-End RNA-Seq Reads

Serghei Mangul
Dept. of Computer Science
Georgia State University
34 Peachtree Street
Atlanta, GA 30303
serghei@cs.gsu.edu

Adrian Caciula
Dept. of Computer Science
Georgia State University
34 Peachtree Street
Atlanta, GA 30303
acaciula@cs.gsu.edu

Sahar Al Seesi
Department of Computer Science & Engineering
University of Connecticut
371 Fairfield Way
Storrs, CT 06269-2155
sahar@engr.uconn.edu

Dumitru Brinza
Ion Bioinformatics
Life Technologies Corporation
850 Lincoln Center Drive
Foster City, CA 94404
Dumitru.Brinza@lifetech.com

Abdul Rouf Banday
Department of Physiology & Neurobiology
University of Connecticut
75 North Eagleville Road
Storrs, CT 06269-3156
abdul.banday@uconn.edu

Rahul Kanadia
Department of Physiology & Neurobiology
University of Connecticut
75 North Eagleville Road
Storrs, CT 06269-3156
rahul.kanadia@uconn.edu

ABSTRACT

Massively parallel whole transcriptome sequencing, commonly referred to as RNA-Seq, has become the technology of choice for performing gene expression profiling. However, reconstruction of full-length novel transcripts from RNA-Seq data remains challenging due to the short read length delivered by most existing sequencing technologies. We propose a novel statistical genome-guided method called "Transcriptome Reconstruction using Integer Programming" (TRIP) that incorporates fragment length distribution into novel transcript reconstruction from paired-end RNA-Seq reads. TRIP 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 (IP) which minimizes the set of selected transcripts yielding a good statistical fit between the fragment length distribution (empirically determined during library preparation) and fragment lengths implied by mapped read pairs. Experimental results on both real and synthetic datasets show that TRIP is more accurate than methods ignoring fragment length distribution information. The software is available at:
$http://www.cs.gsu.edu/\ serghei/?q=trip$

1. INTRODUCTION

Massively parallel whole transcriptome sequencing, commonly referred to as RNA-Seq, has become the technology of choice for performing gene and isoform specific expression profiling. However, accurate normalization of RNA-Seq data critically requires knowledge of expressed transcript sequences [13, 22, 14, 8]. Unfortunately, as shown by recent targeted RNA-Seq studies [12], existing transcript libraries still miss large numbers of transcripts. 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.

ACM-BCB'12, October 7-10, 2012, Orlando, FL, USA

1.1 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" [5]. "Genome-guided" and "annotation-guided" methods typically start by mapping sequencing reads onto the reference genome, reference annotations, exon-exon junction libraries, or combinations thereof. In case of mapping reads onto the reference genome, one needs to use spliced alignment tools, such as TopHat [20] or SpliceMap [2]. "Genome-guided" methods use the spliced genome alignments to identify exons and trnascripts that explains the alignments. These methods first map all reads to the reference genome and then use spliced reads to reconstruct transcriptome. Such methods, also called "ab initio", have been proposed in [21, 9, 7]. Guttman et al. [7] construct a splicing graph from the mapped reads and filter candidate transcripts using paired-end information, after performing spliced alignment of (paired) reads onto the genome. The method of Trapnell et al. [21], 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. Annotation-guided methods such as RABT Assembly [17, 11, 4] explicitly use existing annotations for the transcriptome reconstruction. A recent tool CLIIQ [10] uses an integer linear programming solution that minimizes

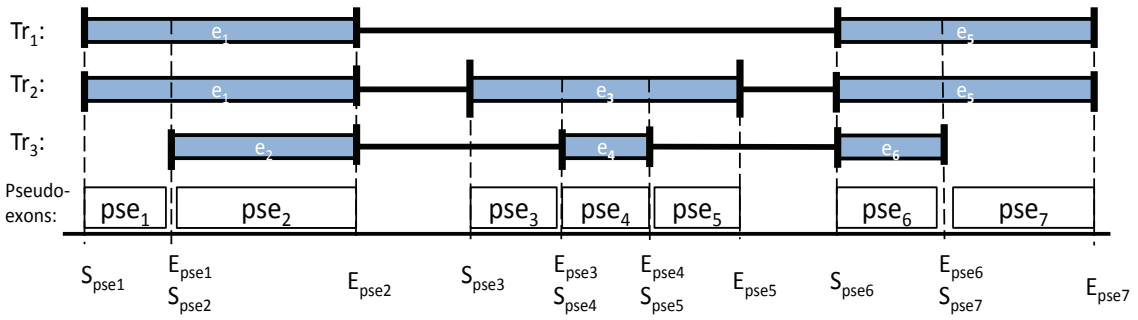


Figure 1: Pseudo-exons(white boxes) : regions of a gene between consecutive transcriptional or splicing events. An example of three transcripts $Tr_i, i = 1, 2, 3$ each sharing exons(blue boxes). S_{psej} and E_{psej} represent the starting and ending position of pseudo-exon j, respectively.

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. Although the CLIIQ results are promising, the tool is not yet available for comparison.

Rather than mapping reads to the reference genome first, genome-independent methods such as Trinity [6] or trans-Abyss [18] directly assemble reads into the transcripts. A commonly used approach for such methods is de Brujin graph [16] utilizing "k-mers" rather than reads for the graph construction. Because these tools do not rely on known reference genome their results are worst than genome guided transcriptome reconstruction tools.

1.2 Our Contributions

In this work, we propose a novel statistical "genome-guided" method called "**T**ranscriptome **R**econstruction using **I**nteger **P**rogramming" (TRIP). The method incorporates information about fragment length distribution of RNA-Seq paired end reads to reconstruct novel transcripts. First, we infer exon boundaries from spliced genome alignments of the reads. Then, we create a splice graph based on inferred exon boundaries. We enumerate all maximal paths in the splice graph corresponding to putative transcripts. The problem of selecting true transcripts is formulated as an integer program (IP) which minimizes the set of selected transcripts subject to a good statistical fit between the fragment length distribution (empirically determined during library preparation) and fragment lengths implied by mapped read pairs.

Experimental results on both real and synthetic datasets generated with various sequencing parameters and distribution assumptions show that TRIP has increased transcriptome reconstruction accuracy compared to previous methods that ignore information about fragment length distribution.

The rest of the paper is organized as follows: Section 2 describes the TRIP method including read mapping, splice graph construction and the IP formulation. The performance of proposed approaches are evaluated and analysed in Section 3. We conclude this work in Section 4.

2. TRANSCRIPTOME RECONSTRUCTION USING INTEGER PROGRAMMING

TRIP is a novel statistical "genome-guided" method that incorporates fragment length distribution into novel transcript reconstruction from paired-end RNA-Seq reads. The method starts from a set of maximal paths corresponding to putative transcripts and selects the subset of candidate transcript 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 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.

2.1 Read Mapping

As with many RNA-Seq analyses, the first step of TRIP is to map the reads. We map reads onto the genome reference using any of the available splice alignment tools (we use TopHat [20] with default parameters in our experiments). Note that a paired read consists of two reads flanking a fragment whose length usually follows normal distribution. The mean and variance of fragment length distribution are usually known in advance or can be inferred from read alignments.

2.2 Construction of Splice Graph and Enumeration of Putative Transcripts

Typically, alternative variants occurs due alternative transcriptional events and alternative splicing events [15] . 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 both serve as a first or last exons we will refer to such event as A5SS or A3SS respectively.

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. Hence every gene has a set of

non-overlapping pseudo-exons, from which it is possible to reconstruct a set of putative transcripts.

The notations used in Figure 1 represents the following:

e_i :	exon i ;
pse_j :	pseudo-exon j ;
S_{pse_j} :	start position of pseudo-exon j, $1 \leq j \leq 2n$;
E_{pse_j} :	end position of pseudo-exon j, $1 \leq j \leq 2n$;
Tr_i :	transcript i ;

A splice graph is a directed acyclic graph (see Fig. 2), 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). We enumerate all maximal paths in the splice graph using a depth-first-search algorithm. These paths correspond to putative transcripts and are the input for the TRIP algorithm. A gene with n pseudo-exons may have $2^n - 1$ possible candidate transcripts, each composed of a subset of the n pseudo-exons.

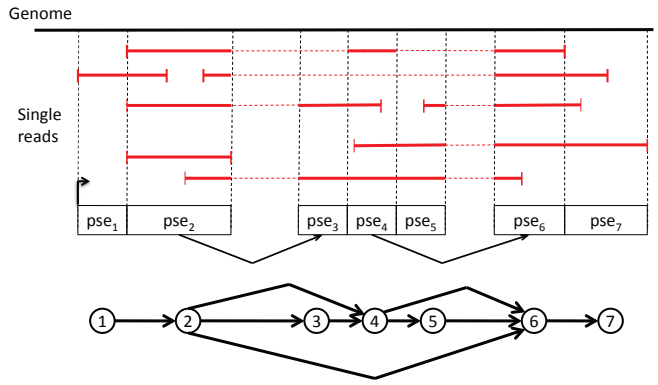


Figure 2: Splice graph. The red horizontal lines represent single reads. Reads interrupted by dashed lines are spliced reads. Each vertex of the splice graph corresponds to a pseudo-exon and each directed edge corresponds to a (splice) junction between two pseudo-exons.

Next section introduces an integer program producing minimal number of transcripts sufficiently well covering observed paired reads.

2.3 Integer Program Formulation

We will use the following notations in the Integer Program (IP) formulation :

N	Total number of reads ;
J_l	l-th splice junction;
p_j	paired-end read, $1 \leq j \leq N$;
t_k	k-th candidate transcript , $1 \leq k \leq K$;
s_i	Expected portion of reads mapped within i standard deviations ($s_1 \approx 68\%$, $s_2 \approx 95\%$, $s_3 \approx 99.7\%$);
ϵ	allowed deviation from the rule ($\epsilon = 0.05$)
$T_i(p_j)$	Set of candidate transcripts where p can be mapped with a fragment length between $i-1$ and i standard deviations, $1 \leq i \leq 3$;
$T_4(p_j)$	Set of candidates transcripts where p_j can be mapped with a fragment length within more than 3 standard deviations;

For a given instance of the transcriptome reconstruction problem, we formulate the integer program.

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

Subject to

(1) $\sum_{t_k \in T_i(p)} y(t) \geq x_i(p), \forall p, i = \overline{1,4}$

(2) $N(s_i - \epsilon) \leq \sum_j x_i(p_j) \leq N(s_i + \epsilon), i = \overline{1,4}$

(3) $\sum_i x_i(p) \leq 1, \forall p$

(4) $\sum_{t_k \in J_l} y(t) \geq 1, \forall J_l$

where the boolean variables are:

$y(t_k) =$	1 if candidate transcript t_k is selected, and 0 otherwise;
$x_i(p_j) =$	1 if the read p_j is mapped between $i-1$ and i standard deviations, and 0 otherwise;

The IP objective is to minimize the number of candidate transcripts subject to the constraints (1) through (4).

Constraint (1) implies that for each paired-end read $p \in n(s_i)$, at least one transcript $t \in T_i(p_j)$ is selected. Constraint (2) restricts the the number of paired-end reads mapped within every category of standard deviation. Constraint (3) ensures that each paired-end read p_j is mapped no more than with one category of standard deviation. Finally, constraint (4) requires that every splice junction to be present in the set of selected transcripts at least once.

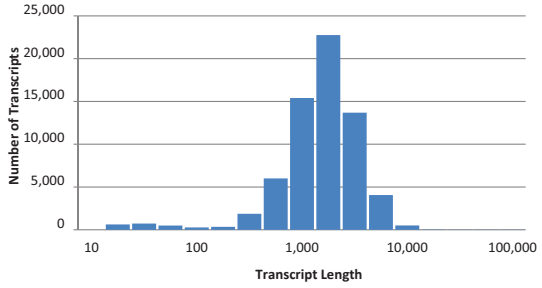
3. EXPERIMENTAL RESULTS

3.1 Simulation Setup and Matching Criteria

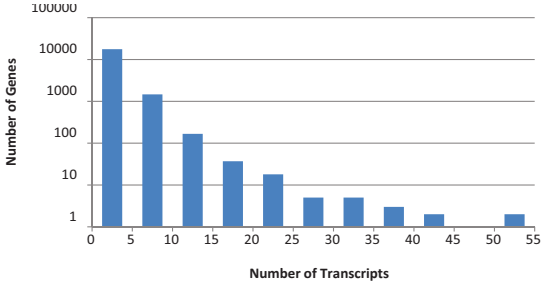
Simulation Setup. We tested TRIP on simulated human RNA-Seq data. The human genome sequence (hg18, NCBI build 36) was downloaded from UCSC together with the the KnownGenes transcripts annotation table. Genes were defined as clusters of known transcripts defined by the GNFAtlas2 table. The dataset contains a total of 66803 transcripts pertaining to 19372 genes. The transcripts length distribution is shown in Figure 3(a) and the number of transcripts per genes is shown in Figure 3.

Error-free paired-end reads of length 100 base pairs were randomly generated per gene by sampling fragments from known transcripts at coverage $100X$ per transcript. Expression levels of transcripts inside gene cluster follows uniform distribution. To address library preparation process for RNA-Seq experiment we simulate fragment lengths from a normal probability distribution with a mean of 500 and standard deviation 50 and 500.

We also include in the comparison variants of our methods that are given the transcription start sites (TSS) and transcription end sites (TES) to assess the benefits of complementing RNA-Seq data with TSS/TES data generated by specialized protocols such as the PolyA-Seq protocol in [3].



(a)



(b)

Figure 3: Distribution of transcript lengths (a) and gene cluster sizes (b) in the UCSC dataset

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 [21] and [9].

Following [1], we use *Sensitivity*, *Positive Predictive Value (PPV)* and *F-Score* to evaluate the performance of different methods. Sensitivity is defined as the proportion of annotated transcript sequences that match reconstructed sequences, i.e.,

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

PPV is defined the proportion of reconstructed sequences that match annotated transcript sequences, i.e.,

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

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

$$F\text{-}Score = 2 \times \frac{PPV \times Sens}{PPV + Sens}$$

3.2 Comparison of Methods on Simulated Data

In this section, we use the sensitivity, PPV, and F-score defined above to compare the TRIP method to the most recent version of Cufflinks (version 2.0.0 downloaded from website: http://cufflinks.cbcb.umd.edu/). We run Cufflinks with the following options: -m (the expected (mean) fragment length) and -s (the standard deviation for the distribution on fragment lengths). For this study, comparison with IsoLasso [9] was omitted. Due to technical problems, results were consistently incomparable to other methods. The integer program for TRIP is solved by IBM ILOG CPLEX (version 12.2.0.0). We also add a method that reports all candidate transcripts in order to illustrate the effectiveness of selection produced by the integer program (IP) in TRIP. It is also very important how much information is used when candidate transcripts are identified.

If annotated alternative transcription start sites (TSS) and transcription end sites (TES) can be used (these can be computationally inferred using read statistics and motifs or generated by specialized protocols such as the PolyA-Seq protocol in [3]) then the candidate transcript set is more accurate and the resulted method is referred as TRIP with TSS/TES. Otherwise, when TRIP does not rely on this information, the method is referred as TRIP.

Figures 4(a)-4(c) compare the performance of 4 methods (Cufflinks, Candidate Transcripts, TRIP with and without TSS/TES) on simulated data with respect to number of transcripts per gene. Note that sensitivity (see Fig. 4(a)) for single-transcript genes is 100% for all methods and with the growth in number of transcripts per gene, TRIP's sensitivity gradually improves over Cufflinks while sensitivity of Candidate Transcripts stays almost 100%. The advantage of TRIP over Cufflinks can be explained by extra statistical constraints in the IP that are not taken into account by Cufflinks.

Fig. 4(b) shows that Cufflinks has an advantage over TRIP in the portion of correctly predicted transcripts but overall comparison using F-score (see Fig. 4(c)) shows that TRIP improves over Cufflinks.

3.3 Influence of Sequencing Parameters

Although high-throughput technologies allow users to make trade-offs between read length and the number of generated reads, very little has been done to determine optimal parameters for fragment length. Additionally, novel Next Generation Sequencing (NGS) technologies such as Ion Torrent may allow to learn exact fragment length. For the case when fragment length is known, we have modified TRIP's IP referring to this new method as TRIP-L.

In this section we compare methods TRIP-L, TRIP and Cufflinks for the mean fragment length 500bp and variance of either 50bp or 500bp, to check how the variance affects the prediction quality. Figures 5(a)-5(c) compare sensitivity,

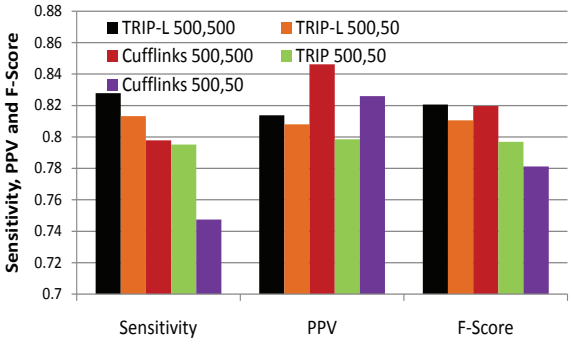


Figure 6: Overall Sensitivity, PPV and F-Score on simulated dataset with different sequencing parameters and distribution assumptions.

PPV and F-score of five methods (TRIP-L 500,500; TRIP-L 500,50; TRIP 500,50; Cufflinks 500,500; Cufflinks 500,50) on simulated data. The results show that as before TRIP has a better sensitivity and F-score while TRIP-L further improves them. Also higher variation in fragment length actually improves performance of all methods.

3.4 Results on Real RNA-Seq Data

We tested TRIP 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 gene was picked and validated experimentally due to interest in its biological function. We plan to have experimental validation at a larger scale in the future. The read alignments falling within the genomic locus of the selected gene were used to construct a splicing graph; then candidate transcripts were selected using TRIP. The dataset used consists of 46906 alignments for 22346 read pairs with read length of 68. TRIP was able to infer 5 out of 10 transcripts that we confirmed using qPCR. For comparison, we ran the same experiment using cufflinks, and it was able to infer 3 out of 10.

4. CONCLUSIONS

In this paper we introduce a novel method for transcriptome reconstruction from paired-end RNA-Seq reads based on Integer Programming. Our method critically exploits the distribution of fragment lengths, and can take advantage of additional experimental data such as TSS/TES and individual fragment lengths estimated, e.g., from ION Torrent [19] flowgram data. Preliminary experimental results on both real and synthetic datasets generated with various sequencing parameters and distribution assumptions show that our IP approach is scalable and has increased transcriptome reconstruction accuracy compared to previous methods that ignore information about fragment length distribution.

5. ACKNOWLEDGMENTS

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.

6. ADDITIONAL AUTHORS

Ion Mandoiu (Department of Computer Science & Engineering, University of Connecticut, email: `ion@engr.uconn.edu`) and Alex Zelikovsky (Department of Computer Science, Georgia State University, email: `alexz@cs.gsu.edu`).

7. REFERENCES

[1] I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. Mandoiu, P. Balfe, and A. Zelikovsky. Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics*, 12(Suppl 6):S1, 2011.

[2] 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.

[3] 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*, 22(6):1173–1183, 2012.

[4] J. Feng, W. Li, and T. Jiang. Inference of isoforms from short sequence reads. In *Proc. RECOMB*, pages 138–157, 2010.

[5] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6):469–477, May 2011.

[6] M. Grabherr. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, 2011.

[7] 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*, 28(5):503–510, 2010.

[8] B. Li, V. Ruotti, R. Stewart, J. Thomson, and C. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, 2010.

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

[10] 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.

[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*, pages 118 –123, nov. 2011.

[12] 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*, 30(1):99–104, 2012.

[13] A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 2008.

[14] M. Nicolae, S. Mangul, I. Mandoiu, and A. Zelikovsky. Estimation of alternative splicing isoform frequencies from rna-seq data. *Algorithms for Molecular Biology*, 6:9, 2011.

[15] 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.

[16] P. A. Pevzner. 1-Tuple DNA sequencing: computer analysis. *J Biomol Struct Dyn*, 7(1):63–73, Aug. 1989.

[17] A. Roberts, H. Pimentel, C. Trapnell, and L. Pachter. Identification of novel transcripts in annotated genomes using rna-seq. *Bioinformatics*, 2011.

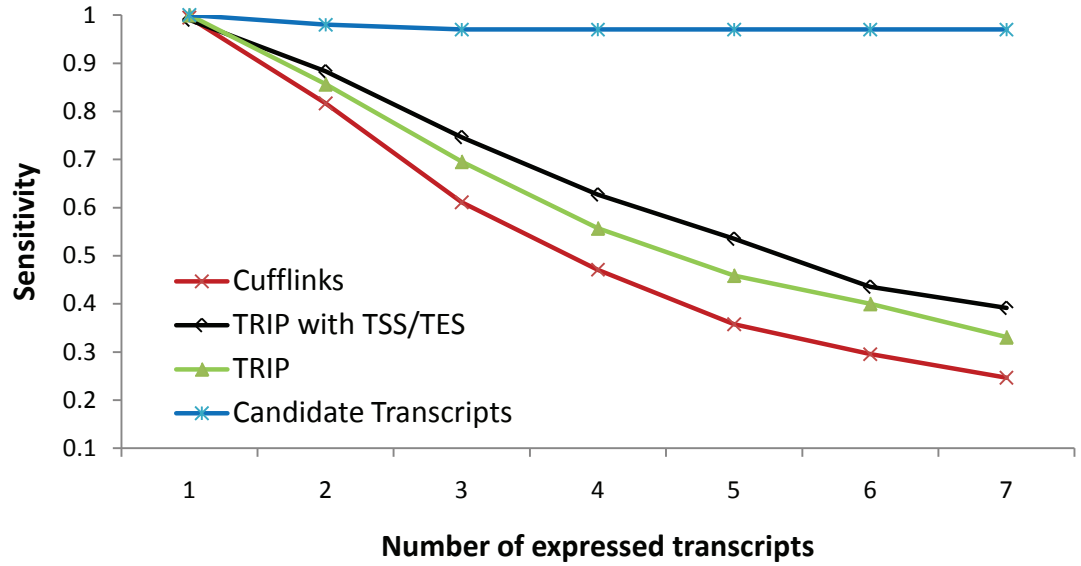
[18] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, and et al. De novo assembly and analysis of rna-seq data. *Nature Methods*, 7(11):909–912, 2010.

[19] J. M. Rothberg, W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, J. Hoon, J. F. Simons, D. Marran, J. W. Myers, J. F. Davidson, A. Branting, J. R. Nobile, B. P. Puc, D. Light, T. A. Clark, M. Huber, J. T. Branciforte, I. B. Stoner, S. E. Cawley, M. Lyons, Y. Fu, N. Homer, M. Sedova, X. Miao, B. Reed, J. Sabina, E. Feierstein, M. Schorn, M. Alanjary, E. Dimalanta, D. Dressman, R. Kasinskas, T. Sokolsky, J. A. Fidanza, E. Namsaraev, K. J. McKernan, A. Williams, G. T. Roth, and J. Bustillo. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011.

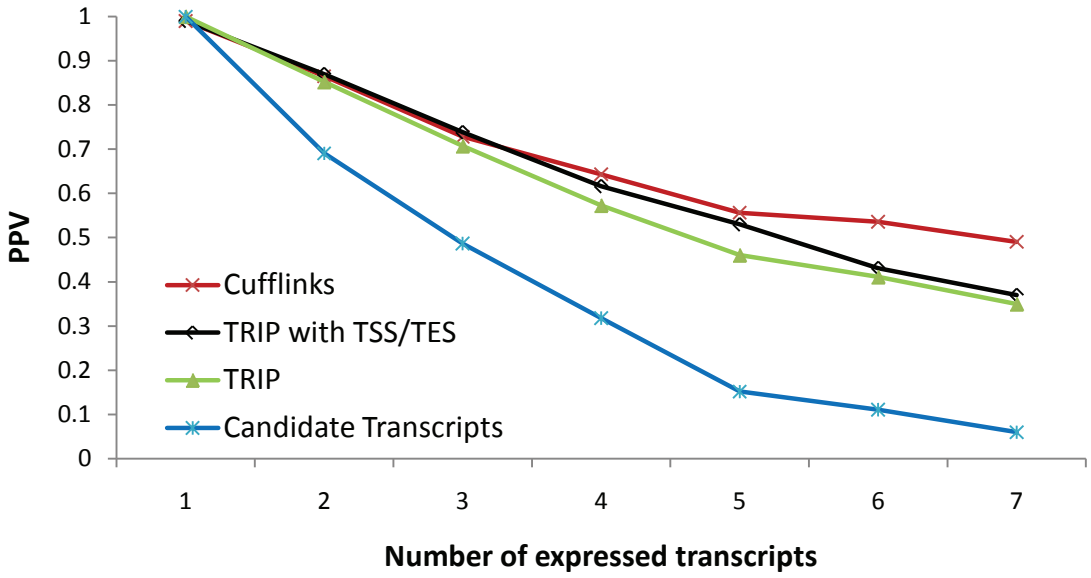
[20] C. Trapnell, L. Pachter, and S. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[21] 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*, 28(5):511–515, 2010.

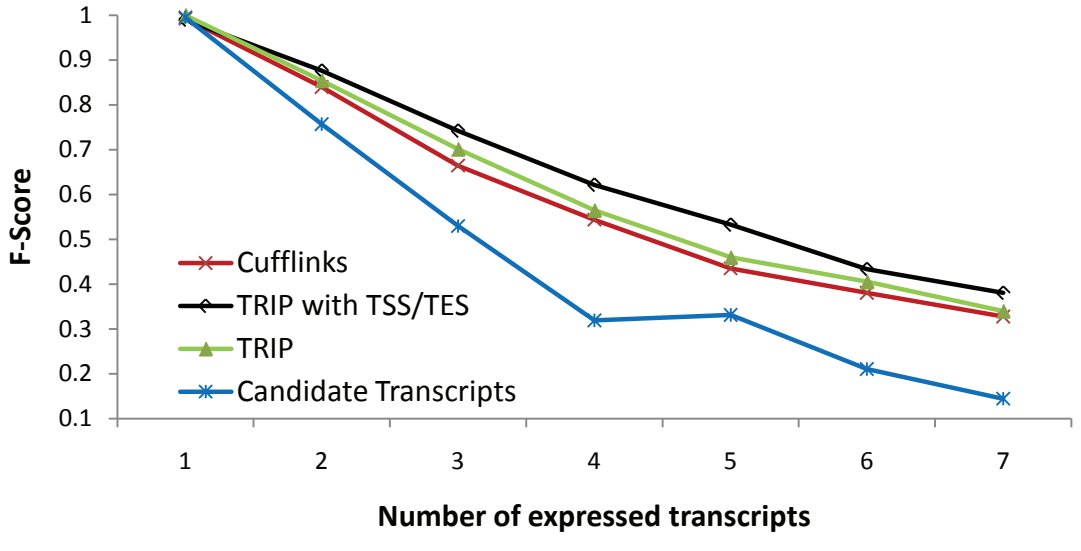
[22] E. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. Kingsmore, G. Schroth, and C. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.



(a)

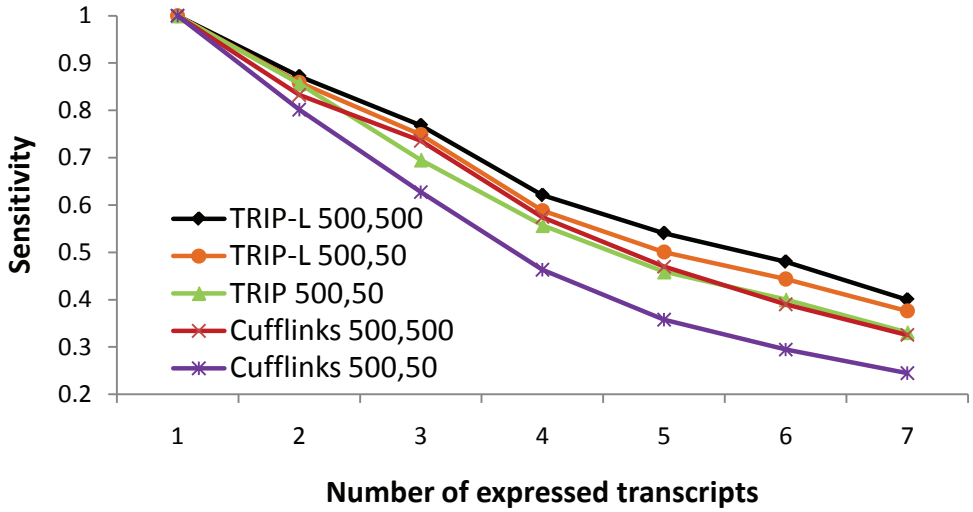


(b)

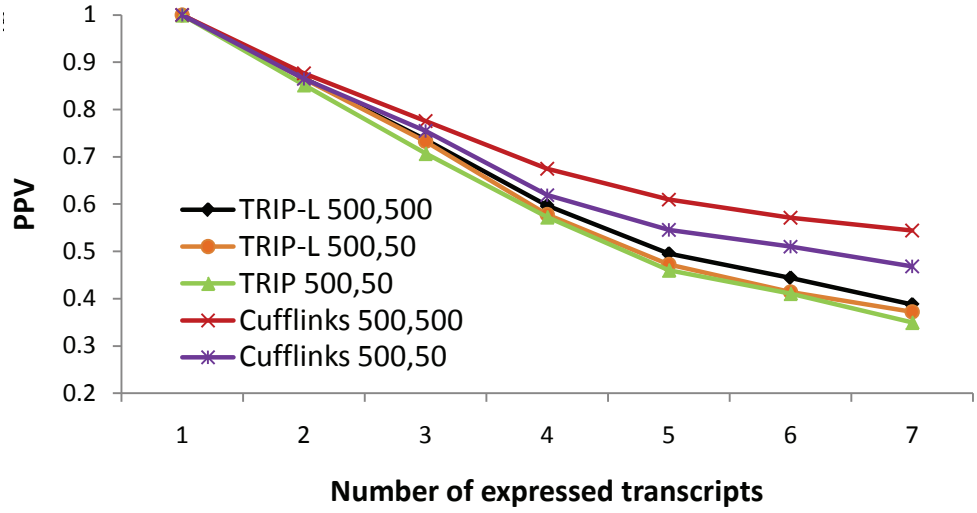


(c)

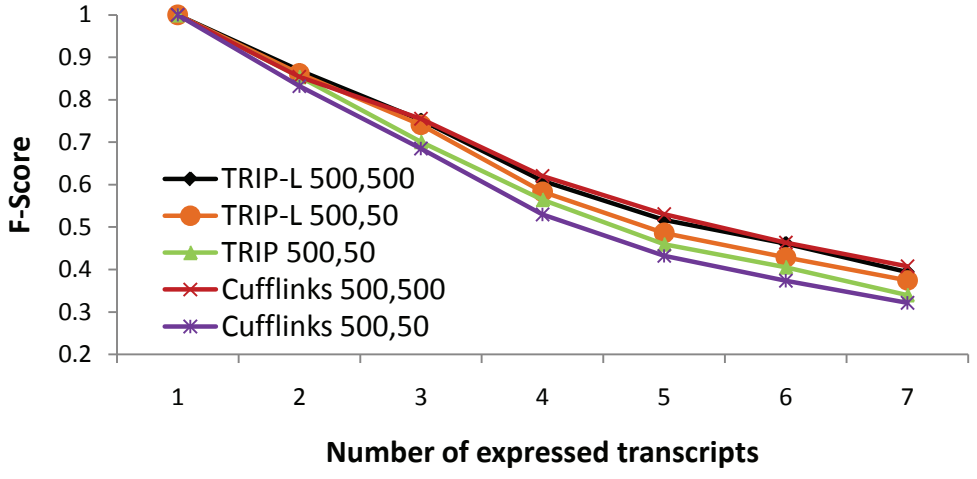
Figure 4: Comparison between methods for groups of genes with n transcripts (n=1,...,>7) on simulated dataset with mean fragment length 500, standard deviation 50 and read length of 100x2: (a) Sensitivity (b) Positive Predictive Value (PPV) and (c) F-Score.



(a)



(b)



(c)

Figure 5: Comparison between methods for groups of genes with n transcripts (n=1,...,>7) on simulated dataset with different sequencing parameters and distribution assumptions: (a) Sensitivity (b) Positive Predictive Value (PPV) and (c) F-Score.