Fall 11-16-2012

Algorithms for Transcriptome Quantification and Reconstruction from RNA-Seq Data

Serghei Mangul

ALGORITHMS FOR TRANSCRIPTOME QUANTIFICATION AND RECONSTRUCTION FROM RNA-SEQ DATA

by

SERGHEI MANGUL

Under the Direction of Dr. Alexander Zelikovsky

ABSTRACT

Massively parallel whole transcriptome sequencing and its ability to generate full transcriptome data at the single transcript level provides a powerful tool with multiple interrelated applications, including transcriptome reconstruction, gene/isoform expression estimation, also known as transcriptome quantification. As a result, whole transcriptome sequencing has become the technology of choice for performing transcriptome analysis, rapidly replacing array-based technologies. The most commonly used transcriptome sequencing protocol, referred to as RNA-Seq, generates short (single or paired) sequencing tags from the ends of randomly generated cDNA fragments. RNA-Seq protocol reduces the sequencing cost and significantly increases data throughput, but is computationally challenging to reconstruct full-length transcripts and accurately estimate their abundances across all cell types.

We focus on two main problems in transcriptome data analysis, namely, transcriptome reconstruction and quantification. Transcriptome reconstruction, also referred to as novel isoform discovery, is the problem of reconstructing the transcript sequences from the sequencing data. Reconstruction can be done de novo or it can be assisted by existing genome and transcriptome annotations. Transcriptome quantification refers to the problem of estimating the expression level of each transcript. We present a genome-guided and annotation-guided transcriptome reconstruction methods as well as methods for transcript and gene expression level estimation. Empirical results on both synthetic and real RNA-seq datasets show that the proposed methods improve transcriptome quantification and reconstruction accuracy compared to previous methods.

INDEX WORDS: Algorithm, transcriptome reconstruction, transcriptome quantification, alternative splicing, RNA-Seq, assembly, isoform expression level, gene expression level, splicing graph, integer programming, expectation maximization, fragment length distribution

ALGORITHMS FOR TRANSCRIPTOME QUANTIFICATION AND RECONSTRUCTION FROM RNA-SEQ DATA

by

SERGHEI MANGUL

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
in the College of Arts and Sciences
Georgia State University
2012

ALGORITHMS FOR TRANSCRIPTOME QUANTIFICATION AND RECONSTRUCTION FROM RNA-SEQ DATA

by

SERGHEI MANGUL

Committee Chair:	Dr. Alexander Zelikovsky
Committee:	Dr. Yi Pan
	Dr. Robert Harrison
	Dr. Ion Mandoiu

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
December 2012

DEDICATION

I dedicate this dissertation in loving memory of my father Ilia D. Mangul (August 2, 1950 - August 21, 2010). I know how important it was for you to see me graduate. It was you who inspired me to get my PhD. As a young boy, I remember attending your PhD defense. I still have the copy of your thesis that you signed for me. What you wrote there has become a motto for my life: ”Aut inveniam viam aut faciam”.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Alex Zelikovsky for his encouragement and constant support over the past four years of my graduate studies at Georgia State University. This thesis could not been done without his guidelines, patience and motivation. For helping to guide this research and provide exceptional feedback and encouragement, I thank my graduate committee members: Dr. Yi Pan, Dr. Robert Harrison and Dr. Ion Mandoiu. Special thanks to Dr. Ion Mandoiu for sharing his knowledge and giving helpful advice. I want to express my gratitude to Life Technologies team in Foster City, CA for discovering exciting world of biotechnology for me. Life Technologies internship changed my life and opened for me great opportunities. I want to say special thanks to Dumitru Brinza and Fiona Hyland and to all my friends from Life Technologies.

Special thanks to Computer Science Department of Georgia State University. I express my special gratitude to Dr. Raj Sunderraman , Mrs. Tammie Dudley and Mr. Shaochieh Ou. I am also grateful to Moldova State University, which I graduate with my Bachelor in Applied Mathematics. Special thanks to Dr. D. Lozovanu and Dr. I. Secrieru.

Thanks to my friends and colleagues Abi, Adrian, Bassam, Blanche, Chad, Cristina, Dinesh, Gulsah, Hamed, Helen, Irina, James, Kelly, Lakshmi, Matt, Marco, Ming, Nick, Olga, Rudy, Sahar, Vanessa. Thanks also to my new friends from all over the world for all what I have learned. Thanks to my friends back home who always supported me.

I am most grateful to my parents Ilya Mangul and Nelly Jardan, whose love was my most important source of strength and determination. I am also grateful to my grandfather Vasilie Jardan, grandmother Alexandra Jardan, aunt Rita, and aunt Natasha. I would like to say very special thanks to my Godfather and an amazing friend Michael Robert Sawyer.

Finally, special recognition goes out to my lovely wife Zoia, for her support, encouragement and patience during my academic journey over the past four years. Her support and encouragement made this dissertation possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- NGS - Next Generation Sequencing
- GE - Gene Expression Level Estimation
- IE - Isoform Expression Level Estimation

PART 1

INTRODUCTION

Massively parallel whole transcriptome sequencing and its ability to generate full transcriptome data at the single transcript level provides a powerful tool with multiple interrelated applications, including transcriptome reconstruction ([3], [4], [5], [6]), gene/isoform expression estimation ([7], [8], [5], [9], also known as transcriptome quantification, studying trans- and cis-regulatory effect [10], studying parent-of origin effect [10], [11], [12], and calling expressed variants ([13]). As a result, whole transcriptome sequencing has become the technology of choice for performing transcriptome analysis, rapidly replacing array-based technologies ([14]).

The most commonly used transcriptome sequencing protocol, referred to as RNA-Seq, generates short (single or paired) sequencing tags from the ends of randomly generated cDNA fragments. Using transcriptome sequencing data, most current research employs methods that depend on existing transcriptome annotations. Unfortunately, as shown by recent studies ([15]), 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 data critically requires knowledge of expressed transcript sequences ([7], [8], [16]. [9]. Another challenge in transcriptomic analysis comes from the ambiguities in read/tag mapping to the reference. My dissertation research focuses on two main problems in transcriptome data analysis, namely, transcriptome reconstruction and quantification, and we show how these challenges are handled. Transcriptome reconstruction, also referred to as novel isoform discovery, is the problem of reconstructing the transcript sequences from the sequencing data. Reconstruction can be done de novo or it can be assisted by existing genome and transcriptome annotations. Transcriptome quantification refers to the problem of estimating the expression level of each transcript.

1.1 RNA-Seq protocol

History of DNA sequencing is rich and diverse. The majority of DNA protocols relied on Sanger capillary-based semi-automated sequencing technology. Sanger biochemistry allows to achieve up to 1,000 bp read length, and per-base “raw” accuracies as high as 99.999%. Due to high accuracy, genomes sequenced by Sanger technology currently are used in modern databases.

Second-generation of DNA sequencing technologies are more parallelizable and have higher throughput compared to Sanger protocol. These technologies are collectively called Next Generation Sequencing (NGS). Many NGS technologies have been realised as a commercial product (e.g., the Illumina HiSeq Systems (marketed by Illumina, San Diego, CA, USA), the SOLiD Systems (marketed by Applied Biosystems by Life Technologies; San Diego, CA, USA), 454 Genome Sequencers (Roche Applied Science; Penzberg, Upper Bavaria, Germany), the HeliScope Single Molecule Sequencer technology (Helicos; Cambridge, MA, USA), Ion Personal Genome Machine Sequencer(marketed by Ion Torrent by Life Technologies, San Diego, CA, USA). These technologies produce reads of length 50 - 500bp and up to 600 Gb of throughput.

Recent advances in DNA sequencing have made it possible to sequence the whole transcriptome by massively parallel sequencing, commonly referred as RNA-Seq [7]. RNA-Seq is quickly becoming the technology of choice for transcriptome research and analyses [14]. RNA-Seq allows reduction of the sequencing cost and significantly increases data throughput, but it is computationally challenging to use such RNA-Seq data for reconstructing of full length transcripts and accurately estimate their abundances across all cell types.

RNA-Seq, uses next generation sequencing technologies, such as SOLiD ([17]), 454 ([18]), Illumina ([19]), or Ion Torrent ([20]). Figure 1.1 depicts the steps in an RNA-Sequencing experiment, including the first step of analysis which is typically mapping the data to a reference. After extracting the RNA sample, it is converted to cDNA fragments. The distribution of the fragment lengths is determined during the RNA-Seq experiment and can

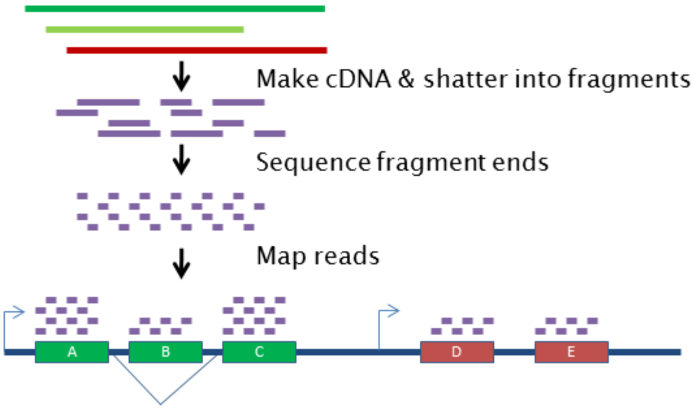


Figure 1.1 A schematic representation of the RNA-Seq protocol.

be useful in downstream analysis. This is usually followed by an amplification step; then one or both ends of the cDNA fragments are sequenced producing either single or paired-end reads. Sequencing can be either directional, meaning that all reads come from the coding strand for single reads. For paired end read, directional sequencing implied that the first read in the pair comes from the coding strand, while the second comes from the non-coding strands. This strand specificity is not maintained in non-directional sequencing. The specifics of the sequencing protocols vary from one technology to the other. Similarly, the length of produced reads varies depending on the technology with newer technologies producing longer reads.

1.2 Applications of RNA-Seq

Ubiquitous regulatory mechanisms such as the use of alternative transcription start and polyadenylation sites, alternative splicing, and RNA editing result in multiple messenger RNA (mRNA) isoforms being generated from a single genomic locus. Most prevalently, alternative splicing is estimated to take place for over 90% of the multi-exon human genes across diverse cell types [8], with as much as 68% of multi-exon genes expressing multiple isoforms in a clonal cell line of colorectal cancer origin [21]. Not surprisingly, the ability to reconstruct full length transcript sequences and accurately estimate their expression levels is widely believed to be critical for unraveling gene functions and transcription regulation mechanisms [22].

The common applications of RNA-seq are gene expression level estimation, isoform expression level estimation, novel transcript discovery, and transcriptome reconstruction. A variety of new methods and tools have been recently developed to tackle these problems.

Estimating transcript and gene expression levels has long been an important application for RNA-Seq analyses. Estimation of isoform expression level is not a trivial task .There is yet no standard protocol for measuring isoforms abundances from RNA-Seq data. The key challenge in transcriptome quantification is accurate assignment of ambiguous reads to isoforms. Most RNA-Seq studies to date still ignore alternative splicing or, similar to splicing array studies, restrict themselves to surveying the expression levels of exons and exon-exon junctions. The main difficulty in inferring expression levels for full-length transcripts lies in the fact that current sequencing technologies generate short reads (from few tens to hundreds of bases), many of which cannot be unambiguously assigned to individual transcripts.

Inferring expression at isoform level provides information for finer-resolution biological studies, and also leads to more accurate estimates of expression at the gene level by allowing rigorous length normalization. Genome-wide gene expression level estimates derived from isoform level estimates are significantly more accurate than those obtained directly from RNA-Seq data using isoform-oblivious GE methods such as the widely used counting of unique reads, the rescue method of [7], or the EM algorithm of [23].

Identifying of all transcripts expressed in a particular sample require the assembly of reads into transcription units. This process is collectively called transcriptome reconstruction. 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 [24]. Genome-independent methods such as Trinity [25] or transAbyss [26] directly assemble reads into transcripts. A commonly used approach for such methods is de Brujin graph [27] 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 [28, 29] 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 [30] or SpliceMap [31]. 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 [4] 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. [3, 32], 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. Cufflinks and Scripture do not target the quantification accuracy. IsoLasso [5] uses the LASSO [33] 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 [34] 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.

Table 3.1 includes classification of the available methods for genome-guided transcriptome reconstruction based on supported parameters and underlying algorithms.

Table 1.1 Classification of transcriptome reconstruction methods

Method	Support paired-end reads	Consider fragment lenght distribution	Require annotation
TRIP	Yes	Yes	No
IsoLasso	Yes	No	No
IsoInfer	No	No	TES/TSS
Cufflinks	Yes	Yes	No
CLIQ	No	No	No
Scripture	Yes	No	No
SLIDE	Yes	No	gene/exon boundaries

1.3 Contributions and Future Work

We present a general framework that includes the genome-guided and annotation-guided transcriptome reconstruction methods as well as methods for transcript and gene expression level estimation.

We propose a novel expectation-maximization algorithm to solve the problem of transcript and gene expression level estimation from RNA-Seq data. Our algorithm, referred to as IsoEM [9], is based on disambiguating of information provided by the distribution of insert sizes generated during sequencing library preparation, and takes advantage of base quality scores, strand and read pairing information when available. To solve the problem of transcriptome quantification in the context of partially annotated genomes we propose enhancement of EM algorithm, "**V**irtual **T**ranscript **E**xpectation **M**aximization(VTEM)" [35]. VTEM detects overexpressed reads and/or exons corresponding to the unannotated transcripts and estimates annotated transcript frequencies.

To address the problem of transcriptome reconstruction we suggest genome-guided and annotation-guided methods. We present a novel annotation-guided method for transcriptome discovery and reconstruction in partially annotated genomes and compare it with existing annotation-guided and genome-guided transcriptome assembly methods. Our method, referred as "**D**iscovery and **R**econstruction of **U**nannotated **T**ranscripts" (DRUT) [36], can be used to enhance existing transcriptome assemblers, such as Cufflinks [3]. It was shown that

Cufflinks enhanced by DRUT has superior quality of reconstruction and frequency estimation of transcripts.

To solve transcitome reconstruction problem assisted by existing genome annotations we propose a novel method called "**T**ransciptome **R**econstruction using **I**nteger **P**rograming" (TRIP [6]). The method incorporates information about fragment length distribution of RNA-Seq paired-end reads to reconstruct novel transcripts. The first step is to infer exon boundaries from spliced genome alignments of the reads. Then, create a splice graph based on inferred exon boundaries. Third step enumerates 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.

Recent advances in sequencing technologies made it possible to produce longer single-end reads with the length comparable to length of fragment for paired-end technology[20] . Novel method was developed to address transcriptome reconstruction problem from single RNA-Seq reads. This method, called " Maximum Likelihood Integer Programming " (MLIP), aims 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 isoform expression estimation.

Empirical results on both synthetic and real RNA-seq datasets show that the proposed methods improve transcriptome quantification and reconstruction accuracy compared to previous methods.

In ongoing work we are exploring possibility of integrating transcriptome quantification and transcriptome reconstruction that will possibly lead to quantification based reconstruction method. Currently, Next Generation Sequencing technologies allow to run library preparation step multiple times varying the fragment length distribution for every step. Additionally, it is possible to perform read barcoding for every library preparation step, which will produce reads with different fragment lengths. To take adventure of this technology we plan to develop the method able to handle reads from multiple libraries. We expect to improve reconstruction accuracy by integrating different fragment length distributions into transcriptome reconstruction algorithm. Also we are planning to release software tool for transcriptome quantification and reconstruction that will include all our methods.

1.4 Organization

Dissertation is organized as follows. Chapter 1 gives a brief description of the RNA-Seq technology and discuss application of this technology for transcriptome quantification and reconstruction problems. Chapter 2 presents the transcriptome quantification problem and motivation behind it. Then two approaches are described: first approach for completely annotated genomes and second one for partially annotated genomes. We finalize this chapter with application of our method to human RNA-Seq data.

Chapter 3 introduces transcriptome reconstruction problem and gives classification of existing methods. Transcriptome reconstruction, also referred to as novel isoform discovery, can be done de novo or it can be assisted by existing genome and transcriptome annotations. We present algorithms for reconstruction of organisms transcriptome from RNA-Seq data assisted by existing genome and transcriptome annotations. Discussion and future directions are provided in the Chapter 4.

1.5 Software Packages

- **IsoEM** - Inferring Alternative Splicing Isoform Frequencies from High-Throughput RNA-Seq Data $http : //dna.engr.uconn.edu/?page_id = 105$
- **VSEM** - Inferring Unannotated Haplotypes Frequencies in Partially Annotated Genomes. Enhacement Tool for IsoEM and ViSpA. $http : //www.cs.gsu.edu/\ serghei/?q = vsem$
- **DRUT** - Discovery and Reconstruction of Unannotated Transcripts in Partially Annotated Genomes from High-Throughput RNA-Seq Data. $http : //www.cs.gsu.edu/\ serghei/?q = drut$

- **TRIP** - Novel Transcript Reconstruction from Paired-End RNA-Seq Reads.
  http : //www.cs.gsu.edu/ serghei/?q = trip

1.6 Related Publications

Refereed Journal Articles and Book Chapters

- S. Al Seesi, **S. Mangul**, A. Caciula, I. Mandoiu and A. Zelikovsky ”Transcriptome reconstruction and quantification from RNA sequencing data”(to appear), **book chapter**, *Genome Analysis: Current Procedures and Applications*, 2013
- **S. Mangul**, A. Caciula, O. Glebova, I. Mandoiu and A. Zelikovsky, ”Improved Transcriptome Quantication and Reconstruction from RNA-Seq Reads using Partial Annotations”(to appear), *In Silico Biology(ISB) : An International Journal on Computational Molecular Biology*, 2012
- **S. Mangul**, A. Caciula, I. Mandoiu and A. Zelikovsky ”RNA-Seq based transcriptome quantification and reconstruction guided by protein coding gene annotation”(to appear), **book chapter**, *Algorithmic and AI Methods for Protein Bioinformatics*, 2012
- M. Nicolae, **S. Mangul**, I. Mandoiu and A. Zelikovsky, ”Estimation of alternative splicing isoform frequencies from RNA-Seq data”, *Algorithms for Molecular Biology*, 2011
- I. Astrovskaya, B. Tork, **S. Mangul**, K. Westbrooks, I. Mandoiu, P. Balfe and A. Zelikovsky, ”Inferring Viral Spectrum from 454 Pyrosequencing Reads”, *BMC Bioinformatics*, 2011

Refereed Conference Articles

- **S. Mangul**, A. Caciula, S. Al Seesi, D. Brinza, A. Banday, R. Kanadia, I. Mandoiu and A. Zelikovsky, ”Flexible Approach for Novel Transcript Reconstruction from RNA-Seq Data using Maximum Likelihood Integer Programming”(submitted), *Proc. 5th International Conference on Bioinformatics and Computational Biology (BICoB 2013)*
- **S. Mangul**, A. Caciula, S. Al Seesi, D. Brinza, A. Banday, R. Kanadia, I. Mandoiu and A. Zelikovsky, ”An Integer Programming Approach to Novel Transcript Reconstruction from Paired-End RNA-Seq Reads”, *Proc. 3rd ACM Conference on Bioinformatics,*

Computational Biology and Biomedicine (ACM-BCB 2012)

- **S. Mangul**, A. Caciula, I. Mandoiu and A. Zelikovsky, ”Novel Transcript Reconstruction from Paired-End RNA-Seq Reads Using Fragment Length Distribution”, *Proc. 2nd IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS 2012)*
- **S. Mangul**, A. Caciula, I. Mandoiu and A. Zelikovsky, ”RNA-Seq based discovery and reconstruction of unannotated transcripts in partially annotated genomes”, *Proc. of Workshop on Computational Advances in Molecular Epidemiology (CAME 2011)*
- **S. Mangul**, I. Astrovskaya, M. Nicolae, B. Tork, I. Mandoiu and A. Zelikovsky, ”Maximum Likelihood Estimation of Incomplete Genomic Spectrum from HTS Data”, *Proc. 11th Workshop on Algorithms in Bioinformatics (WABI 2011)*, Lecture Notes in Bioinformatics, pp.
- M. Nicolae, **S. Mangul**, I. Mandoiu and A. Zelikovsky, ”Estimation of Alternative Splicing isoform Frequencies from RNA-Seq Data”, *Proc. 10th Workshop on Algorithms in Bioinformatics (WABI 2010)*, Lecture Notes in Bioinformatics 6293, pp. 202-214

Posters and Presentations

- **S. Mangul**, A. Caciula, S. Al Seesi, D. Brinza, I. Mandoiu and A. Zelikovsky, ”TRIP : A Method for Novel Transcript Reconstruction from Paired-End RNA-Seq Reads”, *Poster at 11th European Conference on Computational Biology(ECCB 2012)*, Basel, Switzerland, **travel award**
- **S. Mangul**, A. Caciula, S. Al Seesi, O. Sakarya, I. Mandoiu and A. Zelikovsky, ”An Integer Programming Approach to Novel Transcript Reconstruction from Paired-End RNA-Seq Reads”, *Poster at 20th Annual International Conference on Intelligent Systems for Molecular Biology(ISMB 2012)*, Long Beach, CA, **travel award**
- **S. Mangul**, ”An Integer Programming Approach to Novel Transcript Reconstruction from Paired-End RNA-Seq Reads”, **invited talk**, *Student Council Symposium 2012/ISMB 2012*, Long Beach, CA
- **S. Mangul**, A. Caciula, I. Mandoiu and A. Zelikovsky, ”An Integer Programming Ap-

proach to Novel Transcript Reconstruction from Paired-End RNA-Seq Reads", *Poster at 8th International Symposium on Bioinformatics Research and Applications (ISBRA 2012)*, Dallas, TX, **best poster award**

- **S. Mangul**, "Computational Methods for Transcriptome Reconstruction and Quantification using RNA-seq", **public lecture**, *Center of Molecular Biology, University of Academy of Sciences of Moldova(2012)*, Chisinau, Moldova
- **S. Mangul**, "Mathematical and Computational Approaches in High- Throughput Genomics", **public lecture**, *Center of Molecular Biology, University of Academy of Sciences of Moldova(2012)*, Chisinau, Moldova
- **S. Mangul**, "Its a DNA World: An introduction to Next Generation Sequencing", **public lecture**, *Center of Molecular Biology, University of Academy of Sciences of Moldova(2012)*, Chisinau, Moldova
- **S. Mangul**, A. Caciula, I. Mandoiu and A. Zelikovsky, "An Integer Programming Approach to Novel Transcript Reconstruction from Paired-End RNA-Seq Reads", *Poster at Workshop on Biostatistics and Bioinformatics, Department of Mathematics and Statistics, Georgia State University (2012)*, Atlanta, GA
- **S. Mangul**, A. Caciula, N. Mancuso, I. Mandoiu and A. Zelikovsky, "An Integer Programming Approach to Novel Transcript Reconstruction from Paired-End RNA-Seq Reads", *Poster at 16th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2012)*, Barcelona, Spain
- **S. Mangul**, "Novel Transcript Reconstruction from Paired-End RNA-Seq Reads Using Fragment Length Distribution", **invited talk**, *2nd Workshop on Computational Advances for Next Generation Sequencing (CANGS 2012)*, Las Vegas, NV, US
- **S. Mangul**, "The Next, Next Generation Sequencing - From Semiconductor to Single Molecule", *Life Technologies UCLA Embassy Info Session(2011)*, UCLA ISPE, University of California, Los Angeles, CA
- **S. Mangul**, A. Caciula, I. Mandoiu and A. Zelikovsky, "RNA-Seq based novel transcripts identification in partially annotated genomes", *Poster at 8th International Conference on*

Bioinformatics "From Genomics to Synthetic Biology"(2011), Atlanta, GA

- **S. Mangul**, A. Caciula, I. Mandoiu and A. Zelikovsky, "RNA-Seq based discovery and reconstruction of unannotated transcripts in partially annotated genomes", *Poster at Workshop on Next-generation Sequencing Technology and Algorithms for Primary Data Analysis(2011)*, Institute for Pure and Applied Mathematics, University of California, Los Angeles, CA
- **S. Mangul**, I. Astrovskaya, M. Nicolae, B. Tork, I. Mandoiu and A. Zelikovsky, "Maximum Likelihood Estimation of Incomplete Genomic Spectrum from HTS Data", *Mathematical and Computational Approaches in High-Throughput Genomics(2011)*, Institute for Pure and Applied Mathematics, University of California, Los Angeles, CA
- **S. Mangul**, A. Caciula, I. Mandoiu and A. Zelikovsky, "RNA-Seq based discovery and reconstruction of unannotated transcripts in partially annotated genomes", *Poster at Workshop on Next-generation Sequencing Technology and Algorithms for Primary Data Analysis(2011)*, Institute for Pure and Applied Mathematics, University of California, Los Angeles, CA
- **S. Mangul**, D. Brinza, F. Hyland, "Orthogonal Error Correction Codes in Next Generation Sequencing", *Poster at Life Technologies Intern Poster Session(2011)*, Foster City, CA, USA, **best poster award**
- **S. Mangul**, I. Astrovskaya, M. Nicolae, B. Tork, I. Mandoiu and A. Zelikovsky (to appear), "Maximum Likelihood Estimation of Incomplete Genomic Spectrum from HTS Data", *Poster at Georgia Life Sciences Summit : Innovation for a Healthier World (2011*), Atlanta, GA, USA
- **S. Mangul**, I. Astrovskaya, B. Tork, I. Mandoiu and A. Zelikovsky (to appear), "Viral Quasispecies Reconstruction Based on Unassembled Frequency Estimation", *Poster at 7th International Symposium on Bioinformatics Research and Applications (ISBRA 2011)*, Changsha, China
- **S. Mangul** and A. Zelikovsky, "Haplotype Discovery Based on Unassembled Sequences Estimation", *Poster at First Annual RECOMB Satellite Workshop on Massively Parallel*

Sequencing (RECOMBseqCCB 2011), Vancouver, BC, Canada

- **S. Mangul** and A. Zelikovsky, ”Haplotype discovery from high-throughput sequencing data”, *Poster at 1st IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2011)*, Orlando, FL, USA
- I. Astrovskaya, B. Tork, **S. Mangul**, I. Mandoiu, P. Balfe and A. Zelikovsky, ”VISPA: Viral Spectrum Assembling Method”, *Poster at 1st IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2011)*, Orlando, FL, USA, **best poster award**
- M. Nicolae, **S. Mangul**, I. Mandoiu and A. Zelikovsky, ”Estimation of Alternative Splicing Isoform Frequencies From RNA-Seq Data”, *INFORMS Annual Meeting (2010)* , Austin, TX,
- I. Astrovskaya, B. Tork, **S. Mangul**, K. Westbrooks, I. Mandoiu, P. Balfe and A. Zelikovsky, ”HCV Quasispecies Spectrum Reconstruction from 454 Pyrosequencing Reads”, *Poster at Georgia Life Sciences Summit : Innovation for a Healthier World (2010)*, Atlanta, GA, USA
- **S. Mangul** and A. Zelikovsky, ”Haplotype discovery from RNA-Seq data”, *Poster at Georgia Life Sciences Summit : Innovation for a Healthier World (2010)*, Atlanta, GA, USA
- M. Nicolae, **S. Mangul**, I. Mandoiu and A. Zelikovsky, ”Estimation of alternative splicing isoform frequencies from RNA-Seq data”, *Dagstuhl seminar on Structure Discovery in Biology: Motifs, Networks and Phylogenies (2010)*, Dagstuhl, Germany
- M. Nicolae, **S. Mangul**, I.I. Mandoiu and A. Zelikovsky, ”Estimation of alternative splicing isoform frequencies from RNA-Seq data”, *Poster at 6th International Symposium on Bioinformatics Research and Applications (ISBRA 2010)*, Storrs, CT, USA
- **S. Mangul** and A. Zelikovsky, ”Haplotype spectrum reconstruction from sequencing reads”, *Poster at 6th International Symposium on Bioinformatics Research and Applications (ISBRA 2010)*, Storrs, CT, USA
- **S. Mangul** and A. Zelikovsky, ”MLE-based 2SNP Phasing”, *Poster at 5th International*

Symposium on Bioinformatics Research and Applications (ISBRA 2009), Fort Lauderdale, FL, USA

PART 2

TRANSCRIPTOME QUANTIFICATION

2.1 Introduction

Massively parallel whole transcriptome sequencing, commonly referred as RNA-Seq, is quickly becoming the technology of choice for gene expression profiling. However, due to the short read length delivered by sequencing technologies, estimation of expression levels for alternative splicing gene isoforms remains challenging.

2.1.1 Background

Ubiquitous regulatory mechanisms such as the use of alternative transcription start and polyadenylation sites, alternative splicing, and RNA editing result in multiple messenger RNA (mRNA) isoforms being generated from a single genomic locus. Most prevalently, alternative splicing is estimated to take place for over 90% of the multi-exon human genes across diverse cell types [8], with as much as 68% of multi-exon genes expressing multiple isoforms in a clonal cell line of colorectal cancer origin [21]. Not surprisingly, the ability to reconstruct full length isoform sequences and accurately estimate their expression levels is widely believed to be critical for unraveling gene functions and transcription regulation mechanisms [22].

Two key interrelated computational problems arise in the context of transcriptome quantification: *gene expression level estimation (GE)*, and *isoform expression level estimation (IE).* Targeted GE using methods such as quantitative PCR has long been a staple of genetic studies. The completion of the human genome has been a key enabler for genome-wide GE performed using expression microarrays. Since expression microarrays have limited capability of detecting alternative splicing events, specialized splicing arrays have been developed for genome-wide interrogation of both annotated exons and exon-exon junctions.

However, despite sophisticated deconvolution algorithms [37, 38], the fragmentary information provided by splicing arrays is typically insufficient for unambiguous identification of full-length transcripts [39, 40]. Massively parallel whole transcriptome sequencing, commonly referred to as RNA-Seq, is quickly replacing microarrays as the technology of choice for performing GE due to their wider dynamic range and digital quantitation capabilities [14]. Unfortunately, most RNA-Seq studies to date still ignore alternative splicing or, similar to splicing array studies, restrict themselves to surveying the expression levels of exons and exon-exon junctions. The main difficulty in inferring expression levels for full-length isoforms lies in the fact that current sequencing technologies generate short reads (from few tens to hundreds of bases), many of which cannot be unambiguously assigned to individual isoforms.

2.1.2 Previous Work

RNA-Seq analyses typically start by mapping sequencing reads onto the reference genome, transcript libraries, exon-exon junction libraries, or combinations thereof. Early RNA-Seq studies have recognized that limited read lengths result in a significant percentage of so called *multireads*, i.e., reads that map equally well at multiple locations in the genome. A simple (and still commonly used) approach is to discard multireads, and estimate expression levels using only the so called *unique* reads. Mortazavi et al. [7] proposed a multiread "rescue" method whereby initial gene expression levels are estimated from unique reads and used to fractionally allocate multireads, with final expression levels obtained by re-estimation based on total counts obtained after multiread allocation. An expectation-maximization (EM) algorithm that extends this scheme by repeatedly alternating between fractional read allocation and re-estimation of gene expression levels was recently proposed in [23].

A number of recent works have addressed the IE problem, namely isoform expression level estimation from RNA-Seq reads. Under a simplified "exact information" model, [40] showed that neither single nor paired read RNA-Seq data can theoretically guarantee unambiguous inference of isoform expression levels, although paired reads may be sufficient to deconvolute expression levels for the majority of annotated isoforms. The key challenge in IE is accurate assignment of ambiguous reads to isoforms. Compared to the GE context, read ambiguity is much more significant, since it affects not only multireads, but also reads that map at a unique genome location expressed in multiple isoforms. Estimating isoform expression levels based solely on unambiguous reads, as suggested, e.g., in [21], results in splicing-dependent biases similar to the transcript-length bias noted in [41], further complicating the design of unbiased differential expression tests based on RNA-Seq data. To overcome this difficulty, [42] proposed a Poisson model of single-read RNA-Seq data explicitly modeling isoform frequencies. Under their model, maximum likelihood estimates are obtained by solving a convex optimization problem, and uncertainty of estimates is obtained by importance sampling from the posterior distribution. Li et al. [43] introduced an expectation-maximization (EM) algorithm similar to that of [23] but applied to isoforms instead of genes. Unlike the method of [42], which estimates isoform frequencies only from reads that map to a unique location in the genome, the algorithm of [43] incorporates multireads as well. The IE problem for single reads is also tackled in [1], who propose an EM algorithm for inferring isoform expression levels from the read coverage of exons (reads spanning exon junctions are ignored).

2.1.3 Our contributions

In this section we focus on the IE problem, namely estimating isoform expression levels (interchangeably referred to as frequencies) from RNA-Seq reads, under the assumption that a complete list of candidate isoforms is available. Projects such as [44] and [45] have already assembled large libraries of full-length cDNA sequences for humans and other model organisms, and the coverage of these libraries is expected to continue to increase rapidly following ultra-deep paired-end transcriptome sequencing projects such as [3, 4] and the widely anticipated deployment of third-generation sequencing technologies such as [46, 47], which deliver reads with significantly increased length. Inferring expression at isoform level provides information for finer-resolution biological studies, and also leads to more accurate estimates of expression at the gene level by allowing rigorous length normalization. Indeed, as shown in the 'Experimental results' section, genome-wide gene expression level estimates derived from isoform level estimates are significantly more accurate than those obtained directly from RNA-Seq data using isoform-oblivious GE methods such as the widely used counting of unique reads, the rescue method of [7], or the EM algorithm of [23].

Our main contribution is a novel expectation-maximization algorithm for isoform frequency estimation from any mixture of single and paired RNA-Seq reads. A key feature of our algorithm, referred to as IsoEM, is that it exploits information provided by the distribution of insert sizes, which is tightly controlled during sequencing library preparation under current RNA-Seq protocols. Such information is not modeled in the "exact" information models of [39, 40], challenging the validity of their negative results. Guttman et al. [4] take into account insert lengths derived from paired read data, but only for filtering candidate isoforms in ID. Trapnell et al. [3] is the only other work we are aware of that exploits this information for IE, in conjunction with paired read data. We show that modeling insert sizes is highly beneficial for IE even for RNA-Seq data consisting of single reads. Insert sizes contribute to increased estimation accuracy in two different ways. On one hand, they can help disambiguating the isoform of origin for the reads. In IsoEM, insert lengths are combined with base quality scores, and, if available, read pairing and strand information to probabilistically allocate reads to isoforms during the expectation step of the algorithm. As in [43], the genomic locations of multireads are also resolved probabilistically in this step, further contributing to improved overall accuracy compared to methods that ignore or fractionally pre-allocate multireads. On the other hand, insert size distribution is used to accurately adjust isoform lengths during frequency re-estimation in the maximization step of the IsoEM algorithm.

We also present the results of comprehensive experiments conducted to assess the performance of IsoEM on both synthetic and real RNA-Seq datasets. These results show that IsoEM consistently outperforms existing methods under a wide range of sequencing parameters and distribution assumptions. We also report results of experiments empirically evaluating the effect of sequencing parameters such as read length, read pairing, and strand information on estimation accuracy. Our experiments confirm the surprising finding of [43] that, for a fixed total number of sequenced bases, longer reads do not necessarily lead to better accuracy for estimation of isoform and gene expression levels.

2.2 Transcriptome Quantification Algorithms

2.2.1 Mapping RNA-Seq Reads

As with many RNA-Seq analyses, the first step of IsoEM is to map the reads. Our approach is to map them onto the library of known isoforms using any one of the many available ungapped aligners (we used Bowtie [48] with default parameters in our experiments). An alternative strategy is to map the reads onto the genome using a spliced alignment tool such as TopHat [30], as done, e.g., in [3, 4]. However, preliminary experiments with TopHat resulted in fewer mapped reads and significantly increased mapping uncertainty, despite providing TopHat with a complete set of annotated junctions. Since further increases in read length coupled with improvements in spliced alignment algorithms could make mapping onto the genome more attractive in the future, we made our IsoEM implementation compatible with both mapping approaches by always converting read alignments to genome coordinates and performing all IsoEM read-isoform compatibility calculations in genome space.

2.2.2 Finding read-isoform compatibilities

The candidate set of isoforms for each read is obtained by combining all genome coordinates of reads and isoforms, sorting them and using a line sweep technique to detect read-isoform compatibilities (see Algorithm 1). As detailed below, during the line sweep reads are grouped into equivalence classes defined by their isoform compatibility sets; this speeds up the E-step of the IsoEM algorithm by allowing the processing of an entire read class at once.

Some of the reads match multiple positions in the genome, which we refer to as *alignments* (for paired end reads, an alignment consists of the positions where the two reads in the pair align with the genome). Each alignment a can in turn be compatible with multiple isoforms that overlap at that position of the genome. During the line sweep, we compute the relative “weight” of assigning a given read/pair r to isoform j as $w_{r,j} = \sum_a Q_a F_a O_a$, where the sum is over all alignments of r compatible with j, and the factors of the summed products are defined as follows:

- Q_a represents the probability of observing the read from the genome locations described by the alignment. This is computed from the base quality scores as $Q_a = \prod_{k=1}^{|r|}[(1 - \varepsilon_k)M_{a_k} + \frac{\varepsilon_k}{3}(1 - M_{a_k})]$, where $M_{a_k} = 1$ if position k of alignment a matches the reference genome sequence and 0 otherwise, while ε_k denotes the error probability of k-th base of r.

- For paired end reads, F_a represents the probability of the fragment length needed to produce alignment a from isoform j; note that the length of this fragment can be inferred from the genome coordinates of the two aligned reads and the available isoform annotation. For single reads, we can only estimate an upperbound u on the fragment length: if the alignment is on the same strand as the isoform then u is the number of isoform annotated bases between the $5'$ end of the aligned read and the $3'$ end of the isoform, otherwise u is the number of isoform annotated bases between the $5'$ end of the aligned read and the $5'$ end of the isoform. In this case F_a is defined as the probability of observing a fragment with length of u bases or fewer.

- O_a is 1 if alignment a of r is consistent with the orientation of isoform j, and 0 otherwise. Consistency between the orientations of r and j depends on whether or not the library preparation protocol preserves the strand information. For single reads $O_a = 1$ when reads are generated from fragment ends randomly or, for directional RNA-Seq, when they match the known isoform orientation. For paired-end reads, $O_a = 1$ if the two reads come from different strands, point to each other, and, in the case of directional RNA-Seq, the orientation of first read matches the known isoform orientation.

Algorithm 1 The algorithm for identifying isoforms compatible with reads.

```
X = all the coordinates of all the entities (isoforms and reads)
sort X (radix sort; for equal values, isoform coordinates come first)
for $x$ in $X$ do
  $e$ = entityFor($x$)
  if $x$ is an entity end then
    sig = signature[$e$]
    gap = getLastGap(sig)
    if $x$ is an isoform end then
      currentIsoformsForGap[$gap$].remove($e$)
    else if $x$ is a read end then
      isoforms = currentIsoformsForGap[$gap$].keepOnlyMatching(sig)
      if read $e$ is the second read in the pair then
        isoformsForRead[$e$] = isoformsForRead[$e$]$\cap$ isoforms
      else
        isoformsForRead[$e$] = isoforms
      end if
      readClasses[isoformsForRead[$e$]].add($e$)
    end if
    signature.remove($e$)
  else
    signature[$e$].add($x$)
  end if
  if $x$ is an exon start then
    sig = signature[$e$]
    lastButOneGap = getLastButOneGap(sig)
    currentIsoformsForGap[$lastButOneGap$].remove($e$)
    lastGap = getLastGap(sig)
    currentIsoformsForGap[$lastGap$].add($e$, sig)
  end if
end for
```

2.2.3 IsoEM : *E*xpectation *M*aximization Algorithm for Estimation *Iso*form Frequencies

The IsoEM algorithm starts with the set of N known isoforms. For each isoform we denote by $l(j)$ its length and by $f(j)$ its (unknown) frequency. If we denote by $n(j)$ the number of reads coming from isoform j and let $p(k)$ denote the probability of a fragment of length k, then

$$E[n(j)] \propto \sum_{k \leq l(j)} p(k)(l(j) - k + 1) \tag{2.1}$$

since, the number of fragments of length k is expected to be proportional to the number of valid starting positions for a fragment of that length in the isoform. Thus, if the isoform of origin is known for each read, the maximum likelihood estimator for $f(j)$ is given by $c(j)/(c(1) + \ldots + c(N))$, where $c(j) = n(j)/\sum_{k \leq l(j)} p(k)(l(j) - k + 1)$ denotes the length-normalized fragment coverage. Note that the length of most isoforms is significantly larger than the mean fragment length μ typical of current sequencing libraries; for such isoforms $\sum_{k \leq l(j)} p(k)(l(j) - k + 1) \approx l(j) - \mu + 1$ and $c(j)$ can be approximated by $n(j)/(l(j) - \mu + 1)$.

Since some reads match multiple isoforms, their isoform of origin cannot be established unambiguously. The IsoEM algorithm (see Algorithm 2) overcomes this difficulty by simultaneously estimating the frequencies and imputing the missing read origin within an iterative framework. After initializing frequencies $f(j)$ at random, the algorithm repeatedly performs the next two steps until convergence:

- E-step: Compute the expected number $n(j)$ of reads that come from isoform j under the assumption that isoform frequencies $f(j)$ are correct, based on weights $w_{r,j}$ computed as described in the previous section
- M-step: For each j, set the new value of $f(j)$ to $c(j)/(c(1) + \ldots + c(N))$, where normalized coverages $c(j)$ are based on expected counts computed in the prior E-step

2.2.4 IsoEM optimizations

Below we describe two implementation optimizations that significantly improve the performance of IsoEM by reducing both runtime and memory usage.

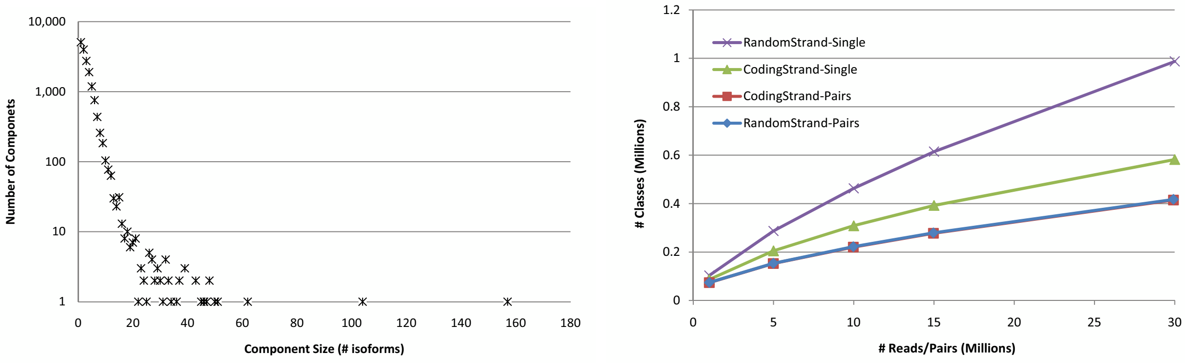
The first optimization consists of partitioning the input into compatibility components. The compatibility between reads and isoforms naturally induces a bipartite read-isoform compatibility graph, with edges connecting each isoform with all reads that can possibly

Algorithm 2 The expectation-maximization algorithm used by IsoEM.

```
assign random values to all $f(i)$
while not converged do
    E-step:
    initialize all $n(j)$ to 0
    for each read r do
        sum = $\sum_{j:w_{r,j}>0} w_{r,j} f(j)$
        for each isoform $j$ with $w_{r,j} > 0$ do
            $n(j) += w_{r,j} f(j)/\text{sum}$
        end for
    end for
    M-step:
    $s = \sum_j n(j)/(l(j) - \mu + 1)$
    for each isoform j do
        $f(j) = \frac{n(j)/(l(j)-\mu+1)}{s}$
    end for
end while
```

originate from it. Connected components of the compatibility graph can be processed independently in IsoEM since the frequencies of isoforms in one connected component do not affect the frequencies of isoforms in any other connected component. Although this optimization can be applied to any EM algorithm, its impact is particularly significant in IsoEM. Indeed, in this context the compatibility graph decomposes in numerous small components (see Figure 2.1(a) for a typical distribution of component sizes). The resulting speed-up comes from the fact that in each iteration of IsoEM we update frequencies of isoforms in a single compatibility component, avoiding needless updates for other isoforms.

The second IsoEM optimization consists of partitioning the set of reads within each compatibility component into equivalence classes. Two reads are equivalent for IsoEM if they are compatible with the same set of isoforms and their compatibility weights to the isoforms are proportional. Keeping only a single representative from each read class (with appropriately adjusted frequency) drastically reduces the number of reads kept in memory (see Figure 2.1(b)). As the number of reads increases, the number of read classes increases much slower. Eventually this reaches saturation and no new read classes appear – at which point the runtime of IsoEM becomes virtually independent of the number of reads. Indeed,



(a) (b)

Figure 2.1 Distribution of compatibility component sizes (defined as the number of isoforms) for 10 million single reads of length 75 (a) and number of read classes for 1 to 30 million single reads or pairs of reads of length 75 (b).

```
E-step for read classes:
initialize all $n(j)$ to 0
for each read class $R$ do
    sum $= \sum_{j:w_{R,j}>0} w_{R,j} f(j)$
    for each isoform $j$ with $w_{R,j} > 0$ do
        $n(j)+ = m(R) * w_{R,j} f(j)/\text{sum}$
    end for
end for
```

Figure 2.2 The E-Step of IsoEM algorithm based on read classes.

in practice the runtime bottlenecks are parsing the reads, computing the compatibility graph and detecting equivalent reads.

Once read classes are constructed, we only need a small modification of the E-step of IsoEM to use read classes instead of reads (Figure 2.2). Next we describe the union-find algorithm used for efficiently finding compatibility components and read classes in IsoEM. A read class is defined as $\langle m, \{(i, w) | i = \text{isoform}, w = \text{weight}\}\rangle$, where m is called the multiplicity of the read class. Given a collection of reads, we want to:

- Find the connected components of the compatibility graph induced by the reads, and

- Collapse equivalent reads into read classes with multiplicity indicating the number of reads in each class.

A straightforward approach is to solve the first problem using a union-find algorithm, then to take the reads corresponding to each connected component and remove equivalent reads, e.g., using hashing. However, there are two drawbacks to this approach:

- First, all reads need to be kept in memory until all connected components have been computed.
- Second, when the number of reads in a connected component is very large the number of collisions increases, which leads to poor performance.

We overcome the two problems presented above using an online version of the union-find algorithm which computes connected components and eliminates equivalent reads on the fly. This way, equivalent reads will never reside too long in memory. Also, we avoid the problem of large hash tables by using multiple smaller hash tables which are guaranteed to be disjoint.

We start our modified version of union-find with an empty set of trees. A new single-node tree is initialized every time a new isoform is found in a read class. In each node we store a hash-table of read classes. Each read is processed as follows:

- *If the isoforms compatible with the read correspond to nodes in more than one tree* unite the corresponding trees. The root of the tallest tree becomes the root of the union tree. Then create a new read class for this read (we can be sure it was not seen before, otherwise the isoforms would have been in the same tree) and add it to the hash table of the root node. Notice that at this point the root node is also (trivially) the Lowest Common Ancestor (LCA) of the nodes corresponding to the isoforms in the read class
- *If the isoforms correspond to nodes in the same tree* find the LCA of all these nodes. If the class of the read is present in the hash table of the LCA, increment its multiplicity

and then drop the read. Otherwise, create a new read class and add it to the LCA's hash table.

Notice that in the second case it suffices to look only in the LCA of the isoforms for an already existing read class. This follows immediately from the fact that we always add reads to the LCA of the nodes (isoforms) compatible with the read. Note that we cannot use path compression to speed up 'find' operations because this would be altering the structure of existing trees. Thus, 'find' operations will take logarithmic (amortized) time. At the end of the algorithm, each tree in the union-find forest corresponds to a connected component. The read classes in each connected component are obtained by traversing the corresponding tree and collecting all the read classes present in the nodes. At this point we are sure that all the read classes are distinct, so the collection process performs simple concatenations. To further speed up the collection process, we can safely use path compression as we traverse the trees, since we no longer care about the exact topology of the subtrees.

Runtime analysis. Each union operation takes $O(1)$ time, so for a read with k compatible isoforms we spend at most $O(k)$ time doing unions. By always making the root of the taller tree to be the root of a union, we ensure that the height of any tree is not bigger than $O(\log n)$ where n is the number of nodes in the tree. Thus, finding the root of a node's tree takes $O(\log n)$. For a read with k compatible isoforms we spend at most $O(k \log n)$ time processing it. The LCA of two nodes can be computed at constant overhead when performing find operations (by marking the nodes on the paths from isoforms to root). Collecting all the read classes is sped-up by using path compression. The whole collecting phase takes $O(n\alpha(n))$ time where n is the total number of isoforms and $\alpha(n)$ is the inverse of the Ackermann function. Overall, for q reads with an average of k isoforms per read and n total distinct isoforms, computing read classes and compatibility components using the modified union-find algorithm takes $O(qk \log n + n\alpha(n))$ time.

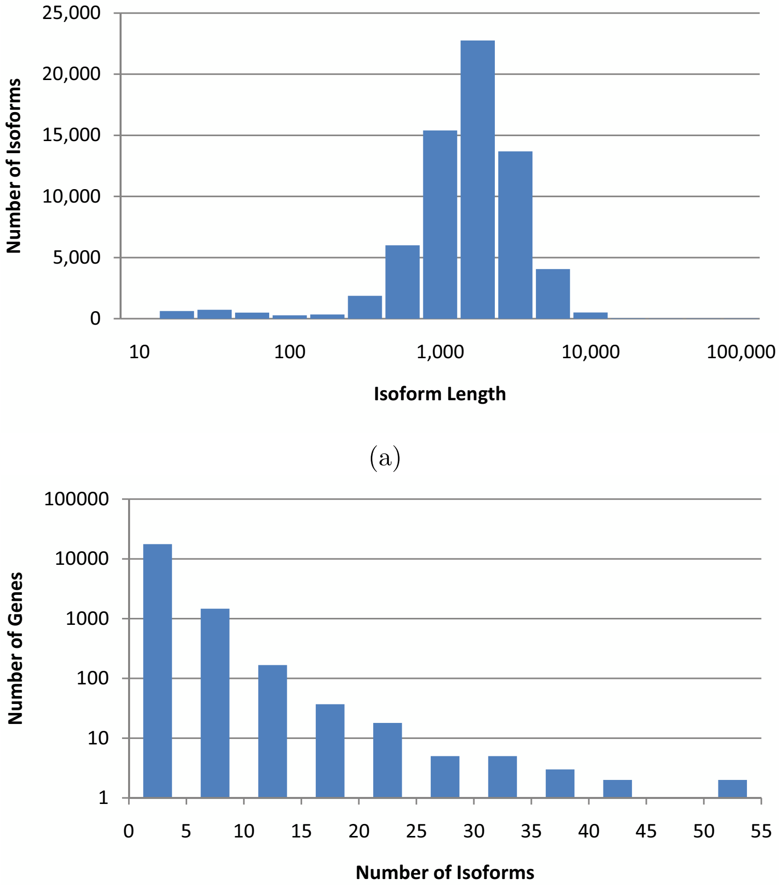
2.2.5 Hexamer and repeat bias corrections

As noted in [49], some commonly used library preparation protocols result in biased sampling of fragments from isoforms due to the random hexamers used to prime reverse transcription. To correct for possible hexamer bias, we implemented a simple re-weighting scheme similar to that proposed in [49]. Each read is assigned a weight $b(h)$ based on its first six bases and computed as follows. Given a set of mapped reads, let $\hat{p}_i$ be the observed distribution of hexamers starting at position i (spanning positions i to $i+5$) of all the reads. Thus, $\hat{p}_i(h)$ is the proportion of reads which have hexamer h at position i and $\hat{p}_1(h)$ is the proportion of reads starting with hexamer h. Let l be the read length. We define the weights b by:

$$b(h) = \frac{\frac{1}{6}\sum_{i=l/2-2}^{l/2+3} \hat{p}_i(h)}{\frac{1}{2}(\hat{p}_1(h) + \hat{p}_2(h))}$$

Since we already collapse equivalent reads into read classes, we can seamlessly incorporate hexamer weights in the algorithm by slightly changing the definition of a read class' multiplicity to $m(R) = \sum_{r\in R} b(h(r))$, where $h(r)$ denotes the starting hexamer of r. The effect of this correction procedure is to reduce (respectively increase) the multiplicity of reads with starting hexamers that are overrepresented (respectively under-represented) at the beginning of reads compared to the middle of reads. The underlying assumption is that the average frequency with which a hexamer appears in the middle of reads is not affected by library preparation biases. Recent methods further target biases in the bases surrounding the sequenced fragments in addition to those at read ends.

To avoid biases from incorrectly mapped reads originating from repetitive regions, IsoEM will also discard reads that overlap annotated repeats. When applying this correction, isoform lengths are automatically adjusted by subtracting the number of positions resulting in reads that would be discarded.



(a)

(b)

Figure 2.3 Distribution of isoform lengths (a) and gene cluster sizes (b) in the UCSC dataset.

2.3 Experimental results

2.3.1 Comparison of methods on simulated datasets

We tested IsoEM on simulated human RNA-Seq data. The human genome sequence (hg18, NCBI build 36) was downloaded from UCSC together with the coordinates of the isoforms in the KnownGenes table. Genes were defined as clusters of known isoforms defined by the GNFAtlas2 table. The dataset contains a total of $66,803$ isoforms pertaining to $19,372$ genes. The isoform length distribution and the number of isoforms per genes are shown in Figure 3.10.

Single and paired-end reads were randomly generated by sampling fragments from the known isoforms. Each isoform was assigned a *true frequency* based on the abundance reported for the corresponding gene in the first human tissue of the GNFAtlas2 table, and a probability distribution over the isoforms inside a gene cluster. Thus, the true frequency of isoform j is $a(g)p(j)$, where $a(g)$ is the abundance of the gene g for which j is an isoform and $p(j)$ is the probability of isoform j among all the isoforms of g. We simulated datasets with uniform, respectively truncated geometric distribution with ratio $r = 1/2$ for the isoforms of each gene. For a gene with k isoforms $p(j) = 1/k$, $j = 1, \ldots, k$, under the uniform distribution. Under the truncated geometric distribution, the respective isoform probabilities are $p(j) = 1/2^j$ for $j = 1, \ldots, k-1$ and $p(k) = 1/2^{k-1}$. Fragment lengths were simulated from a normal probability distribution with mean 250 and standard deviation 25.

We compared IsoEM to several existing algorithms for solving the IE and GE problems. For IE we included in the comparison the isoform analogs of the Uniq and Rescue methods used for GE [7], an improved version of Uniq (UniqLN) that estimates isoform frequencies from unique read counts but normalizes them using adjusted isoform lengths that exclude ambiguous positions, the Cufflinks algorithm of [3] (version 0.8.2), and the RSEM algorithm of [43] (version 0.6). For the GE problem, the comparison included the Uniq and Rescue methods, our implementation of the GeneEM algorithm described in [23], and estimates obtained by summing isoform expression levels inferred by Cufflinks, RSEM, and IsoEM. All methods use alignments obtained by mapping reads onto the library of isoforms with Bowtie [48] and then converting them to genome coordinates, except for Cufflinks which uses alignments obtained by directly mapping the reads onto the genome with TopHat [30], as suggested in [3].

Frequency estimation accuracy was assessed using the coefficient of determination, r^2, along with the *error fraction (EF)* and *median percent error (MPE)* measures used in [43]. However, accuracy was computed against true frequencies, not against estimates derived from true counts as in [43]. If $\hat{f}_i$ is the frequency estimate for an isoform with true frequency f_i, the *relative error* is defined as $|\hat{f}_i - f_i|/f_i$ if $f_i \neq 0$, 0 if $\hat{f}_i = f_i = 0$, and ∞ if $\hat{f}_i > f_i = 0$.

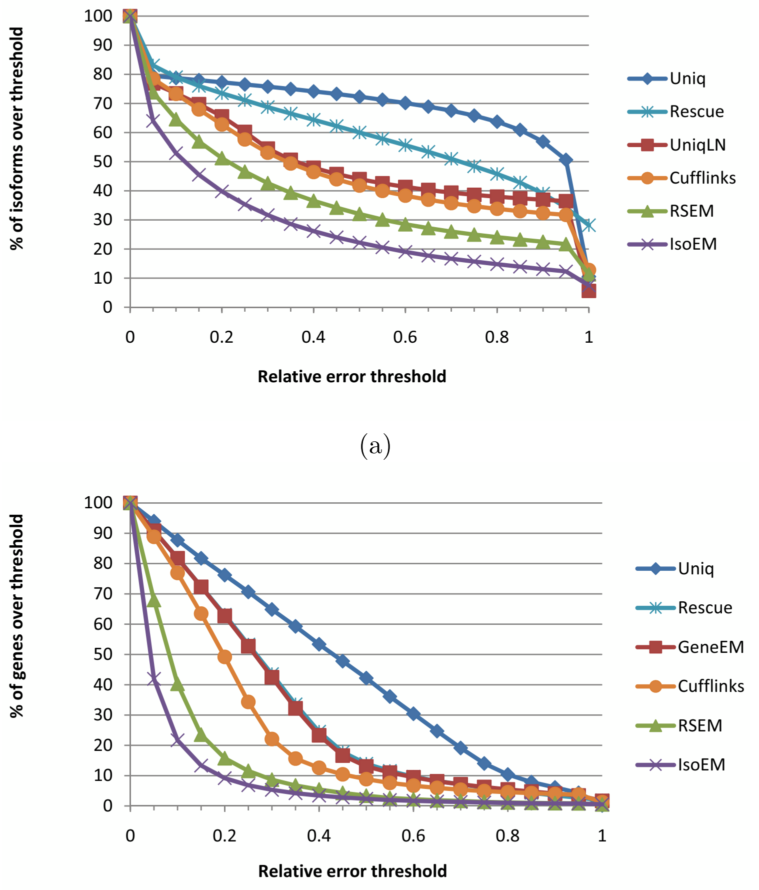
Table 2.1 r^2 for isoform and gene expression levels inferred from 30M reads of length 25 from reads simulated assuming uniform, respectively geometric expression of gene isoforms.

Isoform Expression			Gene Expression		
Algorithm	Uniform	Geometric	Algorithm	Uniform	Geometric
Uniq	0.466	0.447	Uniq	0.579	0.586
Rescue	0.693	0.675	Rescue	0.724	0.724
UniqLN	0.856	0.838	GeneEM	0.636	0.637
Cufflinks	0.661	0.618	Cufflinks	0.778	0.757
RSEM	0.919	0.911	RSEM	0.939	0.934
IsoEM	**0.980**	**0.971**	IsoEM	**0.991**	**0.981**

The error fraction with threshold τ, denoted EF_τ is defined as the percentage of isoforms with relative error greater or equal to τ. The median percent error, denoted MPE, is defined as the threshold τ for which $EF_\tau = 50\%$.

Since not all compared methods could handle paired reads or strand information we focused our comparisons on single read data. Table 1 gives r^2 values for isoform, respectively gene expression levels inferred from 30M reads of length 25, simulated assuming both uniform and geometric isoform expression. IsoEM significantly outperforms the other methods, achieving an r^2 values of over .96 for all datasets. For all methods the accuracy difference between datasets generated assuming uniform and geometric distribution of isoform expression levels is small, with the latter one typically having a slightly worse accuracy. Thus, in the interest of space we present remaining results only for datasets generated using geometric isoform expression.

For a more detailed view of the relative performance of compared IE and GE algorithms, Figure 6 3.3 gives the error fraction at different thresholds ranging between 0 and 1. The variety of methods included in the comparison allows us to tease out the contribution of various algorithmic ideas to overall estimation accuracy. The importance of rigorous length normalization is illustrated by the significant IE accuracy gain of UniqLN over Uniq – clearly larger than that achieved by ambiguous read reallocation as implemented in the IE version of Rescue. Proper length normalization is also explaining the accuracy gain of isoform-aware



(a)

(b)

Figure 2.4 Error fraction at different thresholds for isoform (a) and gene (b) expression levels inferred from 30M reads of length 25 simulated assuming geometric isoform expression.

GE methods (Cufflinks, RSEM, and IsoEM) over isoform oblivious GE methods. Similarly, the importance of modeling insert sizes even for single read data is underscored by the significant IE and GE accuracy gains of IsoEM over RSEM. Indeed, the latest version of the RSEM package, released as this article goes to print, has been updated to include modeling of insert sizes and appears to have accuracy matching that of IsoEM.

For yet another view, Tables 2 and 3 report the MSE and EF$_{.15}$ measures for isoform, respectively gene expression levels inferred from 30M reads of length 25, computed over groups of isoforms with various expression levels. IsoEM consistently outperforms the other

Table 2.2 Median percent error (MPE) and 15% error fraction ($\mathrm{EF}_{.15}$) for isoform expression levels inferred from 30M reads of length 25 simulated assuming geometric isoform expression.

Expression range		0	$(0, 10^{-6}]$	$(10^{-6}, 10^{-5}]$	$(10^{-5}, 10^{-4}]$	$(10^{-4}, 10^{-3}]$	$(10^{-3}, 10^{-2}]$	All
# isoforms		13,290	10,024	23,882	18,359	1,182	66	66,803
MPE	Uniq	**0.0**	**100.0**	98.4	97.1	98.5	96.6	95.4
	Rescue	**0.0**	294.7	75.5	49.2	30.4	28.3	71.9
	UniqLN	**0.0**	**100.0**	80.8	30.3	26.4	24.8	36.0
	Cufflinks	**0.0**	**100.0**	49.7	25.5	27.2	44.6	34.1
	RSEM	**0.0**	**100.0**	31.9	13.5	11.4	13.0	21.2
	IsoEM	**0.0**	**100.0**	**25.3**	**7.3**	**3.2**	**2.2**	**12.0**
$\mathrm{EF}_{.15}$	Uniq	**0.2**	98.4	97.2	96.9	97.0	95.5	78.0
	Rescue	48.4	95.5	86.2	73.1	61.5	56.1	76.0
	UniqLN	**0.2**	97.2	86.2	82.8	83.3	77.3	69.8
	Cufflinks	17.6	96.4	81.3	71.0	74.7	80.3	67.9
	RSEM	19.9	93.7	71.1	46.4	39.8	47.0	56.9
	IsoEM	3.4	**93.1**	**65.1**	**29.1**	**11.1**	**7.6**	**46.1**

IE and GE methods at all expression levels except for isoforms with zero true frequency, where it is dominated by the more conservative Uniq algorithm and its UniqLN variant.

2.3.2 Comparison of methods on two real RNA-Seq datasets

In addition to simulation experiments, we validated IsoEM on two real RNA-Seq datasets. The first dataset consists of two samples with approximately 8 million 27bp Illumina reads each, generated from two human cell lines (embryonic kidney and B cells) as described in [50]. Estimation accuracy was assessed by comparison with quantitative PCR (qPCR) expression levels determined in [1] for 47 genes with evidence of alternative isoform expression. To facilitate comparison with these qPCR results, expression levels were determined using transcript annotations in ENSEMBL version 46. The second dataset consists of approximately 5 million 32bp Illumina reads per sample, generated from the RM11-1a strain of *S. cerevisiae* under two different nutrient conditions [2]. Expression levels were determined using transcript annotations for the reference strain (June 2008 SGD/sacCer2) and compared against qPCR expression levels measured for 192 genes (for a total of 394 datapoints).

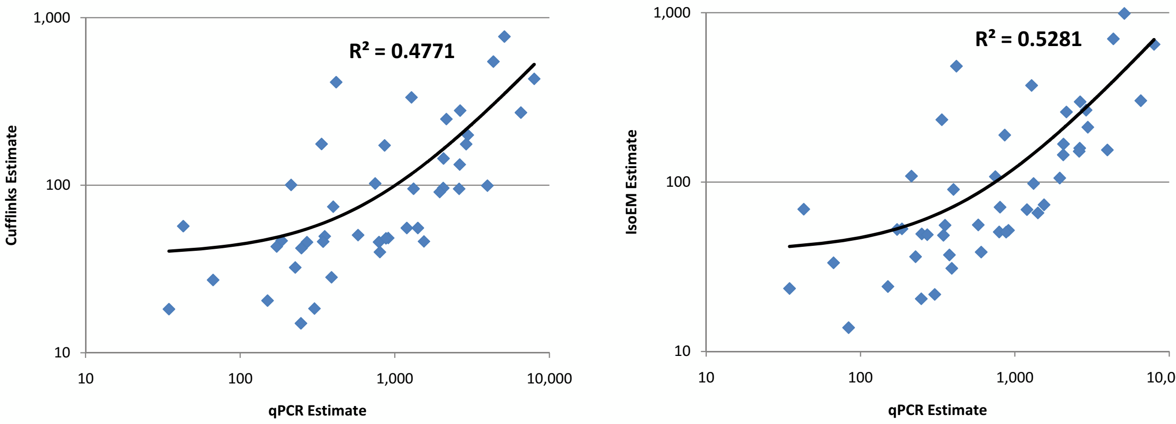
Table 2.3 Median percent error (MPE) and 15% error fraction (EF$_{.15}$) for gene expression levels inferred from 30M reads of length 25 simulated assuming geometric isoform expression.

	Expression range	$(0, 10^{-6}]$	$(10^{-6}, 10^{-5}]$	$(10^{-5}, 10^{-4}]$	$(10^{-4}, 10^{-3}]$	$(10^{-3}, 10^{-2}]$	All
	# genes	120	5,610	11,907	1,632	102	19,372
MPE	Uniq	37.4	43.6	42.7	43.0	48.2	43.0
	Rescue	32.8	28.7	26.0	25.1	28.8	26.7
	GeneEM	30.6	28.2	25.7	25.1	28.0	26.3
	Cufflinks	33.0	21.1	19.0	20.2	40.2	19.7
	RSEM	23.6	11.0	7.2	7.9	11.4	8.1
	IsoEM	**18.2**	**8.4**	**3.2**	**2.0**	**1.9**	**3.9**
EF$_{.15}$	Uniq	77.5	82.4	81.7	79.7	82.4	81.7
	Rescue	74.2	74.0	71.6	72.8	76.5	72.4
	GeneEM	72.5	73.8	71.5	73.0	74.5	72.3
	Cufflinks	73.3	64.7	62.3	66.2	82.3	63.5
	RSEM	64.2	37.3	17.4	16.3	41.2	23.5
	IsoEM	**57.5**	**28.1**	**6.7**	**6.1**	**4.9**	**13.2**

Since the available implementation of RSEM could not be run on transcript sets other than UCSC known genes, in Figures 7 2.5 and 8 2.6 we only compare Cufflinks and IsoEM estimates against qPCR values in [1], respectively [2]. Estimation accuracy of both Cufflinks and IsoEM is significantly lower than that observed in simulations. Likely explanations include poor quality of the transcript libraries used to perform the inference, sequencing library preparation biases not corrected for by the algorithms, and possible inaccuracies in qPCR estimates. Nevertheless, the relative performance of the two algorithms is consistent with simulation results, with IsoEM outperforming Cufflinks on both datasets.

2.3.3 Influence of sequencing parameters and scalability

Although high-throughput technologies allow users to make tradeoffs between read length and the number of generated reads, very little has been done to determine optimal parameters even for common applications such as RNA-Seq. The intuition that longer reads are better certainly holds true for many applications such as *de novo* genome and transcriptome assembly. Surprisingly, [43] found that *shorter* reads are better for IE when the total number of sequenced bases (as a rough approximation for sequencing cost) is fixed. Figure 9 2.7 plots IE estimation accuracy for reads of length between 10 and 100 when the total



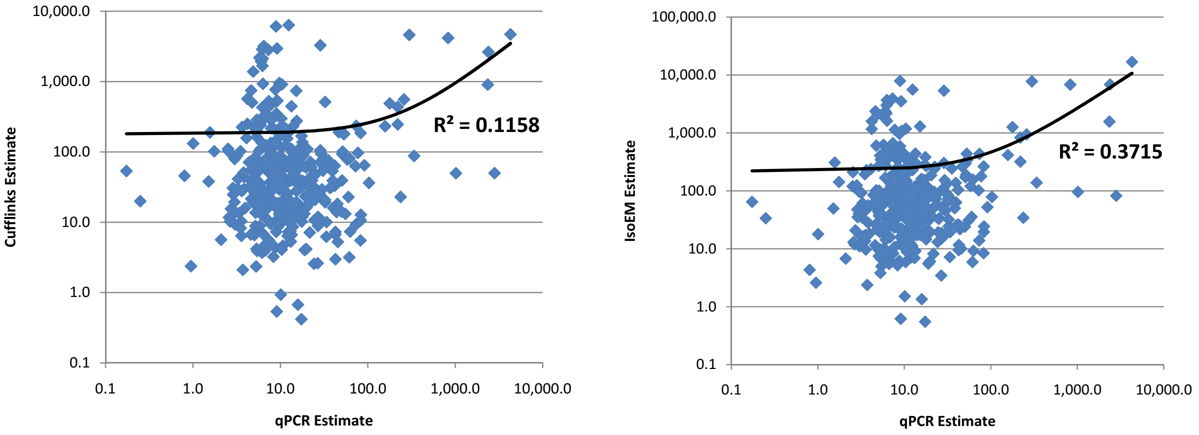
(a)

(b)

Figure 2.5 Comparison of Cufflinks (a) and IsoEM (b) estimates to qPCR expression levels reported in [1].

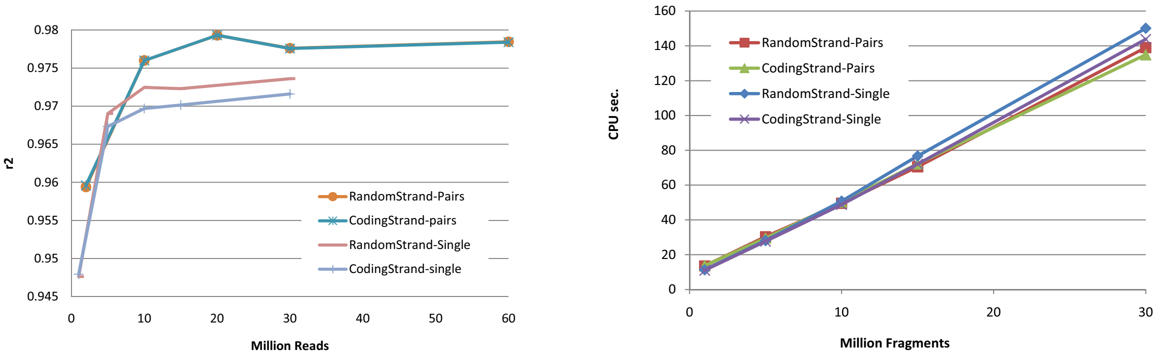
amount of sequence data is kept constant at 750M bases. Our results confirm the finding of [43], although the optimal read length is somewhat sensitive to the accuracy measure used and to the availability of pairing information. While 25bp reads minimize MPE regardless of the availability of paired reads, the read length that maximizes r^2 is 25 for paired reads and 50 for single reads. Although further experiments are needed to determine how the optimum length depends on the amount of sequence data and transcriptome complexity, our simulations do suggest that for isoform and gene expression analysis, increasing the number of reads may be more useful than increasing read length beyond 50 bases.

Figure 2.8(a) shows, for reads of length 75, the effects of paired reads and strand information on estimation accuracy as measured by r^2. Not surprisingly, for a fixed number of reads, paired reads yield better accuracy than single reads. Also not very surprisingly, adding strand information to paired sequencing yields no benefits to genome-wide IE accuracy (although it may be helpful, e.g., in identification of novel transcripts). Quite surprisingly, performing strand-specific single read sequencing is actually *detrimental* to IsoEM IE (and hence GE) accuracy under the simulated scenario, most likely due to the reduction in sampled transcript length.



(a) (b)

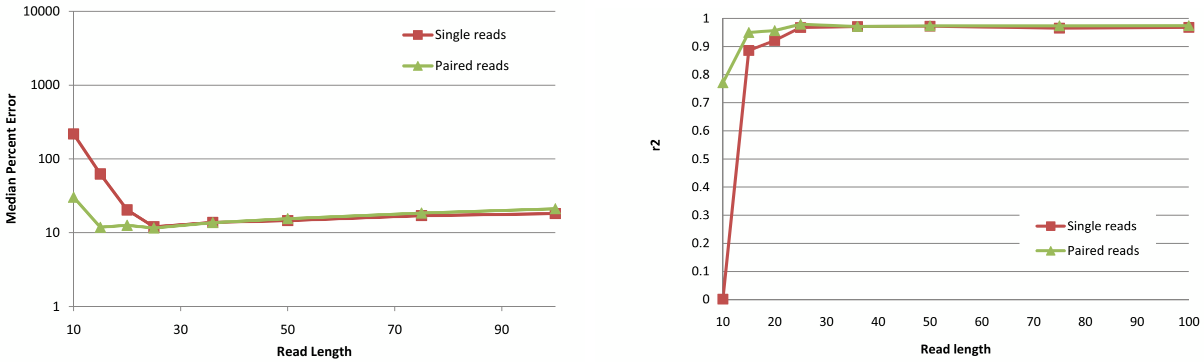
Figure 2.6 Comparison of Cufflinks (a) and IsoEM (b) estimates to qPCR expression levels reported in [2].



(a) (b)

Figure 2.8 IsoEM r^2 (a) and CPU time (b) for 1-60 million single/paired reads of length 75, with or without strand information.

In practice, many RNA-Seq data sets are generated from transcripts with poly(A) tails, and some of the sequenced fragments will contain parts the poly(A) tails. We have added to IsoEM the option to automatically extend annotated transcripts with a poly(A) tail, thus allowing it to use reads coming from such fragments. Table 4 shows the accuracy of isoform and gene expression levels inferred by IsoEM using 30M reads of length 25 simulated from



(a) (b)

Figure 2.7 IsoEM MPE (a) and r^2 values (b) for 750Mb of simulated data generated using single and paired-end reads of length varying between 10 and 100.

transcripts with and without poly(A) tails assuming geometric expression of gene isoforms. The accuracy of IsoEM is practically the same under the two simulation scenarios for paired read data, and decreases only slightly for single reads simulated taking poly(A) tails into account, likely due to the fact that reads overlapping poly(A) tails are more ambiguous.

As shown in Figure 2.8(b), the runtime of IsoEM scales roughly linearly with the number of *fragments*, and is practically insensitive to the type of sequencing data (single or paired reads, directional or non-directional). IsoEM was tested on a Dell PowerEdge R900 server with 4 Six Core E7450Xeon Processors at 2.4Ghz (64 bits) and 128Gb of internal memory. None of the datasets required more than 16GB of memory to complete. It is also true that increasing the available memory significantly decreases runtime by keeping the garbage collection overhead to a minimum. The runtimes in Figure 2.8 were obtained by allowing IsoEM to use up to 32GB of memory, in which case none of the datasets took more than 3 minutes to solve.

2.4 Conclusions

We have introduced an expectation-maximization algorithm for isoform frequency estimation assuming a known set of isoforms. Our algorithm, called IsoEM, explicitly models insert size distribution, base quality scores, strand and read pairing information. Experiments on both real and synthetic RNA-Seq datasets generated using two different assumptions on the isoform distribution show that IsoEM consistently outperforms existing algorithms for isoform and gene expression level estimation with respect to a variety of quality metrics.

PART 3

TRANSCRIPTOME RECONSTRUCTION

3.1 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 [7–9, 43]. Unfortunately, as shown by recent targeted RNA-Seq studies [15], 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.

3.1.1 Background

RNA-Seq is quickly becoming the technology of choice for transcriptome research and analyses [14]. RNA-Seq allows reduction of the sequencing cost and significantly increases data throughput, but it is computationally challenging to use such RNA-Seq data for reconstructing of full length transcripts and accurately estimate their abundances across all cell types. The common computational problems include: gene and isoform expression level estimation, transcriptome quantification, transcriptome discovery and reconstruction. To solve these problems requires scalable computational tools [24]. A variety of new methods and tools have been recently developed to tackle these problems.

3.1.2 Related Work

RNA-Seq analyses 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 [30] or SpliceMap [31].

Identifying of all transcripts expressed in a particular sample require the assembly of reads into transcription units. This process is collectively called transcriptome reconstruction. 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 [24]. Genome-independent methods such as Trinity [25] or transAbyss [26] directly assemble reads into transcripts. A commonly used approach for such methods is de Brujin graph [27] 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 [28] 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 [30] or SpliceMap [31]. 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 [4] 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. [3, 32], 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. Cufflinks and Scripture do not target the quantification accu-

Table 3.1 Classification of transcriptome reconstruction methods

Method	Support paired-end reads	Consider fragment lenght distribution	Require annotation
TRIP	Yes	Yes	No
IsoLasso	Yes	No	No
IsoInfer	No	No	TES/TSS
Cufflinks	Yes	Yes	No
CLIQ	No	No	No
Scripture	Yes	No	No
SLIDE	Yes	No	gene/exon boundaries

racy. IsoLasso [5] uses the LASSO [33] 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 [34] 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.

Table 3.1 includes classification of the available methods for genome-guided transcriptome reconstruction based on supported parameters and underlying algorithms.

3.1.3 Our Contribution

We focus on the problem of transcriptome reconstruction from RNA-Seq data assisted by existing genome and transcriptome annotations. To address transcriptome reconstruction problem we developed annotation-guided and genome-guided methods.

In section 3.2 we propose a novel annotation-guided general framework for transcriptome discovery, reconstruction and quantification in partially annotated genomes, referred as **D**iscovery and **R**econstruction of **U**nannotated **T**ranscripts (DRUT). DRUT framework incorporates an enhancement of EM algorithm,VTEM [35] [36], to detect overexpressed reads and/or exons corresponding to the unannotated transcripts and to estimate annotated transcript frequencies. Our main contribution is an expectation-maximization based method for discovery of unannotated transcripts when partial information about genome annotation is given. A key feature of our algorithm is its usage of the existing genome annotation information to detect reads from unannotated transcripts and accurately estimate annotated transcripts abundances. Moreover, the algorithm applies transcriptome assembler on subset of reads to improve the quality of the transcriptome reconstruction. The recently published paper [28] is the only related work that we are aware of, which exploits information about genome annotations. RABT is an annotation-guided assembler built upon Cufflinks assembler [3] that determines the minimum number of transcripts needed to explain reads mapped to the reference genome.

We also present experimental results on *in silico* datasets generated with various sequencing parameters and distribution assumptions. The results show that DRUT overperforms existing genome-guided transcriptome assemblers and show similar or better performance with existing annotation-guided assemblers. Testing DRUT for transcriptome quantification implies usage of VTEM [35] algorithm for partially annotated transcripts. Our experimental studies show that DRUT significantly improves estimation of transcipts frequencies in comparison to our previous method IsoEM [9] for partially annotated genomes.

In section 3.3 a novel "genome-guided" method called "**T**ranscriptome **R**econstruction using **I**nteger **P**rogramming" (TRIP) is proposed. 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.

3.2 Annotation-guided Transcriptome Reconstruction Algorithms

3.2.1 Mapping RNA-Seq Reads and Exon Counts

As with many RNA-Seq analyses, the first step of DRUT is to map reads (see Fig. 3.2a). Our approach maps reads onto the library of annotated transcripts using any one of the many available ungapped aligners (we used Bowtie [48] with default parameters in our experiments). An alternative strategy is to map the reads onto the genome using a spliced alignment tool such as TopHat [30], as done in [3, 4].

Based on the reads mapped to the set of annotated transcripts it is possible to calculate observed exon counts. Exon counts are calculated based both on the spliced and unspliced reads. For the spliced reads the contribution of the read is equal to the part of the read mapped to particular exon.

3.2.2 VTEM : *V*irtual *T*ranscript *E*xpectation *M*aximization Algorithm

In this section we first formally define the panel and describe expectation - maximization (EM) algorithm for transcriptome quantification, referred as IsoEM [9]. Then we show how to estimate the quality of the model based on EM algorithm and introduce enhancement of IsoEM algorithm with the virtual transcript.

IsoEM is a novel expectation-maximization algorithm for inference of alternative splicing isoform frequencies from high-throughput transcriptome sequencing (RNA-Seq) data proposed in [9]. IsoEM takes advantage of base quality scores, strand information and exploits unambiguous information provided by the distribution of insert sizes generated during sequencing library preparation.

The input data for IsoEM consists of a *panel*, i.e. a bipartite graph $G = (T \bigcup R)$, such that 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 unknown frequency f_s of the transcript, and with each vertex $r \in R$ we associate observed read frequency o_r. Then for the each pair t_i, r_j, we add an edge (t_i, r_j) weighted by the probability of the transcript t_i to emit a read r_j.

Regardless of initial conditions, EM algorithm always converges to a maximum likelihood solution (see [23]). The algorithm starts with the set of N transcripts. After uniform initialization of the frequencies $f_t, t \in T$, the algorithm repeatedly performs the next two steps until convergence:

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

We propose an enhancement of the IsoEM algorithm with the virtual transcript, referred as **V**irtual **T**ranscript **E**xpectation **M**aximization (VTEM). We consider two modification of the panel:

- bipartite graph $G = (T \bigcup R)$, such that each transcript is represented as a vertex $t \in T$, and each read is represented as a vertex $r \in R$.
- bipartite graph $G = (T \bigcup E)$, such that each transcript is represented as a vertex $t \in T$, and each exon is represented as a vertex $e \in E$.

This leads to two new versions of VTEM algorithm. First version, referred as **r**ead **V**irtual **T**ranscript **E**xpectation **M**aximization (rVTEM) algorithm, uses the panel consisting of the set of transcripts and reads with observed counts, similar to IsoEM([9]) algorithm. In the second version, referred as **e**xon **V**irtual **T**ranscript **E**xpectation **M**aximization (eVTEM) algorithm, we replace the reads in the panel by the corresponding exons with the observed counts (calculated as described in 3.2.1). Further we will refer to the reads and exons as *segments* emitted by transcripts; both read frequencies and exon counts will be referred as *segment frequencies*.

In order to decide weather the panel is incomplete we need to measure how well maximum likelihood model explains the segment frequencies. We suggest to measure the model quality by the deviation between expected and observed segment frequencies:

$$D = \frac{\sum_j |o_j - e_j|}{|S|},$$

where $|S|$ is the number of segments, o_j is the observed segment frequencies s_j, and e_j is the expected segment frequencies s_j, calculated as follows:

$$s_j = \sum_i \frac{h_{t_i,j}}{\sum_l h_{t_i,l}} f_i^{ML}, \tag{3.1}$$

where $h_{t_i,j}$ is weighted match between segment s_j and transcript t_i and f_j^{ML} is the maximum-likelihood frequency of the transcript t_i.

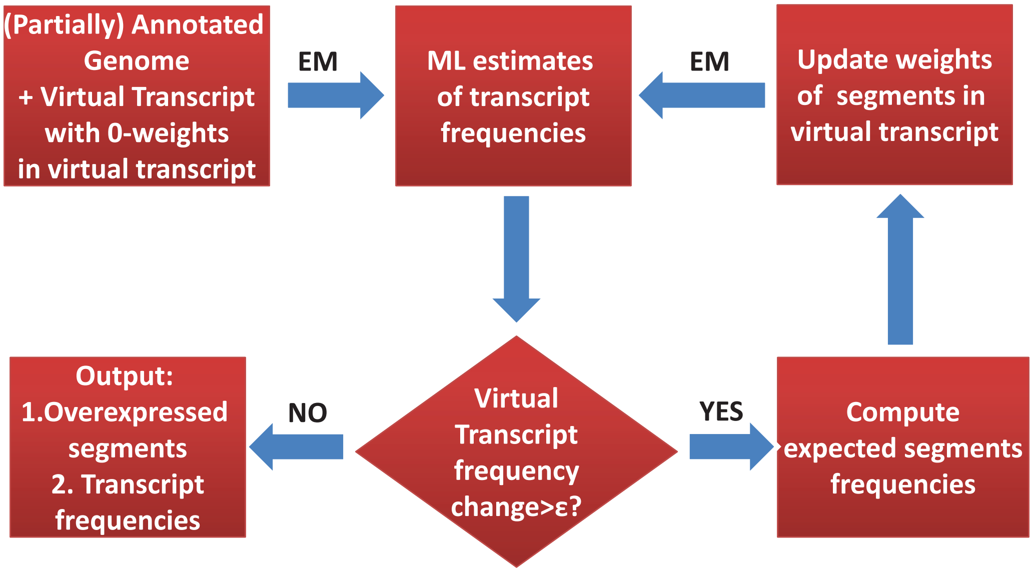

(Partially) Annotated Genome + Virtual Transcript with 0-weights in virtual transcript
EM
ML estimates of transcript frequencies
EM
Update weights of segments in virtual transcript
Output: 1.Overexpressed segments 2. Transcript frequencies
NO
Virtual Transcript frequency change>ε?
YES
Compute expected segments frequencies

Figure 3.1 Flowchart for VTEM.

The main idea of the VTEM algorithm (see Algorithm 3) is to add a virtual transcript vt to the set of known transcripts. This way virtual transcript vt emits segments that do not fit well to annotated transcripts. The flowchart of VSEM, shown in Fig. 3.1, can be explained as follows: initially, all segments are connected to the virtual transcript with weight $h_{t_i,j} = 0$. The first iteration finds the maximum likelihood frequency estimations of candidates transcripts; maximum likelihood frequency estimations of virtual transcript will be equal to 0, since all the edges between the virtual transcript and segments $h_{vt,j} = 0$. Then these estimations are used to compute the expected frequencies of the segments according to (3.3). If the expected segment frequency is less than the observed one (i.e., "underexpressed"), then the lack of the segment expression is added to the weight of the edge between this segment and the virtual transcript. For "overexpressed" segments the excess of segment's expression is subtracted from the corresponding weight (but keeping it non-negative). The iterations are continued while the virtual transcript frequency change is decreasing by more than ϵ.

The frequency f_i of the virtual transcript estimates the total frequency of unannotated transcripts. Therefore, based on the frequency of the virtual transcript, we can decide if the panel is likely to be incomplete, i.e., genome is partially annotated. Furthermore, the output of VTEM contains both the estimated frequency of the virtual transcript and the weights of the edges, connecting segments with the virtual transcript.

These weights can be interpreted as the probabilities of the segments to be the part of the unannotated transcripts. In order to select segments corresponding to these transcripts, it is enough to select f_i most probable overexpressed segments (see Fig. 3.1b).

3.2.3 DRUT : Method for *D*iscovery and *R*econstruction of *U*nannotated *T*ranscripts

In this section, we propose a novel annotation-guided algorithm called "**D**iscovery and **R**econstruction of **U**nannotated **T**ranscripts"(DRUT) [36] for transcriptome discovery, reconstruction and quantification in partially annotated genomes. DRUT incorporates VTEM algorithm to detect overexpressed segments corresponding to the unannotated transcripts and to estimate transcriptome frequencies. In case rVTEM algorithm is used, segments represent reads corresponding to unannotated transcripts. eVTEM algorithm requires one additional step, to select reads corresponding to overexpressed exons. Henceforth we will refer to these reads as overexpressed reads. Spliced read is selected only in the case when it entirely belongs to the "overexpressed" exons.

In this way we add the mapped reads to a new read alignment file (e.g., sam file) that represents a subset of original reads. This subset of reads is merged with reads that failed to map to annotated transcripts. Only reads that failed to map to annotated transcripts are now mapped to the reference genome using spliced alignment tools, e.g. TopHat[30] (see Fig. 3.2c). Merged subsets of reads are used as an input for transcriptome assembler. For DRUT framework we chose Cufflinks [3] as ab initio transcriptome reconstruction tool. Assembled transcripts are merged with annotated transcripts and the resulting set of transcripts is filtered to remove duplicates (see Fig. 3.2d). Finally DRUT reports full set of transcripts and maximum likelihood frequencies of transcripts that the best explain reads.

Algorithm 3 VTEM algorithm

```
add virtual transcript $vt$ to the set of annotated transcripts
initialize weights $h_{vs,j} = 0$
while $\Delta vt > \epsilon$ do
  calculate $f_j^{ML}$ by EM algorithm
  $e_j = \sum_i \frac{h_{t_i,j}}{\sum_l h_{t_i,l}} f_i^{ML}$
  $D = \frac{\sum_j |o_j - e_j|}{|S|}$
  $\delta = o_j - e_j$
  if $\delta > 0$ then
    $h_{vt,j} + = \delta$
  else
    $h_{vt,j} = \max\{0, h_{vt,j} + \delta\}$
  end if
end while
```

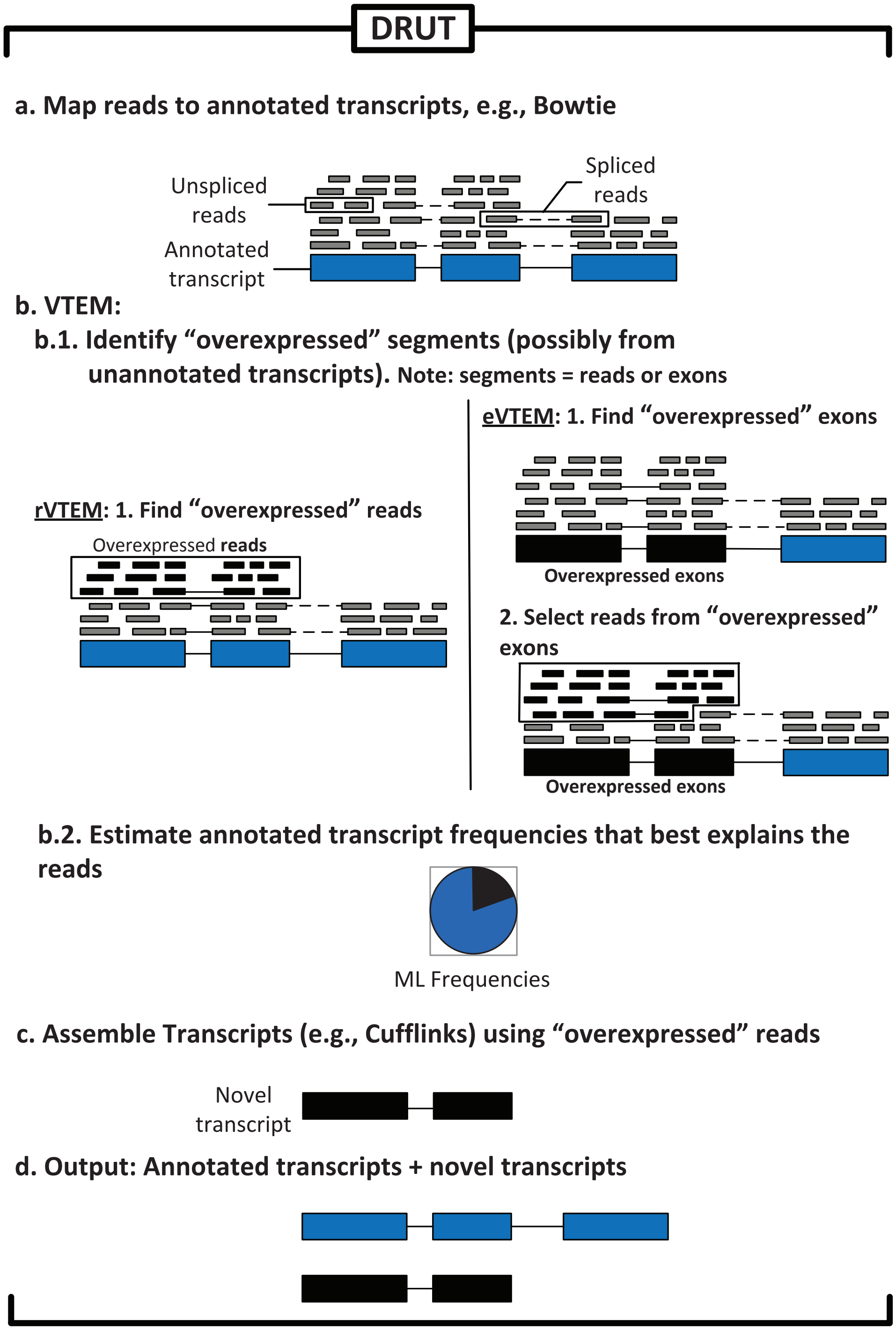

DRUT
a. Map reads to annotated transcripts, e.g., Bowtie
Unspliced reads
Spliced reads
Annotated transcript
b. VTEM:
b.1. Identify "overexpressed" segments (possibly from unannotated transcripts). Note: segments = reads or exons
rVTEM: 1. Find "overexpressed" reads
Overexpressed reads
eVTEM: 1. Find "overexpressed" exons
Overexpressed exons
2. Select reads from "overexpressed" exons
Overexpressed exons
b.2. Estimate annotated transcript frequencies that best explains the reads
ML Frequencies
c. Assemble Transcripts (e.g., Cufflinks) using "overexpressed" reads
Novel transcript
d. Output: Annotated transcripts + novel transcripts

Figure 3.2 Flowchart for DRUT.

3.2.4 Experiment Results.

Our validation of DRUT includes three experiments over human RNA-seq data, two experiments on transcriptome quantification and one experiment on transcriptome discovery and reconstruction. Below we describe the transcriptome data and read simulation and then give the settings for the each experiment and analyze the obtained experimental results.

Simulated human RNA-Seq data. The human genome data (hg19, NCBI build 36) was downloaded from UCSC [51] and CCDS [52], together with the coordinates of the transcripts in the KnownGenes table. The UCSC database contains a total of 66, 803 transcripts pertaining to 19, 372 genes, and CCDS database contains 20, 829 transcripts from 17, 373 genes. The transcript length distribution and the number of transcripts per genes for UCSC are shown in Fig. 3.10. Genes were defined as clusters of known transcripts as in GNFAtlas2 table, such that CCDS data set can be identified with the subset of UCSC data set. 30 millions single reads of length 25bp were randomly generated by sampling fragments of transcripts from UCSC data set. Each transcript was assigned a true frequency based on the abundance reported for the corresponding gene in the first human tissue of the GNFAtlas2 table, and a probability distribution over the transcripts inside a gene cluster [9]. We simulate datasets with geometric (p=0.5) distributions for the transcripts in each gene.

Single error-free reads of length 25bp, 50bp, 100bp and 200bp were randomly generated by sampling fragments of transcripts from UCSC data set. As shown in the [9] for transcriptome quantification purposes it is more beneficial to have shorter reads if the throughput is fixed. At the same time, for transcriptome reconstruction is quite beneficial to have longer reads. Read length of 100bp is the best available option for such next generation sequencing platform as Illumina$^{\text{TM}}$[19]. Current Ion Torrent$^{\text{TM}}$technology is capable of producing reads of length more than 200bp. Ion Torrent$^{\text{TM}}$next generation sequencing technology utilizes integrated circuits capable of detection ions produced by the template-directed DNA polymerase synthesis for sequencing genomes [20].

Accuracy Estimation Transcriptome Quantification Accuracy was assessed using *error fraction (EF)* and *median percent error (MPE)* measures used in [43]. However, accuracy was computed against true frequencies, not against estimates derived from the true counts as in [43]. If $\hat{f}_i$ is the frequency estimate for an transcript with true frequency f_i, the *relative error* is defined as $|\hat{f}_i - f_i|/f_i$ if $f_i \neq 0$, 0 if $\hat{f}_i = f_i = 0$, and ∞ if $\hat{f}_i > f_i = 0$. The error fraction with threshold τ, denoted EF_τ is defined as the percentage of transcripts with relative error greater or equal to τ. The median percent error, denoted MPE, is defined as the threshold τ for which $EF_\tau = 50\%$.

To estimate transcriptome reconstruction accuracy all assembled transcripts (referred to as "candidate transcripts") are matched against annotated transcripts. Two transcripts match if and only if they include the same set of exons. Two single-exon transcripts match if and only if the overlapping area is at least 50% the length of each transcript.

Following [53], we use sensitivity and Positive Predictive Value (PPV) to evaluate the performance of different methods. Sensitivity is defined as portion of the annotated transcript sequences being captured by candidate transcript sequences as follows:

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

PPV is defined portion of annotated transcript sequences among candidate sequences as follows:

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

Comparison on partially annotated UCSC data set. We assumed that in every gene 25% of transcripts are not annotated. In order to create such an instance we assign to the transcripts inside the gene a geometric distribution (p=0.5), assuming a priori that number of transcripts inside the gene is less or equal to 3, we will refer to this experiment as Experiment 1. This way we removed transcripts with frequency 0.25. As a result 11, 339 genes were filtered out, number of transcripts was reduced to 24, 099. Note that in our data set unannotated transcripts do not have unique exon-exon junctions that can emit reads indicating that certain transcripts are not annotated.

We first check how well VTEM estimates the volume of missing transcripts. Although the frequencies of all missing transcripts are the same (25%), the volumes significantly differ because they have different lengths. Therefore, the quality can be measured by correlation between actual unannotated volumes and predicted missing volumes, which represent volumes of virtual transcripts. In this experiment the quality is 61% which is sufficiently high to give an idea which genes have unannotated transcripts in the database.

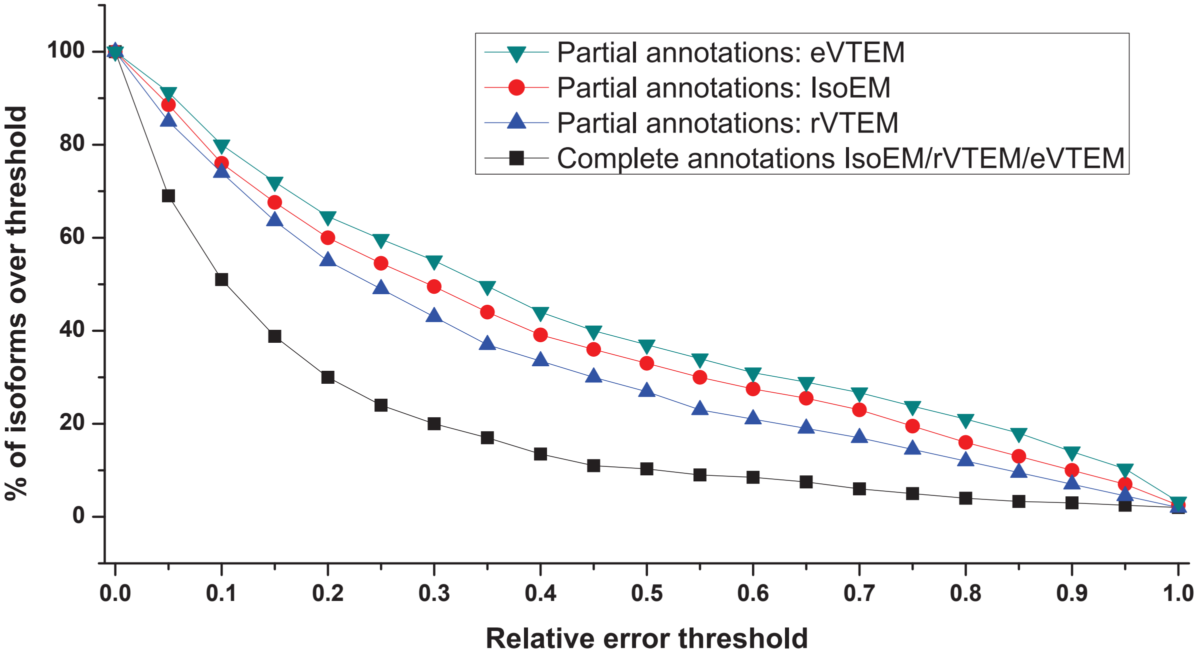


Figure 3.3 Error fraction at different thresholds for isoform expression levels inferred from 30 millions reads of length 25bp simulated assuming geometric isoform expression. Black line corresponds to IsoEM/VTEM with the complete panel, red line is IsoEM with the incomplete panel, blue line is rVTEM and the green line is eVTEM.

Table 3.2 reports the median percent error (MPE) and .15 error fraction EF$_{.15}$ for the isoform expression levels inferred from 30 millions reads of length 25bp, computed over groups of isoforms with various expression levels.

Figure 3.3 gives the error fraction at different thresholds ranging between 0 and 1. Clearly the best performance is achieved when the genome is completely annotated, in which case IsoEM and VTEM (rVTEM and eVTEM) show similar results. This happens due to the fact that the frequency of virtual transcript is not increasing over iterations of VTEM. In case of partial annotated genome using virtual transcript allows rVTEM to achieve better results comparative to IsoEM. eVTEM has worse performance than other methods, the reason is that it uses simplified model based on exons rather than on reads, as is done in IsoEM and rVTEM.

Table 3.2 Median percent error (MPE) and 15% error fraction (EF$_{.15}$) for isoform expression levels in Experiment 1.

	Expression range	0	$(0, 10^{-6}]$	$(10^{-6}, 10^{-5}]$	$(10^{-5}, 10^{-4}]$	$(10^{-4}, 10^{-3}]$	$(10^{-3}, 10^{-2}]$	All
MPE	**Complete annotations:** IsoEM, rVTEM, eVTEM	0.0	61.7	22.0	8.0	3.2	2.1	10.3
	Partial annotations:							
	IsoEM	0.0	59.3	41.3	24.8	19.7	5.9	33.7
	rVTEM	0.0	47.2	33.1	20.7	16.4	8.5	26.9
	eVTEM	0.0	60.5	45.1	25.2	22.1	9.1	35.3
EF$_{.15}$	**Complete annotations:** IsoEM, rVTEM, eVTEM	0.0	81.9	61.3	28.7	7.5	8.5	38.8
	Partial annotations:							
	IsoEM	0.0	81.7	72.4	61.4	56.7	42.1	67.6
	rVTEM	0.0	77.2	68.2	57.6	53.0	36.8	63.6
	eVTEM	0.0	82.8	75.6	64.7	59.2	44.4	70.1

Comparison on on CCDS data set. In this experiment, referred as Experiment 2, UCSC data set represents the complete set of transcripts and CCDS data set represents the partially annotated set of transcripts. Reads were generated from UCSC annotations, while only frequencies of the known transcripts from the CCDS database were estimated. In contrast to Experiment 1, we do not control the frequency of unannotated transcripts (i.e. transcripts from UCSC which are absent in CCDS). Therefore, one cannot expect as good improvements as in Experiment 1.

Table 3.3 reports the median percent error (MPE) and .15 error fraction $\mathrm{EF}_{.15}$ for isoform expression levels inferred from 30 millions reads of length 25bp, computed over groups of isoforms with various expression levels. We do not report the number of transcripts since they are different for UCSC and CCDS panels. Anyway, one can see a reasonable improvement in frequency estimation of rVTEM over IsoEM.

Table 3.3 Median percent error (MPE) and 15% error fraction ($\mathrm{EF}_{.15}$) for isoform expression levels in Experiment 2.

	Expression range	0	$(0, 10^{-6}]$	$(10^{-6}, 10^{-5}]$	$(10^{-5}, 10^{-4}]$	$(10^{-4}, 10^{-3}]$	$(10^{-3}, 10^{-2}]$	All
MPE	**Complete annotations:** IsoEM, rVTEM, eVTEM	0.0	100	22.7	7.3	3.5	2.5	11.8
	Partial annotations:							
	IsoEM	0.0	100	45.5	29.4	28.5	28.7	31.8
	rVTEM	0.0	100	43.2	27.1	25.7	14.3	29.6
	eVTEM	0.0	100	46.3	32.2	33.2	32.1	34.6
$\mathbf{EF}_{.15}$	**Complete annotations:** IsoEM, rVTEM, eVTEM	5.1	91.2	62.8	29.3	15.8	7.6	45.5
	Partial annotations:							
	IsoEM	18.6	95.6	85.6	83.3	89.2	86.7	80.0
	rVTEM	17.6	91.8	81.3	77.9	80.3	75.5	75.2
	eVTEM	19.5	97.4	89.2	87.7	88.3	87.9	82.3

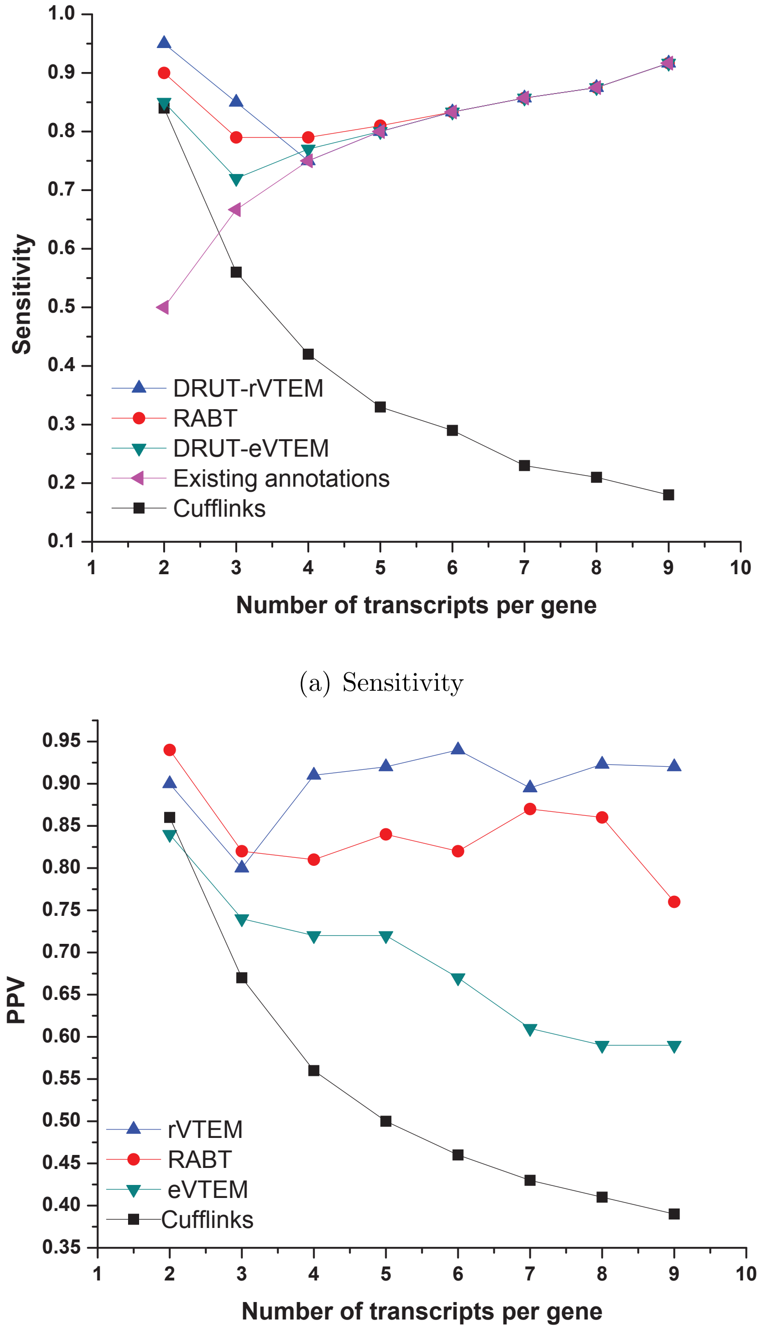
Comparison Between DRUT, RABT and Cufflinks. In order to simulate a partially annotated genome we removed from every gene exactly one transcript. As a result all 19, 372 genes become partially annotated, and number of transcripts was reduced to 47, 431. In this section, we use the sensitivity and PPV defined above to compare our DRUT method to the most recent version of Cufflinks and RABT (version 1.3.0 of Cufflinks and RABT downloaded from website http://cufflinks.cbcb.umd.edu/). Due to the fact that results on 100bp and 200bp are very similar, we decided to present comparison on reads of length 100bp. TopHap [30] is used for Cufflinks and RABT to map simulated reads to the reference genome. For DRUT we used Bowtie [48] to map reads to the set of annotated transcripts. For our simulation setup we assume perfect mapping of simulated reads to the genome in case of Cufflinks and to the annotated transcripts in case of DRUT.

Intuitively, it seems more difficult to predict the transcripts in genes with more transcripts. Following [54] we group all the genes by their number of transcripts and calculate the sensitivity and PPV of the methods on genes with certain number of transcripts as shown in Fig. 3.14.

Next we want to define the portion of known transcripts that is input for annotation-guided methods as “existing annotations”. Please note that sensitivity of annotation-guided methods needs to be compared to the “existing annotations” ratio unlike regular reconstruction methods that do not have any a priori information about annotated transcripts. In our simulation setup “existing annotations” ratio increases as the number of transcripts in genes become larger.

Fig. 3.14(a) shows that for genes with more transcripts it is more difficult to correctly reconstruct all the transcripts. As a result Cufflinks performs better on genes with few transcripts since annotations are not used in it standard settings. DRUT has higher sensitivity on genes with 2 and 3 transcripts, but RABT is better on gene with 4 transcripts. For genes with more than 4 transcripts performance of annotation-guided methods is equal to ”existing annotations ratio”, which means these methods are unable to reconstruct unannotated transcripts.

We compared PPV for all 3 methods (Fig. 3.14(b)), all methods show high PPV for genes with 2 transcripts. DRUT outperforms all methods on genes with more then 3 transcripts and shows comparable performance on gene with 2 and 3 transcripts.



(a) Sensitivity

(b) PPV

Figure 3.4 Comparison between DRUT, RABT, Cufflinks for groups of genes with n transcripts (n=1,...,9) : (a) Sensitivity (b) Positive Predictive Value (PPV)

3.3 Genome-guided Transcriptome Reconstruction Algorithms

3.3.1 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 [30] 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.

3.3.2 TRIP : *T*ranscriptome *R*econstruction using *I*nteger *P*rogramming

TRIP is a novel "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.

Construction of Splice Graph and Enumeration of Putative Transcripts. Typically, alternative variants occurs due alternative transcriptional events and alternative splicing events [55] . 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 3.5. Hence every gene has a set of non-overlapping pseudo-exons, from which it is possible to reconstruct a set of putative transcripts.

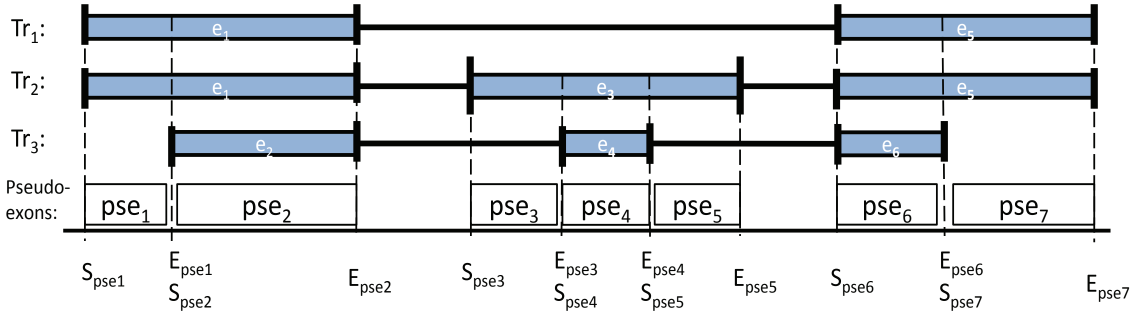


Figure 3.5 Pseudo-exons(white boxes) : regions of a gene between consecutive transcriptional or splicing events. An example of three transcripts $Tr_i, i = 1, 2, 3$ each sharing exons(blue boxes). S_{psej} and E_{psej} represent the starting and ending position of pseudo-exon j, respectively.

The notations used in Figure 3.5 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. 3.6), 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.

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

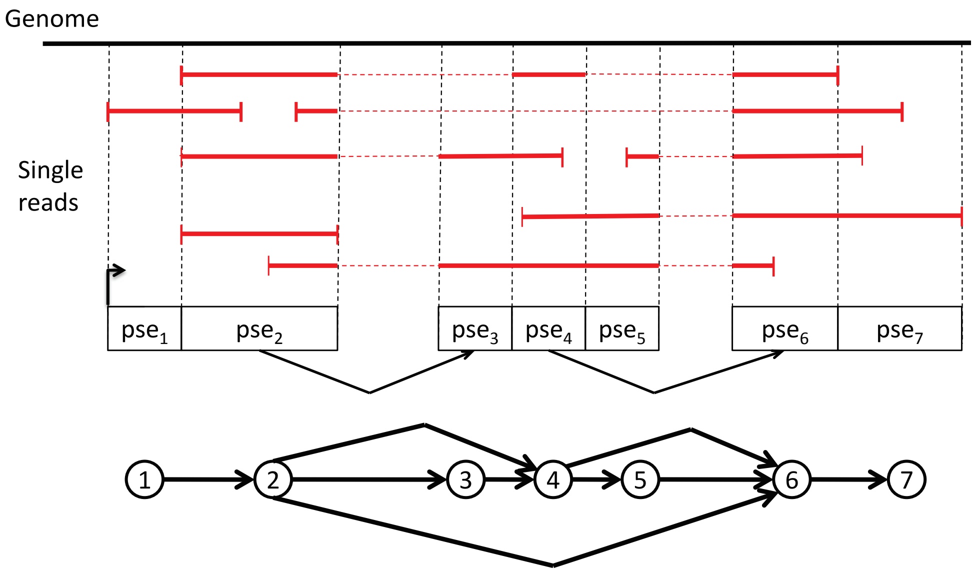


Figure 3.6 Splice graph. The red horizontal lines represent single reads. Reads interrupted by dashed lines are spliced reads. Each vertex of the splice graph corresponds to a pseudo-exon and each directed edge corresponds to a (splice) junction between two pseudo-exons.

Integer Program Formulation. The following notations are used 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) $\sum\limits_{t_k \in T_i(p)} y(t) \geq x_i(p), \forall p, i = \overline{1,4}$

(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) $\sum\limits_{t_k \in J_l} y(t) \geq 1, \forall J_l$

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 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.

3.3.3 MLIP : *M*aximum *L*ikelihood *I*nteger *P*rogramming

Here we present a genome guided method for transcriptome reconstruction from single-end RNA-Seq reads. Our method aims 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 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[20] . Therefore the primary goal of our study is to developed a method for longer single-end reads.

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

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 [55] . 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 3.7-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 3.7-B. Hence every gene has a set of non-overlapping pseudo-exons, from which it is possible to reconstruct a set of putative transcripts.

SG is a directed acyclic graph (see figure 3.7-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 3.7-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 departments on number of exons, this way splitting exons into pseudo-exons has no effect on number of candidate transcripts.

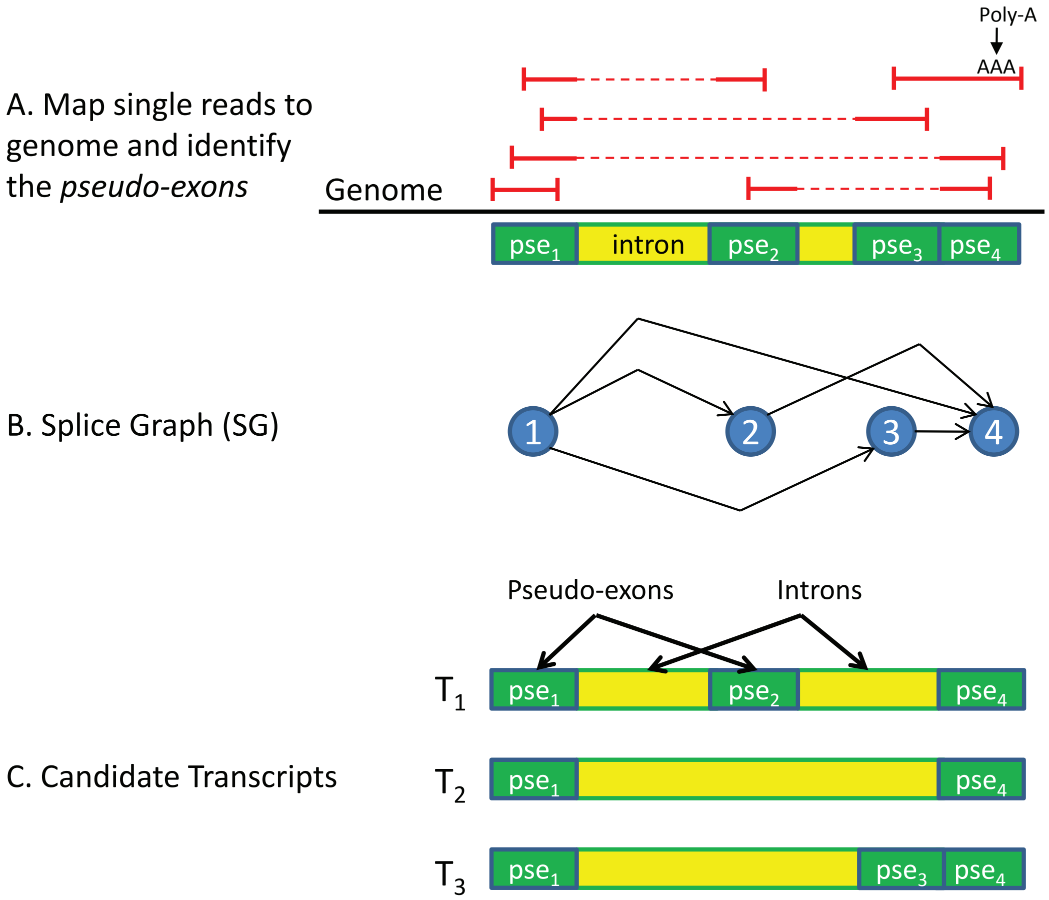


Figure 3.7 Model Description. A - Pseudo-exons. Pseudo-exons(green boxes) : regions of a gene between consecutive transcriptional or splicing events; B - Splice graph. The red horizontal lines represent single-end reads. Reads interrupted by dashed lines are spliced reads. Each vertex of the splice graph 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.

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 [56].

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[30] with default parameters in our experiments).

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: $\sum_{r\in R} x(r) \geq N(1-\epsilon)$

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

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

We use CPLEX [57] to solve the IP, the rest of implementation is done using Boost C++ Libraries and bash scripting language.

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 [9]. Regardless of initial conditions IsoEM algorithm always converge to maximum likelihood solution (see [23]).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:

$$D = \frac{\sum_j |o_j - e_j|}{|R|}, \tag{3.2}$$

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:

$$e_j = \sum_{r_j} \frac{h_{t_k,r_j}}{\sum_{r_j} h_{t_k,r_j}} f_t^{ML} \tag{3.3}$$

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 3.8.

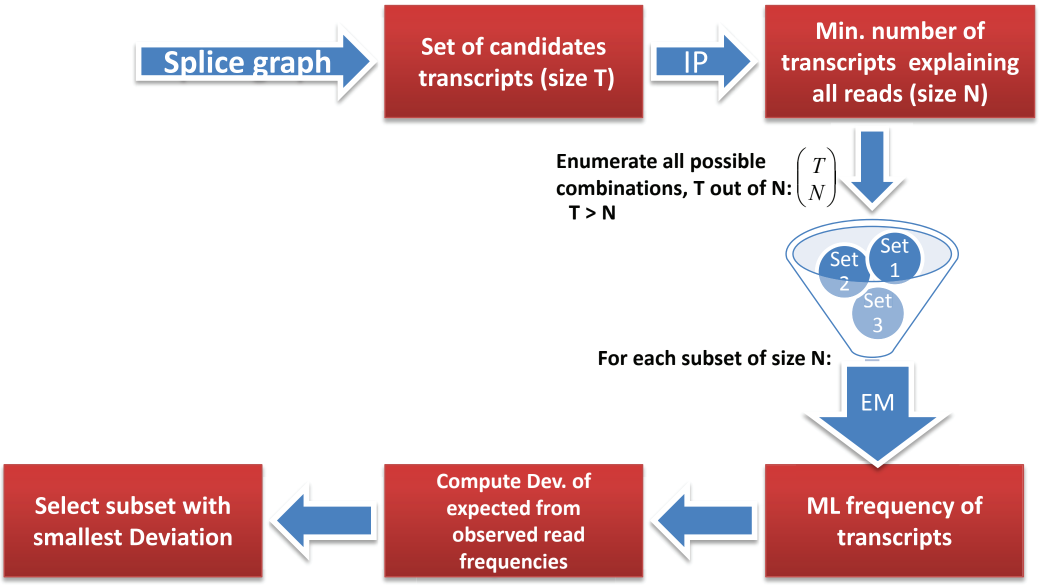
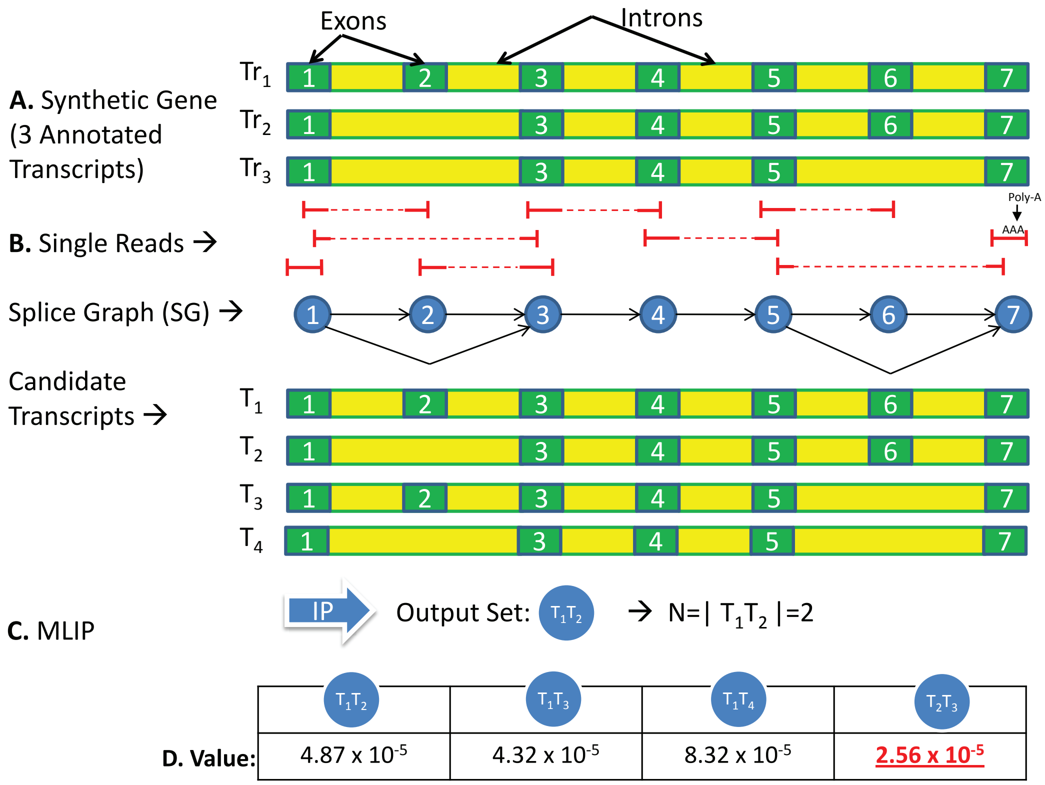

Splice graph
Set of candidates transcripts (size T)
IP
Min. number of transcripts explaining all reads (size N)
Enumerate all possible combinations, T out of N: $\binom{T}{N}$
T > N
Set 1
Set 2
Set 3
For each subset of size N:
EM
ML frequency of transcripts
Compute Dev. of expected from observed read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

Figure 3.8 Flowchart for MLIP. Input : Splice graph. Output: subset of candidate transcripts with the smallest deviation between observed and expected read frequencies.

Figure 3.9 illustrates how MLIP works on a given synthetic gene with 3 transcripts and 7 different exons (see figure 3.9-A). First we use mapped reads to construct the splice graph from which we generate T possible candidate transcripts, as shown in figure 3.9-B. Further we run our 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 will be excluded from consideration. The subset with the smallest deviation between expected and observed read frequencies is selected by the MLIP algorithm.



	T_1T_2	T_1T_3	T_1T_4	T_2T_3
D. Value:	4.87×10^{-5}	4.32×10^{-5}	8.32×10^{-5}	2.56×10^{-5}

Note that Set T_2T_4 and Set T_3T_4 are not chosen since those sets do NOT explain all spliced-junctions

Figure 3.9 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 MLIP will be excluded from consideration. The subset with the smallest deviation between expected and observed read frequencies is selected by the MLIP algorithm.

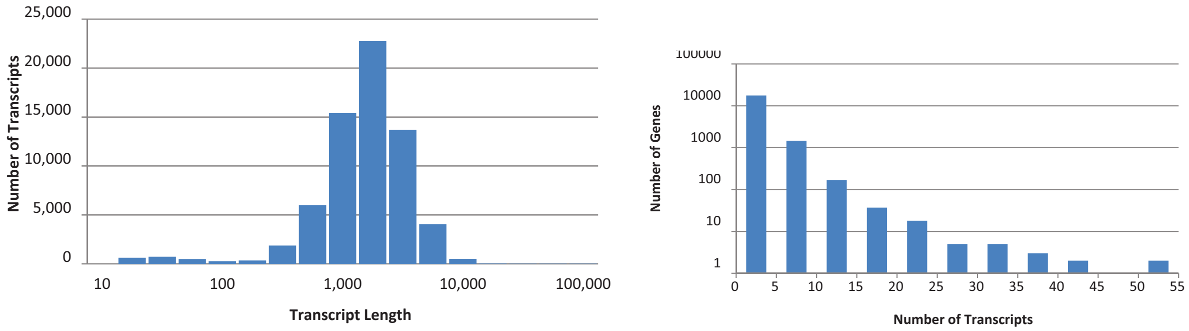
Stringency of Reconstruction. Different level of stringency corresponds to different strategies of transcriptome reconstruction. High stringency has the goal to optimize precision of reconstruction, with some loss in sensitivity. On the other hand, low stringency corresponds to increase in sensitivity 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 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.

3.3.4 Experimental Results

Simulation Setup. We first evaluated performance of TRIP and MLIP methods on simulated human RNA-Seq data. The human genome sequence (hg18, $NCBI$ build 36) was downloaded from $UCSC$ together with 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 and paired-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.



(a) (b)

Figure 3.10 Distribution of transcript lengths (a) and gene cluster sizes (b) in the UCSC dataset

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 [56].

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 [3] and [54].

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.,

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

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.,

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

Comparison Between TRIP and Cufflinks on Paired-End RNA-Seq Reads. 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 [54] 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 [56]) 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.

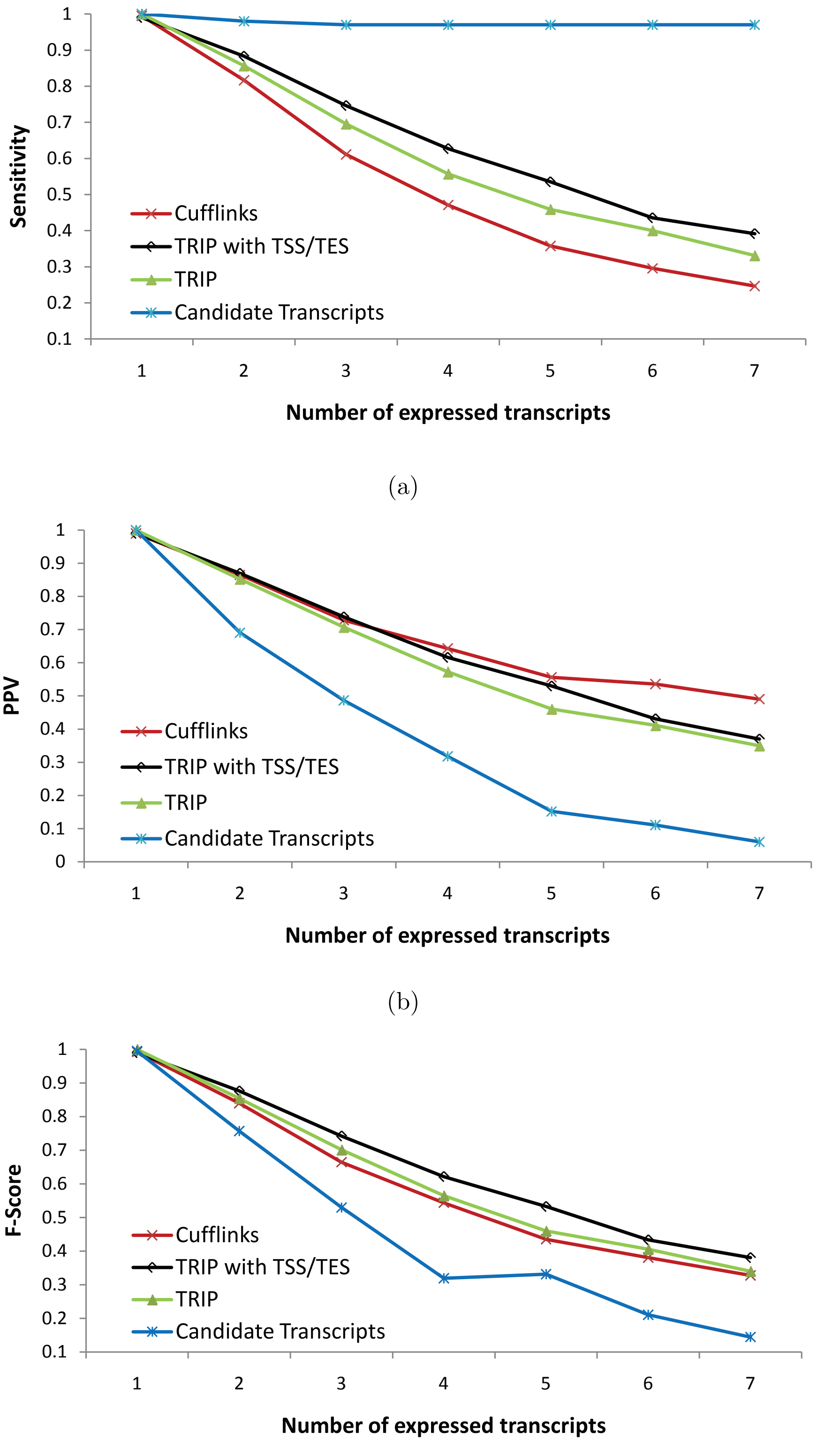
Figures 3.11(a)-3.11(c) compare the performance of 4 methods (Cufflinks, Candidate Transcripts, TRIP with and without TSS/TES) on simulated data with respect to number of transcripts per gene. Note that sensitivity (see Fig. 3.11(a)) for single-transcript genes is 100% for all methods and with the growth 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. 3.11(b) shows that Cufflinks has an advantage over TRIP in the portion of correctly predicted transcripts but overall comparison using F-score (see Fig. 3.11(c)) shows that TRIP improves over Cufflinks. Comparison of TRIP using known fragment length in the ILP formulation is represented by $TRIP - L$.

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 3.12(a)-3.12(c) compare sensitivity, 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.

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.

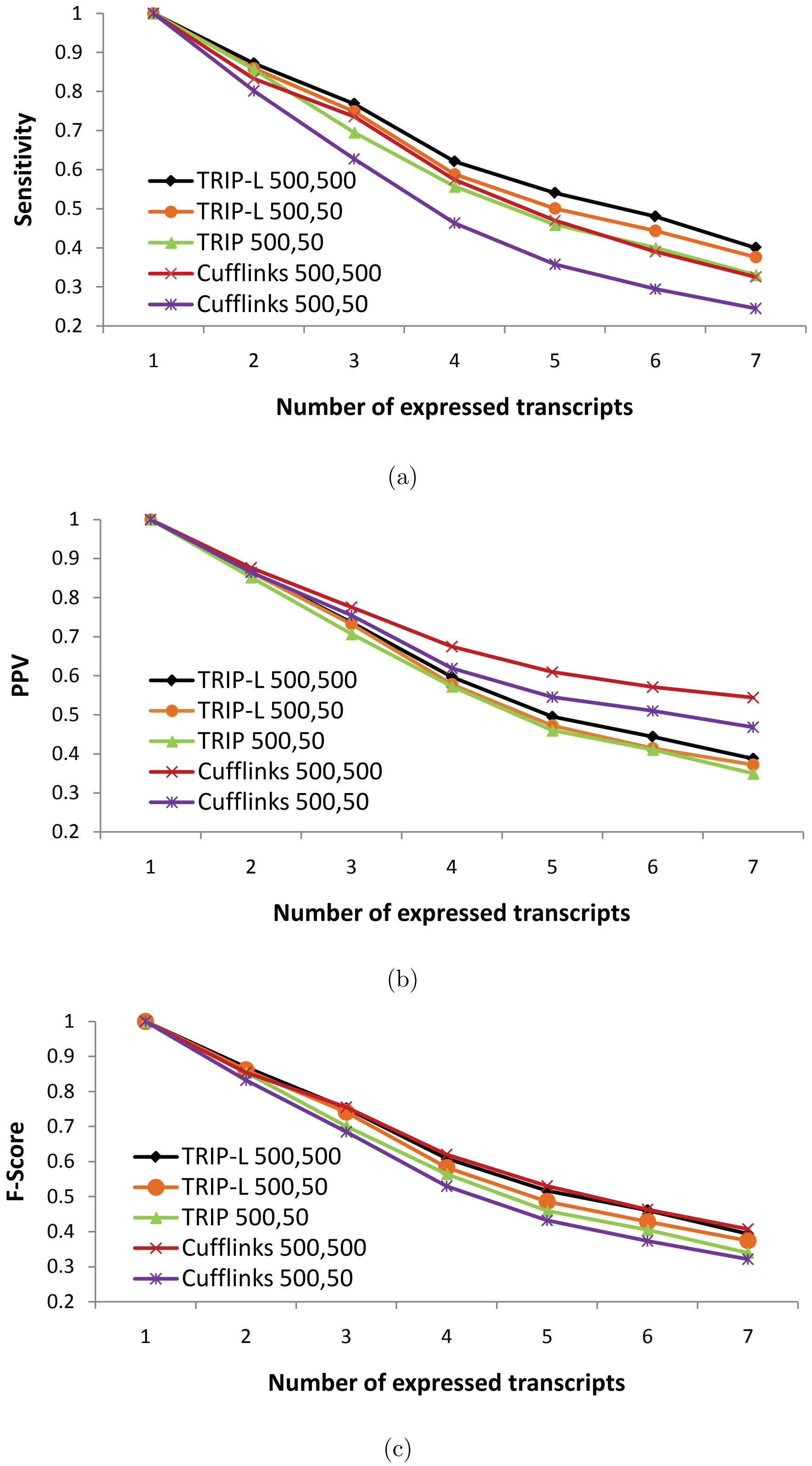


(a)

(b)

(c)

Figure 3.11 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 3.12 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.

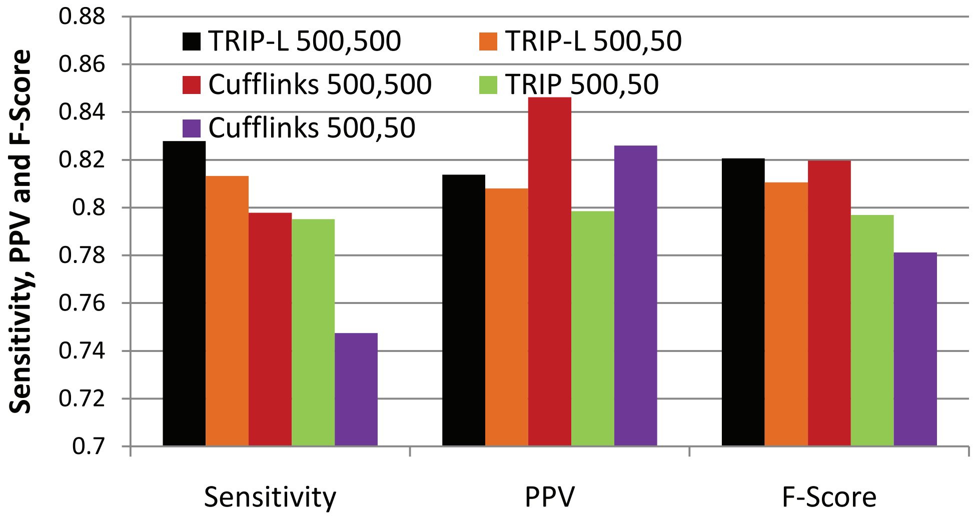


Figure 3.13 Overall Sensitivity, PPV and F-Score on simulated dataset with different sequencing parameters and distribution assumptions.

Comparison between MLIP, IsoLasso and Cufflinks on Single-End RNA-Seq Reads. 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 [3] and IsoLasso (v 2.6.0) from [54] are used for comparison. We explore the influence of read length, fragment length, and coverage on reconstruction accuracy.

Table 3.4 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.

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

Table 3.4 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.

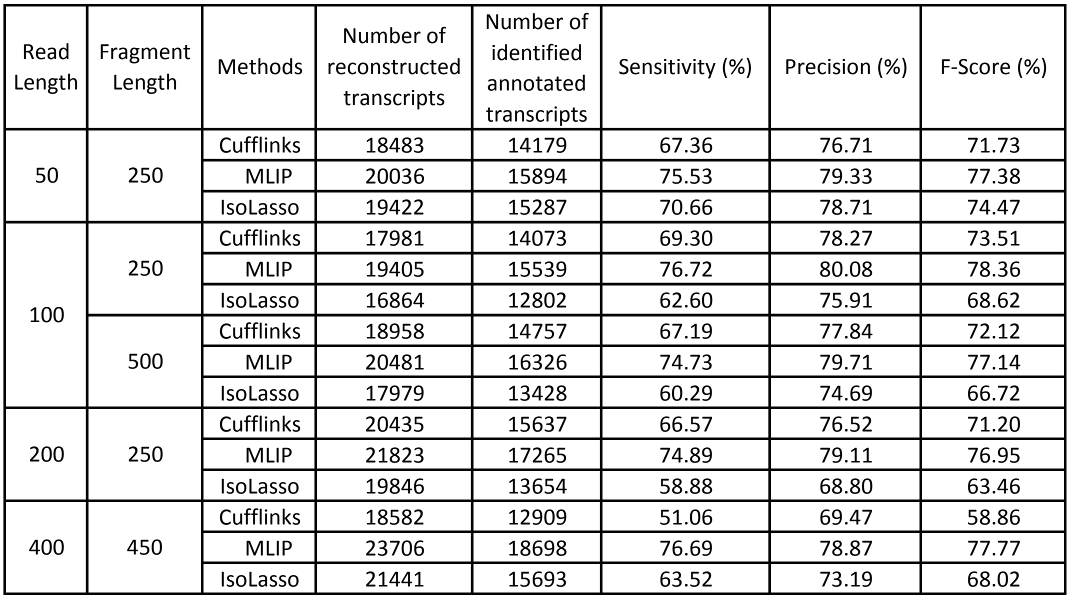
Isoform Distribution	Methods	Number of reconstructed transcripts	Number of identified annotated transcripts	Sensitivity (%)	Precision (%)	F-Score (%)
Uniform	Cufflinks	18582	12909	51.06	69.47	58.86
	MLIP	23706	18698	76.69	78.87	77.77
	IsoLasso	21441	15693	63.52	73.19	68.02
Geometric	Cufflinks	17377	12449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

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.

Table 3.5 compares sensitivity, precision and F-score of Cufflinks, IsoLasso, and MLIP for different combinations of read and fragment lengths: (50bp,250bp), (100bp,250bp),

Table 3.5 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).

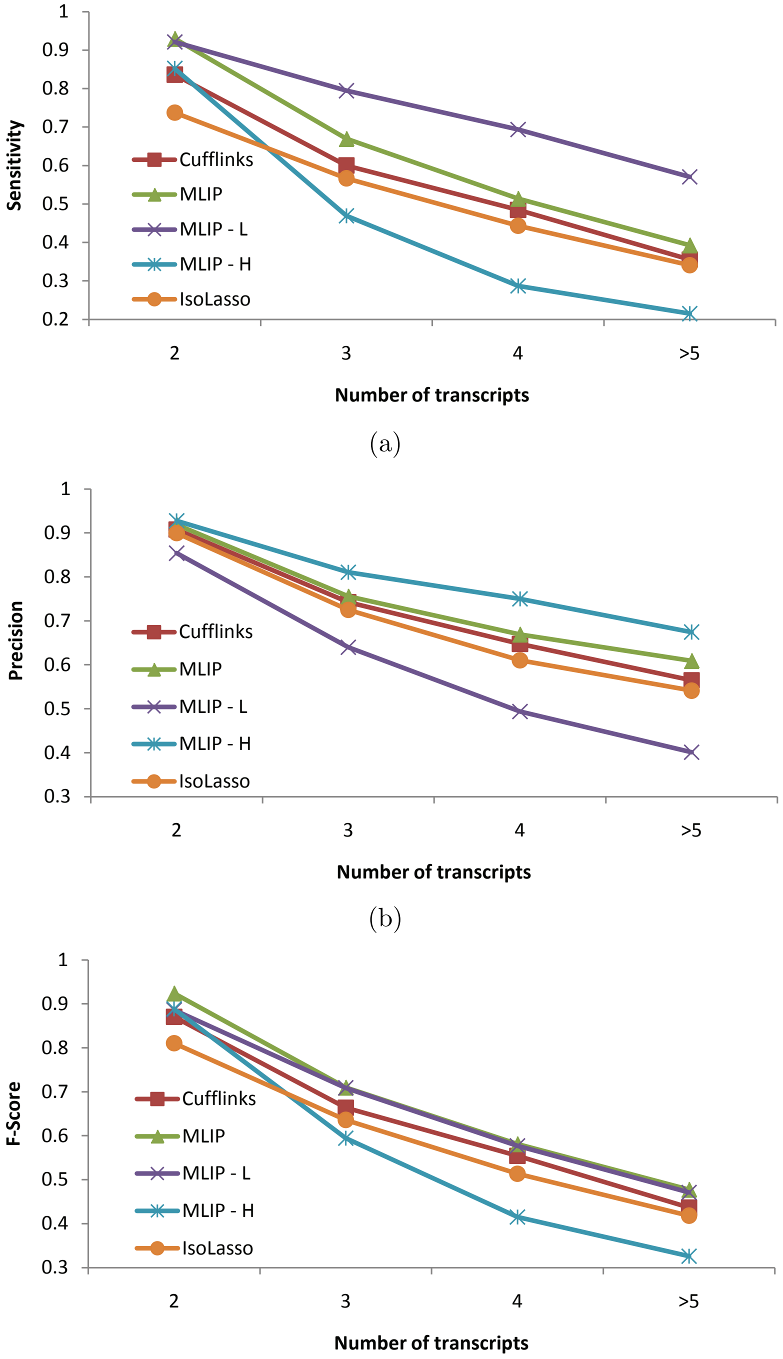


Read Length	Fragment Length	Methods	Number of reconstructed transcripts	Number of identified annotated transcripts	Sensitivity (%)	Precision (%)	F-Score (%)
50	250	Cufflinks	18483	14179	67.36	76.71	71.73
		MLIP	20036	15894	75.53	79.33	77.38
		IsoLasso	19422	15287	70.66	78.71	74.47
100	250	Cufflinks	17981	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

(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. Table 3.6 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.

Results on Real RNA-Seq Data. 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 settings(medium



(a)

(b)

(c)

Figure 3.14 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,..., ≥ 5) on simulated dataset with mean fragment length 250bp, standard deviation 25bp and read length of 100bp.

Table 3.6 Transcriptome reconstruction results with respect to different coverage. 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.

Coverage	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

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.

3.4 Conclusion

Here we have proposed two versions of DRUT, a novel annotation-guided method for transcriptome discovery, reconstruction and quantification in partially annotated genomes. Experiments on *in silico* RNA-Seq datasets confirm that DRUT overperforms existing genome-guided transcriptome assemblers and show similar or better performance with existing annotation-guided assemblers. We also tested DRUT as stand-alone method for transcriptome quantification in partially annotated data sets. Our experimental studies show that DRUT significantly improves the quality of the transcriptome quantification comparative to our previous approach IsoEM.

To address transcriptome reconstruction problem assisted by genome annotation we introduced novel genome-guided method for paired-end RNA-Seq read. 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 [58] 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. Also we introduce MLIP method for genome-guided transcriptome reconstruction from single-end 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. Experimental results on both real and synthetic datasets generated with various sequencing parameters and distribution assumptions show that both genome-guided methods are scalable and has increased transcriptome reconstruction accuracy compared to previous approaches.

PART 4

DISCUSSION AND FUTURE WORK

In ongoing work we are exploring possibility of integrating transcriptome quantification and transcriptome reconstruction that will possibly lead to quantification based reconstruction method. Currently, Next Generation Sequencing technologies allow to run library preparation step multiple times varying the fragment length distribution for every step. Additionally, it is possible to perform read barcoding for every library preparation step, which will produce reads with different fragment lengths. To take adventure of this technology we plan to develop the method able to handle reads from multiple libraries. We expect to improve reconstruction accuracy by integrating different fragment length distributions into transcriptome reconstruction algorithm. Also we are planning to release software tool for transcriptome quantification and reconstruction that will include all our methods.

REFERENCES

[1] H. Richard, M. H. Schulz, M. Sultan, A. Nurnberger, S. Schrinner, D. Balzereit, E. Dagand, A. Rasche, H. Lehrach, M. Vingron, S. Haas, and M.-L. Yaspo, "Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments," *Nucl. Acids Res.*, vol. 38, no. 10, pp. e112+, 2010. [Online]. Available: http://dx.doi.org/10.1093/nar/gkq041

[2] J. Bloom, Z. Khan, L. Kruglyak, M. Singh, and A. Caudy, "Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays," *BMC Genomics*, vol. 10, no. 1, p. 221, 2009. [Online]. Available: http://www.biomedcentral.com/1471-2164/10/221

[3] 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

[4] 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

[5] 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

[6] 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.

[7] 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

[8] 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

[9] 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

[10] J. F. Degner, J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard, "Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data." *Bioinformatics*, vol. 25, no. 24, pp. 3207–3212, 2009. [Online]. Available: http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics25.html#DegnerMPPNGP09

[11] C. Gregg, J. Zhang, J. Butler, D. Haig, and C. Dulac, "Sex-specific parent-of-origin allelic expression in the mouse brain," *Science*, vol. 329, no. 5992, pp. 682–685, 2010.

[12] C. McManus, J. Coolon, M. Duff, J. Eipper-Mains, B. Graveley, and P. Wittkopp, "Regulatory divergence in drosophila revealed by mrna-seq," *Genome research*, vol. 20, no. 6, pp. 816–825, 2010.

[13] J. Duitama, P. Srivastava, and I. Măndoiu, "Towards accurate detection and genotyp-

ing of expressed variants from whole transcriptome sequencing data," *BMC genomics*, vol. 13, no. Suppl 2, p. S6, 2012.

[14] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics." *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, 2009. [Online]. Available: http://dx.doi.org/10.1038/nrg2484

[15] 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.

[16] 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.

[17] V. Pandey, R. Nutter, and E. Prediger, "Applied biosystems solid? system: Ligation-based sequencing," *Next Generation Genome Sequencing: Towards Personalized Medicine*, pp. 29–42, 2008.

[18] R. Thomas, E. Nickerson, J. Simons, P. Jänne, T. Tengs, Y. Yuza, L. Garraway, T. LaFramboise, J. Lee, K. Shah *et al.*, "Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing," *Nature medicine*, vol. 12, no. 7, pp. 852–855, 2006.

[19] D. Bentley, S. Balasubramanian, H. Swerdlow, G. Smith, J. Milton, C. Brown, K. Hall, D. Evers, C. Barnes, H. Bignell *et al.*, "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, vol. 456, no. 7218, pp. 53–59, 2008.

[20] 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

[21] M. Griffith *et al.*, "Alternative expression analysis by RNA sequencing," *Nature Methods*, vol. 7, no. 10, pp. 843–847, 2010. [Online]. Available: http://dx.doi.org/10.1038/nmeth.1503

[22] C. Ponting and T. Belgard, "Transcribed dark matter: meaning or myth?" *Human Molecular Genetics*, August 2010. [Online]. Available: http://dx.doi.org/10.1093/hmg/ddq362

[23] 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.

[24] 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

[25] 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

[26] 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

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

[28] 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

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

[30] 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

[31] 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

[32] 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.

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

[34] 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.

[35] S. Mangul, I. Astrovskaya, M. Nicolae, B. Tork, I. Mandoiu, and A. Zelikovsky, “Maximum likelihood estimation of incomplete genomic spectrum from hts data,” in *Proc. 11th Workshop on Algorithms in Bioinformatics*, ser. Lecture Notes in Computer Science, September 5-7 2011. [Online]. Available: http://pbil.univ-lyon1.fr/members/sagot/htdocs/wabi2011/wabi2011.html

[36] 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.

[37] M. Anton, D. Gorostiaga, E. Guruceaga, V. Segura, P. Carmona-Saez, A. Pascual-Montano, R. Pio, L. Montuenga, and A. Rubio, “SPACE: an algorithm to predict and quantify alternatively spliced isoforms using microarrays,” *Genome Biology*, vol. 9, no. 2, p. R46, 2008. [Online]. Available: http://genomebiology.com/2008/9/2/R46

[38] Y. She, E. Hubbell, and H. Wang, “Resolving deconvolution ambiguity in gene alternative splicing,” *BMC Bioinformatics*, vol. 10, no. 1, p. 237, 2009. [Online]. Available: http://www.biomedcentral.com/1471-2105/10/237

[39] D. Hiller, H. Jiang, W. Xu, and W. Wong, “Identifiability of isoform deconvolution from junction arrays and RNA-Seq,” *Bioinformatics*, vol. 25, no. 23, pp. 3056–3059, 2009.

[40] V. Lacroix, M. Sammeth, R. Guigo, and A. Bergeron, “Exact transcriptome reconstruction from short sequence reads,” in *Proc. WABI*, 2008, pp. 50–63.

[41] A. Oshlack and M. Wakefield, “Transcript length bias in RNA-seq data confounds systems biology,” *Biology Direct*, vol. 4, no. 1, p. 14, 2009. [Online]. Available: http://www.biology-direct.com/content/4/1/14

[42] H. Jiang and W. Wong, “Statistical inferences for isoform expression in RNA-Seq,” *Bioinformatics*, vol. 25, no. 8, pp. 1026–1032, 2009. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btp113

[43] 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

[44] P. Carninci *et al.*, "The Transcriptional Landscape of the Mammalian Genome," *Science*, vol. 309, no. 5740, pp. 1559–1563, 2005. [Online]. Available: http://www.sciencemag.org/cgi/content/abstract/309/5740/1559

[45] G. Temple *et al.*, "The completion of the Mammalian Gene Collection (MGC)," *Genome Research*, vol. 19, no. 12, pp. 2324–2333, 2009. [Online]. Available: http://genome.cshlp.org/content/19/12/2324.abstract

[46] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley, "Continuous base identification for single-molecule nanopore DNA sequencing," *Nature Nanotechnology*, vol. 4, no. 4, pp. 265–270, 2009. [Online]. Available: http://dx.doi.org/10.1038/nnano.2009.12

[47] J. Eid *et al.*, "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009. [Online]. Available: http://dx.doi.org/10.1126/science.1162986

[48] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, p. R25, 2009. [Online]. Available: http://genomebiology.com/2009/10/3/R25

[49] K. D. Hansen, S. E. Brenner, and S. Dudoit, "Biases in Illumina transcriptome sequencing caused by random hexamer priming," *Nucl. Acids Res.*, vol. 38, no. 12, pp. e131+, 2010. [Online]. Available: http://dx.doi.org/10.1093/nar/gkq224

[50] M. Sultan *et al.*, "A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome," *Science*, vol. 321, no. 5891, pp. 956–960, 2008. [Online]. Available: http://www.sciencemag.org/cgi/content/abstract/321/5891/956

[51] UCSC Genome Database, http://genome.ucsc.edu.

[52] CCDS Genome Database, http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi.

[53] 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*, vol. 12, no. Suppl 6, p. S1, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/S6/S1

[54] 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.

[55] 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

[56] 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.

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

[58] 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*, vol. 475, no. 7356, pp. 348–352, 2011.