

Summer 8-12-2014

# Optimization Techniques For Next-Generation Sequencing Data Analysis

Adrian Caciula

Follow this and additional works at: [https://scholarworks.gsu.edu/cs\\_diss](https://scholarworks.gsu.edu/cs_diss)

---

## Recommended Citation

Caciula, Adrian, "Optimization Techniques For Next-Generation Sequencing Data Analysis." Dissertation, Georgia State University, 2014.  
[https://scholarworks.gsu.edu/cs\\_diss/87](https://scholarworks.gsu.edu/cs_diss/87)

This Dissertation is brought to you for free and open access by the Department of Computer Science at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Computer Science Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact [scholarworks@gsu.edu](mailto:scholarworks@gsu.edu).

# OPTIMIZATION TECHNIQUES FOR NEXT-GENERATION SEQUENCING DATA ANALYSIS

by

ADRIAN CACIULA

Under the Direction of Dr. Alexander Zelikovsky

## ABSTRACT

High-throughput RNA sequencing (RNA-Seq) is a popular cost-efficient technology with many medical and biological applications. This technology, however, presents a number of computational challenges in reconstructing full-length transcripts and accurately estimate their abundances across all cell types.

Our contributions include (1) transcript and gene expression level estimation methods, (2) methods for genome-guided and annotation-guided transcriptome reconstruction, and (3)

*de novo* assembly and annotation of real data sets. Transcript expression level estimation, also referred to as transcriptome quantification, tackle the problem of estimating the expression level of each transcript. Transcriptome quantification analysis is crucial to determine similar transcripts or unraveling gene functions and transcription regulation mechanisms. We propose a novel simulated regression based method for transcriptome frequency estimation from RNA-Seq reads. Transcriptome reconstruction refers to the problem of reconstructing the transcript sequences from the RNA-Seq data. We present genome-guided and annotation-guided transcriptome reconstruction methods. Empirical results on both synthetic and real RNA-seq datasets show that the proposed methods improve transcriptome quantification and reconstruction accuracy compared to currently state of the art methods. We further present the assembly and annotation of *Bugula neritina* transcriptome (a marine colonial animal), and *Tallapoosa* darter genome (a species-rich radiation freshwater fish).

INDEX WORDS: Transcriptome quantification, Regression, Transcriptome reconstruction, Alternative splicing, RNA-Seq, Assembly and annotation.

OPTIMIZATION TECHNIQUES FOR NEXT-GENERATION SEQUENCING DATA  
ANALYSIS

by

ADRIAN CACIULA

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

2014

Copyright by  
Adrian Caciula  
2014

OPTIMIZATION TECHNIQUES FOR NEXT-GENERATION SEQUENCING DATA  
ANALYSIS

by

ADRIAN CACIULA

Committee Chair: Alexander Zelikovsky

Committee: Yi Pan

Rajshekhar Sunderraman

Ion Mandoiu

Electronic Version Approved:

Office of Graduate Studies  
College of Arts and Sciences  
Georgia State University  
August 2014

## DEDICATION

I dedicate my dissertation to my son, Andrei. I hope this work will give him some motivation and encouragement to believe in himself, to work hard, to think over the limits and to follow his heart. All things are possible for those who believe.

## ACKNOWLEDGEMENTS

I would like to thank my adviser Dr. Alex Zelikovsky for his encouragement and constant support over my graduate studies at Georgia State University. I express my special gratitude to my graduate committee members: Dr. Raj Sunderraman, Dr. Yi Pan, and Dr. Ion Mandoiu (our co-adviser from University of Connecticut).

Special thanks to my colleagues and friends from Georgia State University for all their support and care: Serghei, Nick, Bassam, Blanche, Olga, Sasha, Igor, Sahar, Marco, Dinesh, Abi, Yang, Kebina, Sunny and all others who encouraged me throughout this journey. Thanks to the entire department of Computer Science, especially to: Dr. Xiaojun Cao, Mrs. Tammie Dudley, Mr. Shaochieh Ou, Mrs. Adrienne Martin, Mrs. Venette Rice, Mrs. Celena Pittman, and all professors and staff which contribute to my achievement.

Thanks to University of North Georgia, the place where I got my first faculty position and to Dr. Markus Hitz and Dr. Bryson Payne for sharing all their teaching experience.

I am also grateful to Transilvania University of Brasov and to all my undergrad professors. Thanks to all my undergrad classmates, which contribute to my achievement, especially to Bogdan, Razvan and Sorin. Thanks to all my friends who believed in me.

Thanks to my brother, Laurentiu, and my sister-in-law, Ersilia, for their continuous support. I will always appreciate all they have done for me. I am grateful to my parents Mariana and Ioan, whose words of encouragement and push for tenacity got me into this graduate program. Thanks to my parents/brother-in-law. I am also grateful to my grandmother Safta, aunt Mihaela, uncle Vio, cousin Diana, and aunt Ioana.

A special feeling of gratitude to my loving wife Manuela Cristina, for her incredible support, encouragement and patience.



## TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS</b> . . . . .	<b>v</b>
<b>LIST OF TABLES</b> . . . . .	<b>ix</b>
<b>LIST OF FIGURES</b> . . . . .	<b>x</b>
<b>LIST OF ABBREVIATIONS</b> . . . . .	<b>xiv</b>
<b>PART 1 INTRODUCTION</b> . . . . .	<b>1</b>
<b>1.1 High-Throughput Sequencing</b> . . . . .	<b>2</b>
<b>1.2 RNA-Seq protocol</b> . . . . .	<b>2</b>
1.2.1 Transcriptome Quantification . . . . .	4
1.2.2 Transcriptome Reconstruction . . . . .	4
<b>1.3 Contributions</b> . . . . .	<b>5</b>
<b>1.4 Future Work</b> . . . . .	<b>7</b>
<b>1.5 Organization</b> . . . . .	<b>7</b>
<b>1.6 Related Publications</b> . . . . .	<b>7</b>
<b>PART 2 TRANSCRIPTOME QUANTIFICATION</b> . . . . .	<b>12</b>
<b>2.1 Introduction</b> . . . . .	<b>12</b>
2.1.1 Background . . . . .	12
2.1.2 Related work . . . . .	13
2.1.3 Our contributions . . . . .	17
<b>2.2 SimReg : Simulated Regression Algorithm for Transcriptome         Quantification from RNA-Seq Data</b> . . . . .	<b>17</b>
2.2.1 Mapping RNA-Seq reads . . . . .	17
2.2.2 Partition reads into read classes . . . . .	18

2.2.3	Splitting the transcripts and reads into independent connected components.	18
2.2.4	Estimating transcript frequencies within each connected component.	20
2.2.5	Update initial estimates of transcript frequencies. . . . .	21
2.2.6	Combining transcript frequency estimates from all connected components.	22
<b>2.3</b>	<b>Experimental results . . . . .</b>	<b>23</b>
<b>2.4</b>	<b>Conclusions . . . . .</b>	<b>24</b>
<b>PART 3</b>	<b>TRANSCRIPTOME RECONSTRUCTION . . . . .</b>	<b>25</b>
<b>3.1</b>	<b>Introduction . . . . .</b>	<b>25</b>
3.1.1	Background . . . . .	25
3.1.2	Related Work . . . . .	26
3.1.3	Our Contribution . . . . .	27
<b>3.2</b>	<b>Annotation-guided Transcriptome Reconstruction Algorithms .</b>	<b>29</b>
3.2.1	Mapping RNA-Seq Reads and Exon Counts . . . . .	29
3.2.2	DRUT : Method for <i>Discovery</i> and <i>Reconstruction</i> of <i>Unannotated</i> Transcripts . . . . .	29
3.2.3	Experiment Results. . . . .	32
<b>3.3</b>	<b>Genome-guided Transcriptome Reconstruction Algorithms . . .</b>	<b>38</b>
3.3.1	Read Mapping . . . . .	38
3.3.2	MaLTA: Maximum Likelihood Transcriptome Assembly . . . . .	38
3.3.3	TRIP : <i>Transcriptome Reconstruction</i> using <i>Integer Programming</i>	39
3.3.4	MLIP : <i>Maximum Likelihood Integer Programming</i> . . . . .	43
3.3.5	Experimental Results . . . . .	51
<b>3.4</b>	<b>Conclusion . . . . .</b>	<b>60</b>
<b>PART 4</b>	<b><i>DE NOVO</i> ASSEMBLY AND ANNOTATION OF REAL DATA SETS. . . . .</b>	<b>64</b>
<b>4.1</b>	<b>Assembly of Illumina RNA-Seq Reads and Contig Annotation for Bugula neritina . . . . .</b>	<b>64</b>

4.1.1	Background . . . . .	64
4.1.2	Methods . . . . .	65
4.1.3	Assembly and annotation of <i>B. neritina</i> transcriptome sequences .	66
4.1.4	<i>Bugula neritina</i> Flows . . . . .	68
4.1.5	Analysis of results of each flow . . . . .	70
4.1.6	Sequence analysis of genes and C1b domains from <i>Bugula</i> species	72
4.1.7	Results . . . . .	72
4.1.8	Conclusions and Future work . . . . .	74
<b>4.2</b>	<b>Assembly and Annotation of the <i>Etheostoma tallapoosae</i> Genome</b>	<b>75</b>
4.2.1	Introduction . . . . .	75
4.2.2	Sequencing . . . . .	76
4.2.3	Assembly . . . . .	76
4.2.4	Utility of Assembly for Annotation . . . . .	78
4.2.5	Setting up WebApollo . . . . .	78
4.2.6	Tallapoosa Darter Genome Annotation with WebApollo . . . . .	80
<b>PART 5</b>	<b>SOFTWARE PACKAGES . . . . .</b>	<b>84</b>
<b>5.1</b>	<b>Transcriptome Quantification . . . . .</b>	<b>84</b>
5.1.1	SimReg . . . . .	84
<b>5.2</b>	<b>Transcriptome Reconstruction . . . . .</b>	<b>84</b>
5.2.1	MaLTA . . . . .	84
<b>5.3</b>	<b>Genome Assembly and Annotation . . . . .</b>	<b>84</b>
5.3.1	<i>Etheostoma tallapoosae</i> Genome . . . . .	84
<b>PART 6</b>	<b>DISCUSSION AND FUTURE WORK . . . . .</b>	<b>85</b>
<b>REFERENCES</b>	<b>. . . . .</b>	<b>86</b>

## LIST OF TABLES

Table 2.1	Comparison results between SimReg and RSEM . . . . .	24
Table 3.1	Classification of transcriptome reconstruction methods . . . . .	27
Table 3.2	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. . . . .	58
Table 3.3	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). . . . .	59
Table 3.4	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. . . . .	60
Table 4.1	Sharring Shallow . . . . .	69
Table 4.2	BlastX Results . . . . .	71
Table 4.3	PKC homologs identified by <i>Bugula neritina</i> transcriptome sequencing.	73

## LIST OF FIGURES

Figure 1.1	A schematic representation of the genome-guided RNA-Seq protocol.	3
Figure 2.1	Screenshot from Genome browser [1] . . . . .	13
Figure 2.2	Paired reads $r$ and $r'$ are simulated from the transcript $T1$ . Each read is mapped to all other transcripts ( $T2, T3, T4$ ). Mapping of the read $r$ into the transcript $T2$ is not valid since the fragment length is 4 standard deviations away from the mean. Then each read is assigned to the corresponding read class – the read $r$ is placed in the read class $T1\_T3$ and the read $r'$ is placed in the read class $T1\_T3\_T4$ . . . .	20
Figure 2.3	Screenshot from Genome browser [1] of a gene with 21 sub-transcripts	23
Figure 3.1	Flowchart for VTEM. . . . .	31
Figure 3.2	Flowchart for DRUT. . . . .	33
Figure 3.4	Comparison between DRUT, RABT, Cufflinks for groups of genes with $n$ transcripts ( $n=1, \dots, 9$ ) : (a) Sensitivity (b) Positive Predictive Value (PPV) . . . . .	37
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.	40
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. . . .	41

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. . . . .	45
Figure 3.8	Flowchart for MLIP. Input : Splice graph. Output: subset of candidate transcripts with the smallest deviation between observed and expected read frequencies. . . . .	49
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. . . . .	50
Figure 3.10	Distribution of transcript lengths (a) and gene cluster sizes (b) in the UCSC dataset . . . . .	52

Figure 3.11	Comparison between methods for groups of genes with $n$ transcripts ( $n=1,\dots,7$ ) 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. . . . .	55
Figure 3.12	Comparison between methods for groups of genes with $n$ transcripts ( $n=1,\dots,7$ ) on simulated dataset with different sequencing parameters and distribution assumptions: (a) Sensitivity (b) Positive Predictive Value (PPV) and (c) F-Score. . . . .	56
Figure 3.13	Overall Sensitivity, PPV and F-Score on simulated dataset with different sequencing parameters and distribution assumptions. . .	57
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. . . . .	62
Figure 3.14	Transcriptome reconstruction results with respect to number of transcripts per gene. Comparison between 5 methods (Cufflinks, IsoLasso, MLIP - medium stringency settings, <i>MLIP - L</i> - low stringency settings, <i>MLIP - H</i> - high stringency settings) for groups of genes with $n$ transcripts ( $n=1,\dots, \geq 5$ ) on simulated dataset with mean fragment length 250bp, standard deviation 25bp and read length of 100bp.	63
Figure 4.1	Screenshot from Metagenomics [2] . . . . .	68
Figure 4.2	Alignment of reads to previously cloned genomic fragments . . . .	77

Figure 4.3	Examples of instances where genes were identified within the scaffolds: (a) Urate Oxidase contained within one scaffold. (b) - (c) Neprilysin (NEP1) gene spanning several scaffolds . . . . .	79
Figure 4.4	WebApollo Work Flow . . . . .	81
Figure 4.5	Scaffold selection . . . . .	82
Figure 4.6	Gene model . . . . .	82
Figure 4.7	Amino Acid . . . . .	82
Figure 4.8	Annotation with WebApollo: (a) Homologous protein. (b) Gene prediction analysis (c) Gene adjustment . . . . .	83



## LIST OF ABBREVIATIONS

- NGS - Next Generation Sequencing
- GE - Gene Expression Level Estimation
- IE - Isoform Expression Level Estimation
- Q-PCR - Quantitative real-time Polymerase Chain Reaction.
- PKC - Protein kinase C

## 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 High-Throughput Sequencing

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” accuracy 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.

## 1.2 RNA-Seq protocol

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, or deep sequencing of RNAs, is a cost-efficient high-coverage powerful technology for transcriptome analysis [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

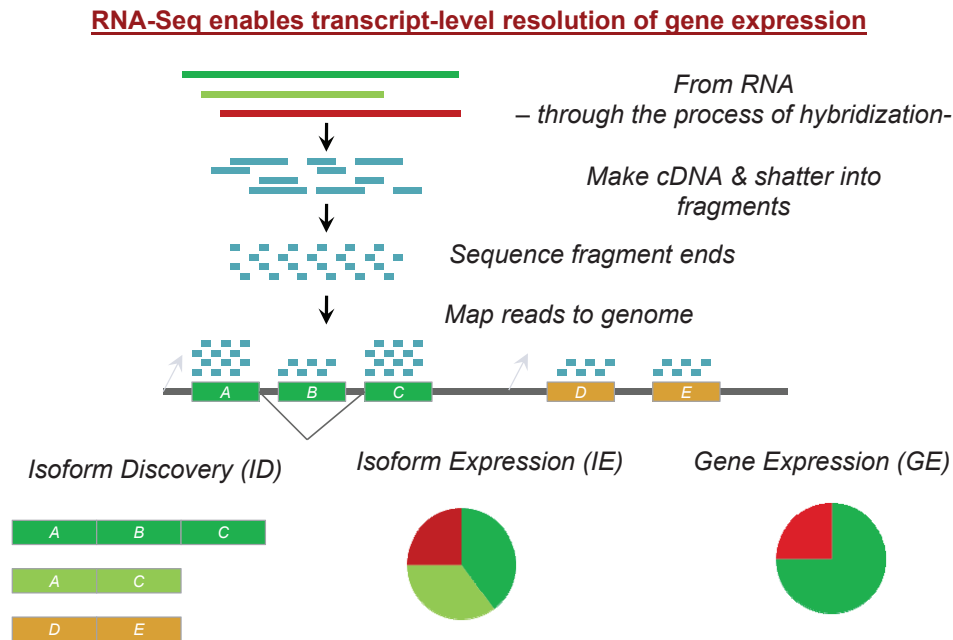


Figure 1.1 A schematic representation of the genome-guided RNA-Seq protocol.

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 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.

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 (i.e. estimate the expression level of each transcript), novel transcript discovery, and transcriptome reconstruction. A variety of new methods and tools have been recently developed to tackle these problems.

### 1.2.1 Transcriptome Quantification

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. 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.

### 1.2.2 Transcriptome Reconstruction

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 [23]. Genome-independent methods such as Trinity [24] or transAbyss [25] directly assemble reads into transcripts. A commonly

used approach for such methods is de Bruijn graph [26] 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 [27, 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.

### 1.3 Contributions

Our contributions include (1) transcript and gene expression level estimation methods, (2) methods for genome-guided and annotation-guided transcriptome reconstruction, and (3) *de novo* assembly and annotation of real data sets. In particular:

- *SimReg*: A novel **Simulated Regression** based algorithm for transcriptome quantification. To solve the problem of transcript and gene expression level estimation from RNA-Seq data, we propose *SimReg*, a Monte-Carlo simulated regression based method, that uses a more accurate simulation of read emission. Simulated data experiments demonstrate superior frequency estimation accuracy of *SimReg* comparatively to that of the existing tools.
- *DRUT*: "**D**iscovery and **R**econstruction of **U**nannotated **T**ranscripts" (DRUT) [29], a novel annotation-guided method for transcriptome discovery and reconstruction in partially annotated genomes. *DRUT* 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.
- Genome-guided transcriptome reconstruction methods:

*MaLTA*: **M**aximum **L**ikelihood **T**ranscriptome **A**ssembly, incorporates maximum likelihood model for candidate transcript expression estimation.

*TRIP* : “**T**ranscriptome **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.

*MLIP* : “ Maximum Likelihood Integer Programming ”. 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. 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.

- *De novo* assembly and annotation of real data sets:

Assembly of Illumina RNA-Seq Reads and Contig Annotation for *Bugula neritina*

Assembly and Annotation of the *Etheostoma tallapoosae* Genome

I am the leading contributor to the development of the transcriptome quantification method, *SimReg*. For the proposed reconstruction methods (2), I have contributed to all developmental stages but Serghei Mangul was the leading contributor. For the assembly

and annotation of real data sets (3), I had equal contribution in doing the bioinformatics analyses.

#### 1.4 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 advantage 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.

#### 1.5 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. Chapter 3 introduces transcriptome reconstruction problem and gives classification of existing methods. Chapter 4 presents *de novo* assembly and annotation of two real data sets. Discussion and future directions are provided in the Chapter 5.

#### 1.6 Related Publications

##### Refereed Journal Articles and Book Chapters

1. **A. Caciula**, O. Glebova, A. Artyomenko, S. Mangul, J. Lindsay, I. Mandoiu and A. Zelikovsky, Simulated Regression Algorithm for Transcriptome Quantification from



- RNA-Seq Data, *BMC Bioinformatics*, 2014, invited.
2. S. Mangul, S. Al Seesi, **A. Caciula**, D. Brinza, I.I. Mandoiu, and A. Zelikovsky, "Transcriptome Assembly and Quantification from Ion Torrent RNA-Seq Data", *BMC Genomics* 15 (Suppl 5), pp. S7, 2014
  3. M. Mathew, K. I Bean, Y. Tiagueu, **A. Caciula**, I. I Mandoiu, A. Zelikovsky and N. B Lopanik Importance of symbiont-produced bioactive natural products to holobiont fitness, *BMC Biology*, 2014, submitted.
  4. S. Al Seesi, S. Mangul, **A. Caciula**, A. Zelikovsky and I.I. Mandoiu, "Transcriptome reconstruction and quantification from RNA sequencing data", In Maria Poptsova, *Genome Analysis: Current Procedures and Applications*, Caister Academic Press, pp. 39-60, 2014
  5. S. Mangul, **A. Caciula**, O. Glebova, I. Mandoiu and A. Zelikovsky, "Improved Transcriptome Quantification and Reconstruction from RNA-Seq Reads using Partial Annotations", *In Silico Biology(ISB) 11 : An International Journal on Computational Molecular Biology*, pp. 251-261, 2012.

#### **Refereed Conference Articles**

6. 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", *Proc. 5th International Conference on Bioinformatics and Computational Biology (BICoB)*, pp. 25-34, 2013.
7. 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)*, pp. 369-376, 2012.

8. 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.

#### **Workshop Articles, Conference Abstracts, and Posters**

9. **A. Caciula**, O. Glebova, A. Artyomenko, S. Mangul, J. Lindsay, I. Mandoiu, and A. Zelikovsky, Simulated Regression Algorithm for Transcriptome Quantification, *10th International Symposium on Bioinformatics Research and Applications (ISBRA)*, pp. 405, 2014.
10. **A. Caciula**, O. Glebova, A. Artyomenko, S. Mangul, J. Lindsay, I. Mandoiu, and A. Zelikovsky, "Deterministic Regression Algorithm for Transcriptome Frequency Estimation", *Proc. 4th Workshop on Computational Advances for Next Generation Sequencing (CANGS 2014)*.
11. L. G. Kral, **A. Caciula**, Y. B. Temate, A. Zelikovsky, "Assembly and Annotation of the Etheostoma tallapoosae Genome", *Plant and Animal Genome XXII (PAG 2014)*
12. Y. B. Temate-Tiagueu, S. Al Seesi, **A. Caciula**, M. Mathew, O. Glebova, I. Mandoiu, N. B. Lopanik and A. Zelikovsky, "Bioinformatics Analysis of RNA-Seq Data for Bugula Neritina", *Proc. 4th IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2014)*
13. **A. Caciula**, S. Mangul, I. Mandoiu and A. Zelikovsky, "Transcriptome Reconstruction from Single RNA-Seq Reads Using Expectation Maximization Algorithm with Expected Deviation Minimization Enhancement", *9th International Symposium on Bioinformatics Research and Applications (ISBRA 2013)*
14. S. Mangul, S. Al Seesi, **A. Caciula**, D. Brinza, I. Mandoiu and A. Zelikovsky, "Transcriptome Assembly and Quantification from Ion Torrent RNA-Seq Data", *Proc.*

*3rd Workshop on Computational Advances for Next Generation Sequencing (CANGS 2013)*

15. 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)*
16. 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**
17. 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**
18. 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 8th International Symposium on Bioinformatics Research and Applications (ISBRA 2012)*, Dallas, TX, **best poster award**
19. 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
20. 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

21. 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
22. 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
23. 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

## 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, 21]. 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 [30, 31], the fragmentary information provided by splicing arrays is typically insufficient for unambiguous identification

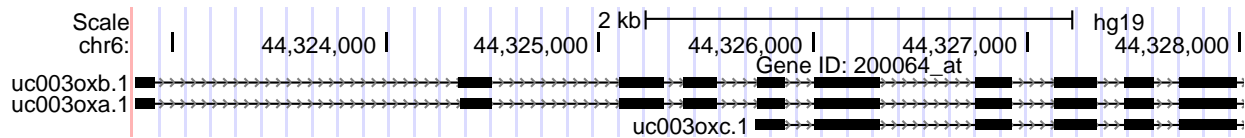


Figure 2.1 Screenshot from Genome browser [1]

of full-length transcripts [32, 33]. 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.

Recent review of computational methods for transcriptome quantification from RNA-Seq data reports several problems with the current state of transcriptome quantification, among them a significant variation in expressions level distributions throughout transcriptome reconstruction and quantification tools [34]. Transcriptome quantification from RNA-Seq data highly depends on read depth. Due to the sparse read support at some loci, many tools fail to report all/some of the exons or exon-intron junctions.

Improving isoform frequency estimation error rate is critical for detecting similar transcripts or unraveling gene functions and transcription regulation mechanisms, especially in those cases when one isoform is a subset of another. Figure 2.1 shows a gene with sub-transcripts from human genome (hg19).

### 2.1.2 Related work

From optimization point of view, the variety of approaches to transcriptome quantification and reconstruction is very wide. The most popular approach is maximizing likelihood using different variants of expectation-maximization (EM) [9, 35, 36], integer linear program (LP)

based methods [6, 37], min-cost flow [38], and regression [39].

RNA-Seq by Expectation Maximization (RSEM) is an Expectation-Maximization (EM) algorithm that works on the isoform level. The initial version of RSEM only handled single-end reads, however, the latest version [35] has been extended to support paired-end reads, variable-length reads, and incorporates fragment length distribution and quality scores in its modeling. In addition to the maximum likelihood estimates of isoform expressions, RSEM also calculates 95% confidence intervals and posterior mean estimates. RSEM is the best algorithm presented so far, so we compare our tool SimReg to RSEM in Results and Discussion section.

The main limitation of statistically-sound EM approach is that it does not include uniformity of transcript coverage, i.e., it is not clear how to make sure that a solution with more uniform coverage of transcripts will be preferred to the one where coverage is volatile. LP and integer LP based methods overcome this limitation but cannot handle many isoforms simultaneously.

More recently, the authors of [36] proposed a quasi-multinomial model with a single parameter to capture positional, sequence and mapping biases. Tomescu et al. [40] proposed a method based on network flows for a multiassembly problem arising from transcript identification and quantification with RNA-Seq. This approach is good at keeping overall uniformity coverage but is not suitable for likelihood maximization.

Regression based approaches are the most related to the proposed method. The most representative of these is IsoLasso approach [39]. IsoLasso mathematically model a gene partitions into segments (a segment is a consecutive exon region while a subexon is a non-spliced region).

IsoLasso approach also assumes reads being uniformly sampled from transcripts. The Poisson distribution [41] then used to approximate the binomial distribution for the number of reads falling into each segment or subexon. The following quadratic program [39] is well-known as a LASSO approach [42]:

$$\begin{aligned} \text{minimize:} \quad & \sum_{i=1}^M \left( \frac{r_i}{l_i} - \sum_{j=1}^N a_{ji} x_j \right)^2 \\ & (2.1) \end{aligned}$$

$$\text{subject to: } x_j \geq 0, 1 \leq j \leq N, \sum_{j=1}^N x_j \leq \lambda, \forall t = 1 \dots |T|$$

and two more “completeness” constraints (namely that each segment or junction with mapped reads is covered by at least one isoform; and the sum of expression levels of all isoforms that contain this segment or junction should be strictly positive[39]) were added to this program in IsoLasso. The main over-simplification is an assumption that each segment receives from containing transcripts the number of reads proportional to its length. For example, it is not clear how to handle very short subexons and take in account position of a subexon in a transcript. Fragment length distribution also can discriminate one subexon from another. Especially difficult to accurately estimate portions of pair-end reads emitted from each subexon since in fact such reads are frequently emitted by multiple subexons collectively. Furthermore, mapping of the reads into transcripts is frequently ambiguous which is consciously ignored in [39].

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 [43].

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 [43].

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, [33] 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 [44], further complicating the design of unbiased differential expression tests based on RNA-Seq data. To overcome this difficulty, [41] 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. [45] introduced an expectation-maximization (EM) algorithm similar to that of [43] but applied to isoforms instead of genes. Unlike the method of [41], which estimates isoform frequencies only from reads that map to a unique location in the genome, the algorithm of [45] incorporates multireads as well. The IE problem for single reads is also tackled in [46], 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 [47] and [48] 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 [49, 50], 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 [43].

## 2.2 SimReg : Simulated Regression Algorithm for Transcriptome Quantification from RNA-Seq Data

The proposed method for estimating frequencies of transcripts is based on the novel approach for estimating expected read frequencies. First we describe the essence of our approach and contrast it with IsoLasso.

### 2.2.1 Mapping RNA-Seq reads

As with many RNA-Seq analyses, the first step of SimReg 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 [51] 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 [52], 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.

### 2.2.2 Partition reads into read classes

As discussed above, it is very difficult (if at all possible) to accurately estimate portions of pair-end reads emitted from each subexon. Instead, rather than distinguishing reads by their gene position, we partition reads into *classes* each consisting of reads consistent with each element of a particular subset of transcripts. In other words, two reads are assigned to the same class if they are consistent with exactly the same transcripts. Our second innovation is to use Monte-Carlo simulations instead of attempting to formally estimate contributions of each transcript to each read class. For any particular read class  $R$ , the expected frequency is estimated based on the frequencies of contributing transcripts as well as portions of reads that fall into the class  $R$ . Finally, using the standard regression method, we estimate transcript frequencies by minimizing deviation between expected and observed read class frequencies.

### 2.2.3 Splitting the transcripts and reads into independent connected components.

We assume that alignment of a read to transcript is valid if the fragment length deviates from the mean by less than 4 standard deviations. Our simulations show that the Monte-Carlo estimates become accurate enough only when simulated coverage is sufficiently high, i.e., approaching 1000x. Such high coverage is time consuming since each simulated read needs to be aligned with each possible transcript. In order to reduce runtime, we split transcripts into small related subsets roughly corresponding to sets of overlapping genes. First, we build the matching graph  $M = (\mathcal{T} \cup \mathcal{R}, E)$ , where  $\mathcal{T}$  and  $\mathcal{R}$  are the sets of all transcripts and reads, respectively, and each edge  $e = (r, T) \in E$  corresponds to a valid alignment of a read  $r$  to a transcript  $T \in \mathcal{T}$ . Transcript frequencies within each connected

---

**Algorithm 1** SimReg Algorithm

---

- 1. Split transcripts and reads into independent connected components:**
    - Estimate mean  $\mu$  and standard deviation  $\sigma$  of read fragment distribution
    - Find valid alignment of all observed reads to all transcripts
    - Construct matching graph  $M = (\mathcal{T} \cup \mathcal{R}, E)$  with edges corresponding to valid alignments
    - Find connected components of  $M$
    - Find observed read classes  $R$ 's in  $\mathcal{R}$
  - 2. Estimate transcript frequencies inside each connected component:**
    - for** each component  $C$  of  $M$  **do**
      - for** each transcript  $T$  in  $C$  **do**
        - Simulate reads with 1000x coverage from  $T$
        - Map simulated reads to all other transcripts in  $C$
        - Find simulated read classes from reads mapped to the same subset of transcripts in  $C$
        - Find  $D_{\mathcal{R},T} = \{d_{R,T}\}$ , distribution of reads simulated from  $T$  between read classes in  $\mathcal{R}$
      - end for**
      - Combine observed read classes and simulated read classes
      - Find crude transcript frequencies  $F'_T$  in  $C$  minimizing deviation between observed read class frequencies  $O_{\mathcal{R}} = \{o_R\}$  and expected read class frequencies
        - $F'_T \leftarrow \arg \min (D_{\mathcal{R},T} \times F'_T - O_{\mathcal{R}})^2$
    - end for**
  - 3. Update initial estimates of transcript frequencies:**
    - for** each component  $C$  of  $M$  **do**
      - Initilize aimed read class frequencies  $A = \{a_R\}$  with observed frequencies:  $a_R \leftarrow o_R$
      - repeat**
        - For  $i = 0, \dots$
        - Simulate reads with 100x coverage based on crude transcript frequency  $F'_T$ 
          - $s_R \leftarrow$  simulated frequency of read class  $R$
        - Compute deviations between observed and simulated read class frequencies
          - $\Delta \leftarrow S - O$
        - Update aimed read class frequencies  $a_R$  –
          - $A \leftarrow A_{\mathcal{R}} - \Delta/2$
        - Compute crude transcript frequencies  $F'_T$  based on corrected read class frequencies  $\{c_R\}$ , i.e.,
          - $F'_T \leftarrow \arg \min (D_{\mathcal{R},T} \times F'_T - A_{\mathcal{R}})^2$
      - until**  $\Delta^2 < \epsilon$
      - Obtain transcript frequencies from crude transcript frequencies
    - end for**
  - 4. Combine transcript frequency estimates from all connected components**
-

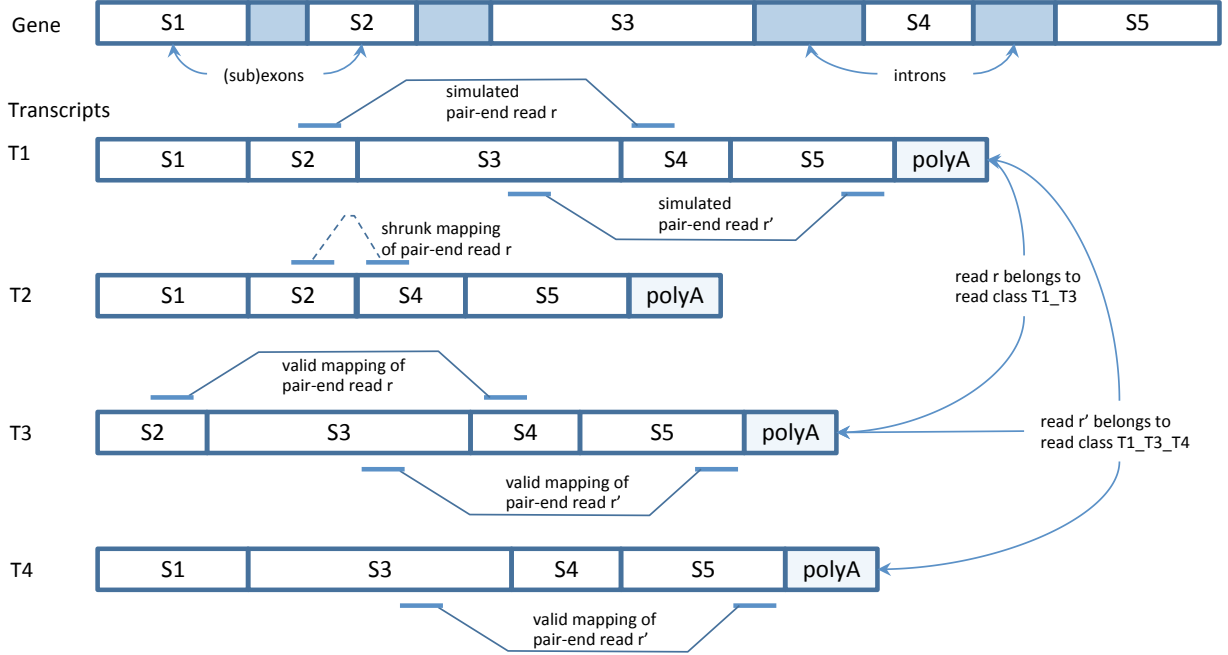


Figure 2.2 Paired reads  $r$  and  $r'$  are simulated from the transcript  $T1$ . Each read is mapped to all other transcripts ( $T2, T3, T4$ ). Mapping of the read  $r$  into the transcript  $T2$  is not valid since the fragment length is 4 standard deviations away from the mean. Then each read is assigned to the corresponding read class – the read  $r$  is placed in the read class  $T1\_T3$  and the read  $r'$  is placed in the read class  $T1\_T3\_T4$ .

component of  $M$  do not depend on transcript frequencies within other connected components and can be estimated separately. A significant portion of connected components contains just a single transcript for which the next step is trivial. Finally, the observed reads are partitioned into read classes each consisting of reads mapped to the same transcripts (see Figure 2.2).

#### 2.2.4 Estimating transcript frequencies within each connected component.

As discussed above, in each connected component  $C$  we simulate reads with  $1000x$  coverage for each transcript (see Figure 2.2). Thus for a transcript  $T$  with the length  $|T|$  we generate  $N_T = 1000l_T$  reads, where  $l_T = |T| - \mu + 1$  is the adjusted length of  $T$ . Similar to observed reads, we allow only alignments with fragment length less than  $4\sigma$  away from  $\mu$ . The reads that belong to exactly the same transcripts are collapsed into a single read

class. Let  $\mathcal{R} = \{R\}$  be all read classes found in the connected component  $C$  and let  $R_T$  be the number of reads simulated from the transcript  $T$  that fall in the read class  $R$ . The first inner loop outputs the set of coefficients  $D_{\mathcal{R},\mathcal{T}} = \{d_{R,T}\}$ , where  $d_{R,T}$  is the portion of reads generated from  $T$  belonging to  $R$

$$D_{\mathcal{R},\mathcal{T}} = \left\{ \frac{|R_T|}{N_T} \right\}$$

Let  $F'_{\mathcal{T}} = \{f'_T\}$  be the *crude* transcript frequency, i.e., the portions of reads emitted by transcripts in the connected component  $C$ . Then the expected read class frequency  $E_{\mathcal{R}}$  can be estimated as

$$E_{\mathcal{R}} = D_{\mathcal{R},\mathcal{T}} \times F'_{\mathcal{T}} \quad (2.2)$$

Regression-based estimation of  $f'_t$ 's minimizes squared deviation

$$(D_{\mathcal{R},\mathcal{T}} \times F'_{\mathcal{T}} - O_{\mathcal{R}})^2 = \sum_{R \in \mathcal{R}} (e_R - o_R)^2 \quad (2.3)$$

between expected read class frequencies  $e_R$ 's and observed read class frequencies  $o_R$ 's. Minimizing (2.3) is equivalent to the following quadratic program that can be solved with any constrained quadratic programming solver.

$$\begin{aligned} &\text{minimize:} && \sum_{R \in \mathcal{R}} \left( \sum_{T \in C} d_{R,T} f'_T - o_R \right)^2 \\ &\text{subject to:} && \sum_{T \in C} f'_T = 1 \text{ and } f'_T \geq 0, \forall T \in C \end{aligned} \quad (2.4)$$

### 2.2.5 Update initial estimates of transcript frequencies.

The obtained crude transcript frequency estimation  $F'_{\mathcal{T}}$  can deviate from the true crude frequency since the minimization of deviation is done uniformly. Indeed, the deviation in frequency is minimized on the same scale for each read class while different read classes have different size, as well as contribute to different subsets of transcripts. Instead of

estimating unknown coefficients, we propose to directly obtain  $F'_T$  for which simulated read class frequencies  $S_R = \{s_R\}$  match the observed frequencies  $O_R$  accurately enough as follows.

Until the deviation between simulated and observed read class frequencies is small enough, we repeatedly

- simulate reads according to  $F'_R$ ,
- find deviation between simulated and observed reads,  $\Delta_R = S_R - O_R$ ,
- obtain read frequencies  $C_R = O_R - \Delta_R/2$  corrected half-way in the direction opposite to the deviation
- update estimated crude transcript frequencies  $F'_T$  based on corrected read class frequencies  $\{C_R\}$

Finally, the transcript frequencies  $f_T$ 's can be obtained from crude frequencies  $f'_T$ 's as follows

$$f_T = \frac{f'_T/l_T}{\sum_{T' \in C} f'_{T'}/l_{T'}} \quad (2.5)$$

### 2.2.6 Combining transcript frequency estimates from all connected components.

Finally, we combine together individual solutions for each connected component. Let  $f_T^{glob}$  and  $f_T^{loc}$  be the global frequency of the transcript  $T$  and local frequency of the transcript  $T$  in its connected component  $C$ . Then the global frequency can be computed as follows

$$f_T^{glob} = f_T^{loc} \times \frac{|R_C| / \sum_{T' \in C} f_{T'}^{loc} l_{T'}}{\sum_{C' \in \mathcal{C}} \frac{|R_{C'}|}{\sum_{T' \in C'} f_{T'}^{loc} l_{T'}}} \quad (2.6)$$

where  $\mathcal{C}$  is the set of all connected components in the graph  $M$ ,  $|R_C|$  is the number of reads emitted by the transcripts contained in the connected component  $C$ .

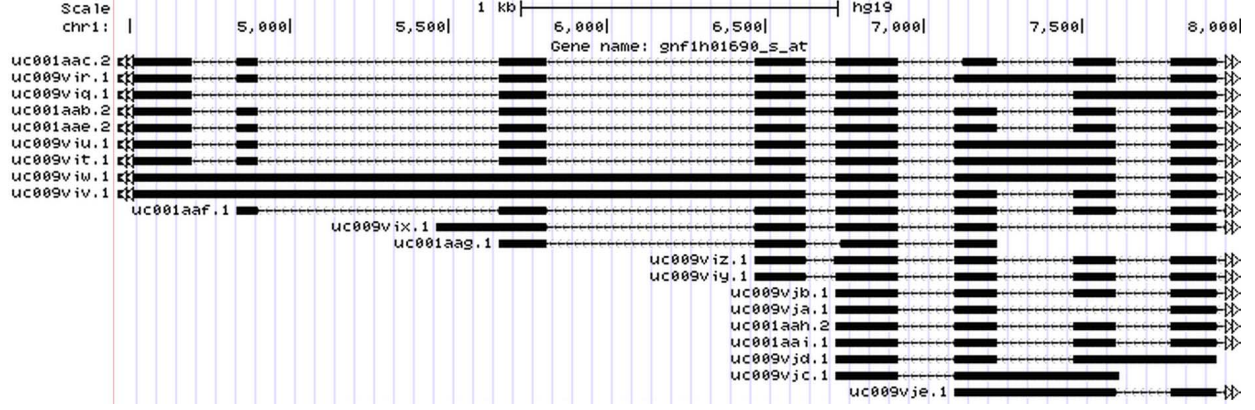


Figure 2.3 Screenshot from Genome browser [1] of a gene with 21 sub-transcripts

## 2.3 Experimental results

We tested *SimReg* on several test cases using simulated human RNA-Seq data. The RNA-Seq data was simulated from UCSC annotation (hg18 Build 36.1) using Grinder read simulator (version 0.5.0) [53], with a uniform 0.1% error rate. Experiments on synthetic RNA-seq datasets show that the proposed method improves transcriptome quantification accuracy compared to previous methods.

The following three test cases have been used to validate *SimReg*:

*Case 1:* consists of a single gene with 21 transcripts extracted from chromosome 1 (see Figure 2.3). From this gene we have simulated around 3000 (coverage 100 $\times$ ) paired-end reads of length 100bp and mean fragment length  $\mu = 300$ .

*Case 2:* we have randomly chosen 100 genes from which we have simulated reads using same parameters as in case 1.

*Case 3:* we have run our tool on the entire chromosome 1 which contains a total of 5509 transcripts (from 1990 genes) from where we have simulated 10M paired-end reads of length 100bp.

We have compared our results with *RSEM*, one of the best tool for transcriptome quantification. Frequency estimation accuracy was assessed using  $r^2$  and the comparison results are presented in Table 1. The results show better correlation compared with *RSEM*



especially because of those cases of sub-transcripts where *RSEM* skewed the estimated frequency toward super-transcripts.

Table 2.1 Comparison results between SimReg and RSEM

Isoform Expression - $r^2$ values					
Case 1: 1 gene		Case 2: 100 genes		Case 3: chr. 1	
SimReg	RSEM	SimReg	RSEM	SimReg	RSEM
0.958	0.923	0.999	0.93	0.995	0.924

SimReg is freely available at <http://alan.cs.gsu.edu/NGS/?q=adrian/simreg>

## 2.4 Conclusions

We propose a novel regression based algorithm to solve the problem of transcript and gene expression level estimation from RNA-Seq data. Our novel algorithm falls into the category of regression based methods: namely, SimReg is a Monte-Carlo based regression method. We propose to apply a more accurate simulation of read emission. The results on several simulated datasets show better correlation compared with *RSEM* especially because of those cases of sub-transcripts where *RSEM* skewed the estimated frequency toward super-transcripts. For the real dataset a subset of human transcripts was used, where transcripts were quantified independently by NanoString assay [?] (a total of 109 genes were targeted by 141 distinct probes). *MCREg2* reports a correlation of 0.8 showing a better performance than *RSEM* which reports only 0.75 correlation. However, for this particular dataset, the *IsoEM* performance is of overall best (0.85).

## 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, 45]. 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 [23]. 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 [52] or SpliceMap [54].

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 [23]. Genome-independent methods such as Trinity [24] or transAbyss [25] directly assemble reads into transcripts. A commonly used approach for such methods is de Bruijn graph [26] 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 [27] 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 [52] or SpliceMap [54]. 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.

Table 3.1 Classification of transcriptome reconstruction methods

Method	Support paired-end reads	Consider fragment length 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

The method of Trapnell et al. [3, 55], 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 [42] 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. CLIQ [37] 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 4.8 we propose a novel annotation-guided general framework for transcriptome discovery, reconstruction and quantification in partially annotated genomes, referred as

**Discovery and Reconstruction of Unannotated Transcripts (DRUT).** DRUT framework incorporates an enhancement of EM algorithm, VTEM [56] [29], 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 [27] 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 [56] algorithm for partially annotated transcripts. Our experimental studies show that DRUT significantly improves estimation of transcripts 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 [51] 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 [52], 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 DRUT : Method for *Discovery* and *Reconstruction* of *Unannotated Transcripts*

In this section, we propose a novel annotation-guided algorithm called "**D**iscovery and **R**econstruction of **U**nannotated **T**ranscripts" (DRUT) [29] for transcriptome discovery, reconstruction and quantification in partially annotated genomes.

In this section we first formally define the problem and briefly describe expectation-maximization method for transcriptome quantification, referred as IsoEM [9] . Then we show how to estimate the quality of the model, i.e. how well model explains the relationship between transcripts and emitted exons. Finally we describe the DRUT method that is based on modification of **V**irtual **S**tring **E**xpectation **M**aximization(VSEM) Algorithm [56].

The input data for EM method consists of a *panel*, i.e., a bipartite graph  $G = \{S \cup R, E\}$  such that each string is represented as a vertex  $s \in S$ , and each read is represented as a vertex  $r \in R$ . With each vertex  $s \in S$ , we associate unknown frequency  $f_s$  of the string. And with each vertex  $r \in R$ , we associate observed read frequency  $o_r$ . Then for each pair  $s_i, r_j$ , we add an edge  $(s_i, r_j)$  weighted by probability of string  $s_i$  to emit read  $r_j$ .

Regardless of initial conditions EM algorithm always converge to maximum likelihood solution (see [43]). The algorithm starts with the set of  $N$  strings. After uniform initialization of frequencies  $f_s, s \in S$ , the algorithm repeatedly performs the next two steps until convergence:

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

In this modification of VSEM algorithm, refereed as **Virtual Transcript Expectation Maximization(VSET)** algorithm , we replace the reads in the panel by corresponding exons with the observed counts(calculated as described in ??). In order to decide if the panel is incomplete we need to measure how well maximum likelihood model explains the exon counts. We suggest to measure the model quality by the deviation between expected and observed exon counts as follows:

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

where  $|R|$  is number of exons,  $o_j$  is the observed exon count  $E_j$  and  $e_j$  is the expected exon count  $r_j$  calculated as follows:

$$e_j = \sum_i \frac{h_{s_i,j}}{\sum_l h_{s_i,l}} f_i^{ML} \quad (3.1)$$

where  $h_{s_i,j} = \{0 - \text{exon } E_j \text{ doesn't belong to transcript } t_i, 1 - \text{otherwise}\}$ , and  $f_j^{ML}$  is the maximum-likelihood frequency of the transcript  $s_i$ .

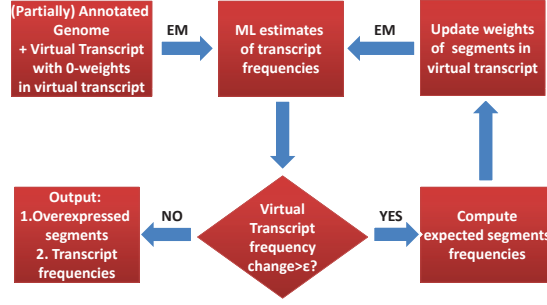


Figure 3.1 Flowchart for VTEM.

The DRUT method is based on our VTEM method, described above, that is created for maximum likelihood estimation of incomplete genomic spectrum. The main idea of DRUT algorithm (see Algorithm 2) is to add to the set of annotated transcripts a virtual transcript which emits exons that do not fit well to annotated transcript sequences. The flowchart of DRUT is on Fig. 3.1. Initially, all exons are connected to the virtual string with weight  $h_{s_i,j} = 0$ . The first iteration finds the ML frequency estimations of annotated transcripts, ML frequency estimations of virtual transcript will be equal to 0, since all edges between virtual transcript and exons  $h_{vs,j} = 0$ . Then these estimation are used to compute expected frequency of the exons according to (3.3). If the expected exon frequency is less than the observed one (under-estimated), then the lack of the exon expression is added to the weight of the read connection to the virtual transcript. For over-estimated exons, the excess of exon expression is subtracted from the corresponding weight (but keeping it non-negative). The iterations are continued while the virtual string frequency is decreasing by more than  $\epsilon$ .

The frequency  $f_i$  of virtual transcript estimates the total frequency of unannotated transcript. Therefore, based on the frequency of virtual transcript we can decide if the panel is likely to be incomplete or not. Furthermore the output of DRUT besides estimated frequency of the virtual transcript also contain the weights of edges connecting exons to the virtual transcript. These weights can be interpreted as probabilities of exon to be part by the unannotated transcripts. In order to select exons corresponding to unannotated transcripts



---

**Algorithm 2** 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
   $e_j = \sum_i \frac{h_{si,j}}{\sum_l h_{si,l}} f_i^{ML}$ 
   $D = \frac{\sum_j |o_j - e_j|}{|R|}$ 
   $\delta = o_j - e_j$ 
  if  $\delta > 0$  then
     $h_{vt,j} += \delta$ 
  else
     $h_{vt,j} = \max\{0, h_{vt,j} + \delta\}$ 
  end if
end while

```

---

it is enough to select  $f_i$  most probable exons. Let's call these exons "*overexpressed*" (see Fig. 3.2b).

After "*overexpressed*" exons corresponding to unannotated transcripts were detected, it becomes possible to select reads corresponding to these exons. Spliced reads are selected only in the case when all spliced parts belong to "*overexpressed*" exons. This way we add the reads to a new read file that represents a subset of original reads. Also, this subset of reads is merged with reads that failed to map to annotated transcripts a priori, mapping these reads to the reference genome with spliced alignment tool e.g. TopHat[52] (see Fig. 3.2c). Merged subset of reads are used as an input for transcriptome assembler. For our DRUT method we choose ab initio transcriptome reconstruction tools - Cufflinks [3]. Assembled transcripts are merged with annotated transcripts and the resulting set of transcripts is filtered to remove duplicates (see Fig. 3.2d).

### 3.2.3 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

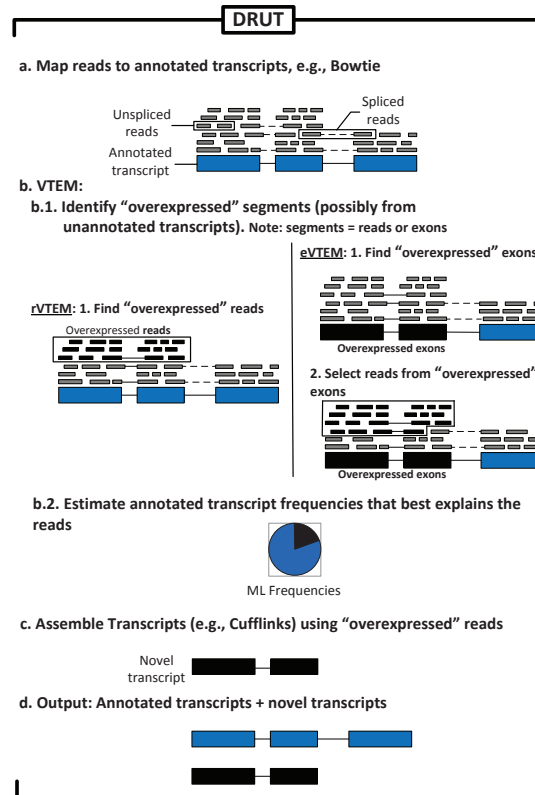


Figure 3.2 Flowchart for DRUT.

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 [57] and CCDS [58], 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 GNFAAtlas2 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<sup>TM</sup>[19]. Current Ion Torrent<sup>TM</sup> technology is capable of producing reads of length more than 200bp. Ion Torrent<sup>TM</sup> 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 [45]. However, accuracy was computed against true frequencies, not against estimates derived from the true counts as in [45]. 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  $\infty$  if  $\hat{f}_i > f_i = 0$ . The error fraction with threshold  $\tau$ , denoted  $EF_\tau$  is defined as the percentage of transcripts with relative error greater or equal to  $\tau$ . The median percent error, denoted MPE, is defined as the threshold  $\tau$  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 [59], 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.

Table ?? 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.

**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

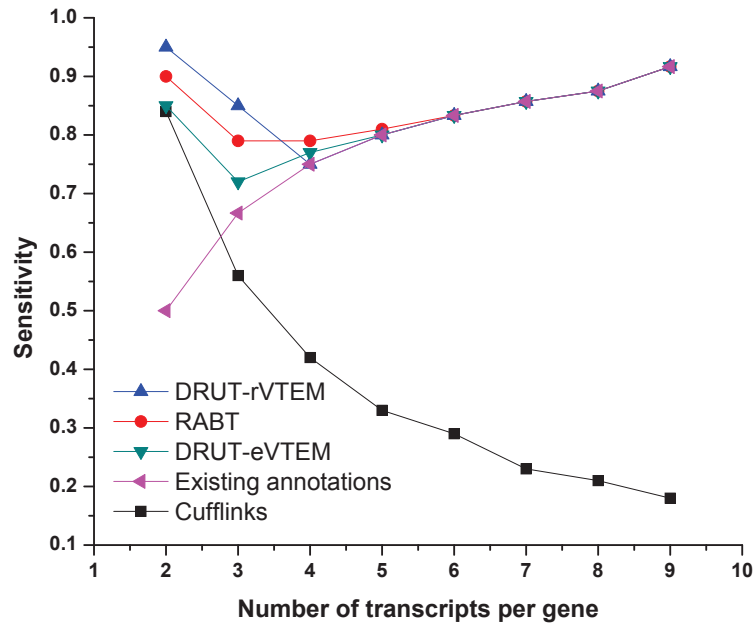
47, 431. In this section, we use the sensitivity and PPV defined above to compare our DRUT method to the most recent version of Cufflinks and RABT (version 1.3.0 of Cufflinks and RABT downloaded from website <http://cufflinks.cbcb.umd.edu/>). Due to the fact that results on 100bp and 200bp are very similar, we decided to present comparison on reads of length 100bp. TopHap [52] is used for Cufflinks and RABT to map simulated reads to the reference genome. For DRUT we used Bowtie [51] 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 [60] 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.

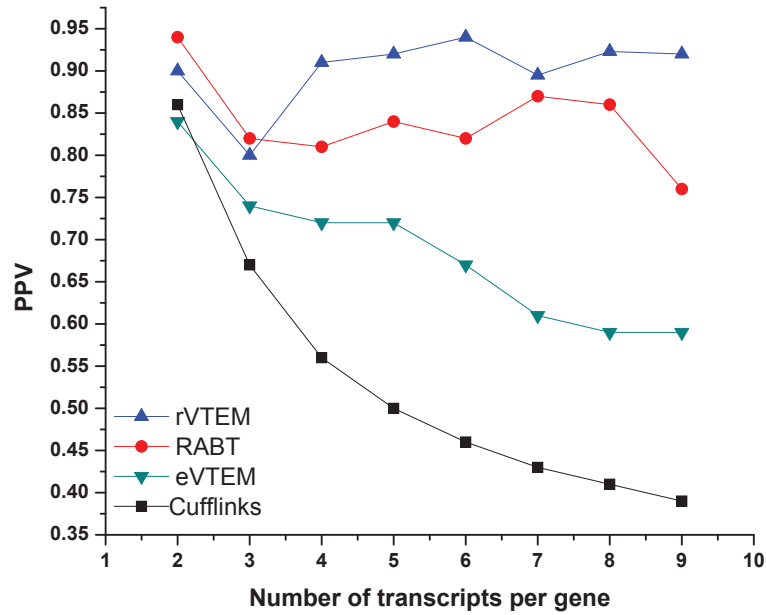
Next we want to define the portion of known transcripts that is input for annotation-guided methods as “existing annotations”. Please note that sensitivity of annotation-guided methods needs to be compared to the “existing annotations” ratio unlike regular reconstruction methods that do not have any 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 its 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 than 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 ( $n=1, \dots, 9$ ) : (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 [52] 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 MaLTA: Maximum Likelihood Transcriptome Assembly

Existing transcriptome methods([3],[6]) use read pairing information and fragment length distribution to accurately assemble set of transcripts expressed in a sample. This information is not available for current Ion Torrent technology, which can makes it challenging to assemble transcripts. Ion Torrent PGM platform is able to produce single reads with read length in 50-300bp range. We present MaLTA, method for simultaneous transcriptome assembly and quantification from Ion Torrent RNA-Seq data. Our approach explores transcriptome structure and incorporates maximum likelihood model into assembly procedure. MaLTA starts from a set of putative transcripts and selects the subset of this transcripts with the highest support from the RNA-Seq data. Maximum likelihood estimates of putative transcripts are computed using Expectation Maximization(EM) algorithm which take into account alternative splicing and mapping ambiguities. EM algorithm is state-of-the-art approach for transcriptome quantification from RNA-Seq data and are proven to outperform count-based approaches. Several independent implementations of EM algorithm exist in the literature ( [9], [35]).

We developed a new version of IsoEM [9] suitable for Ion Torrent RNA-Seq reads. IsoEM is an expectation-maximization algorithm for transcript frequency estimation. It overcomes the problem of reads mapping to multiple transcripts using iterative framework

which simultaneously estimates transcript frequencies and imputes the missing origin of the reads. A key feature of 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. In [9], we showed that modeling insert sizes is highly beneficial for transcript expression level estimation even for RNA-Seq data consisting of single reads, as in the case of Ion Torrent. Insert sizes contribute to increased estimation accuracy. They can help disambiguating the transcript 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 transcripts during the expectation step of the algorithm. Since most Ion Torrent sequencing errors are insertions and deletions, we developed a version of IsoEM capable of handling insertions and deletions in read alignments.

The main idea of the MaLTA approach is to cover all transcriptional and splicing variants presented in the sample with the minimum set of putative transcripts. We use new version of IsoEM algorithm, described above, to estimate expression levels of putative transcripts. Since we infer all possible transcripts in the sample, selecting all of them with non zero frequency will lead to unfeasible solution. Here, we suggest to select only such transcripts that contain novel variants and have highest support from sequencing data. To realize this idea we suggest a greedy algorithm which traverses the list of transcripts (sorted by expression levels in descending order) and select a transcript only if it contains a novel transcriptional or splicing event.

### 3.3.3 TRIP : Transcriptome *Re*construction 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 [61] . 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.

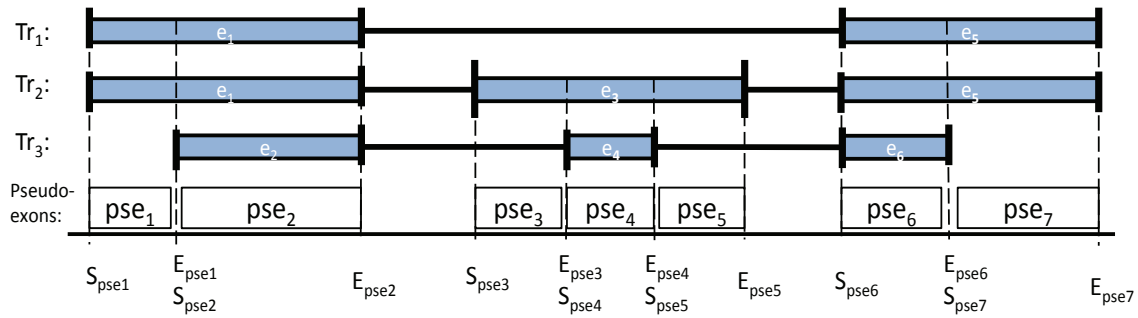


Figure 3.5 Pseudo-exons(white boxes) : regions of a gene between consecutive transcriptional or splicing events. An example of three transcripts  $Tr_i, i = 1, 2, 3$  each sharing exons(blue boxes).  $S_{psej}$  and  $E_{psej}$  represent the starting and ending position of pseudo-exon  $j$ , respectively.

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.

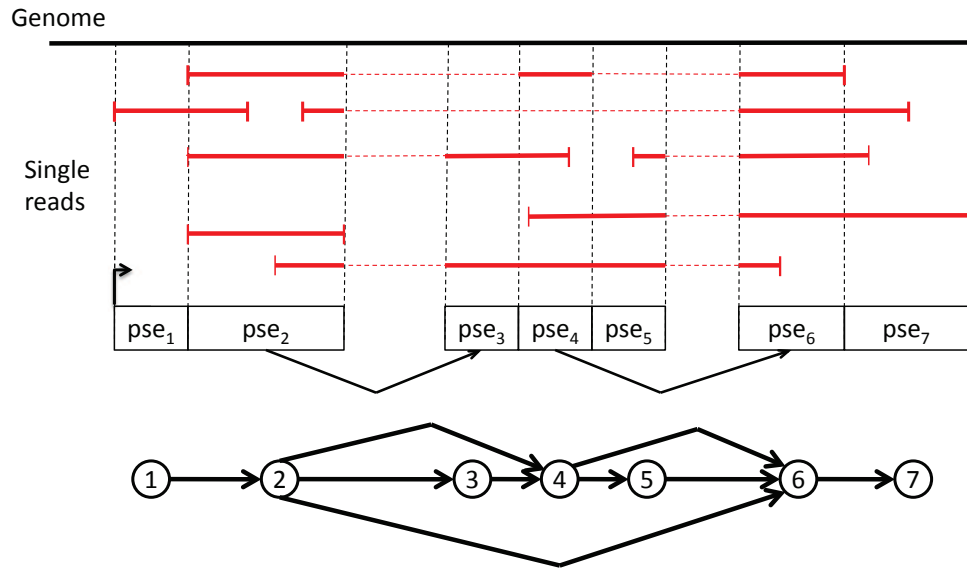


Figure 3.6 Splice graph. The red horizontal lines represent single reads. Reads interrupted by dashed lines are spliced reads. Each vertex of the splice graph corresponds to a pseudo-exon and each directed edge corresponds to a (splice) junction between two pseudo-exons.

**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  
( $s_1 \approx 68\%$ ,  $s_2 \approx 95\%$ ,  $s_3 \approx 99.7\%$ );
- $\epsilon$  allowed deviation from the rule ( $\epsilon = 0.05$ )
- $T_i(p_j)$  Set of candidate transcripts where  $p$  can be mapped with a fragment  
length between  $i - 1$  and  $i$  standard deviations,  $1 \leq i \leq 3$ ;
- $T_4(p_j)$  Set of candidates transcripts where  $p_j$  can be mapped with a  
fragment length within more than 3 standard deviations;

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

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

Subject to

- (1)  $\sum_{t_k \in T_i(p)} y(t) \geq x_i(p), \forall p, i = \overline{1, 4}$
- (2)  $N(s_i - \epsilon) \leq \sum_j x_i(p_j) \leq N(s_i + \epsilon), i = \overline{1, 4}$
- (3)  $\sum_i x_i(p) \leq 1, \forall p$
- (4)  $\sum_{t_k \in J_l} y(t) \geq 1, \forall J_l$

where the boolean variables are:

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

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

Constraint (1) implies that for each paired-end read  $p \in n(s_i)$ , at least one transcript  $t \in T_i(p_j)$  is selected. Constraint (2) restricts the 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.

### 3.3.4 MLIP : *Maximum Likelihood Integer Programming*

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 [61]. 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.

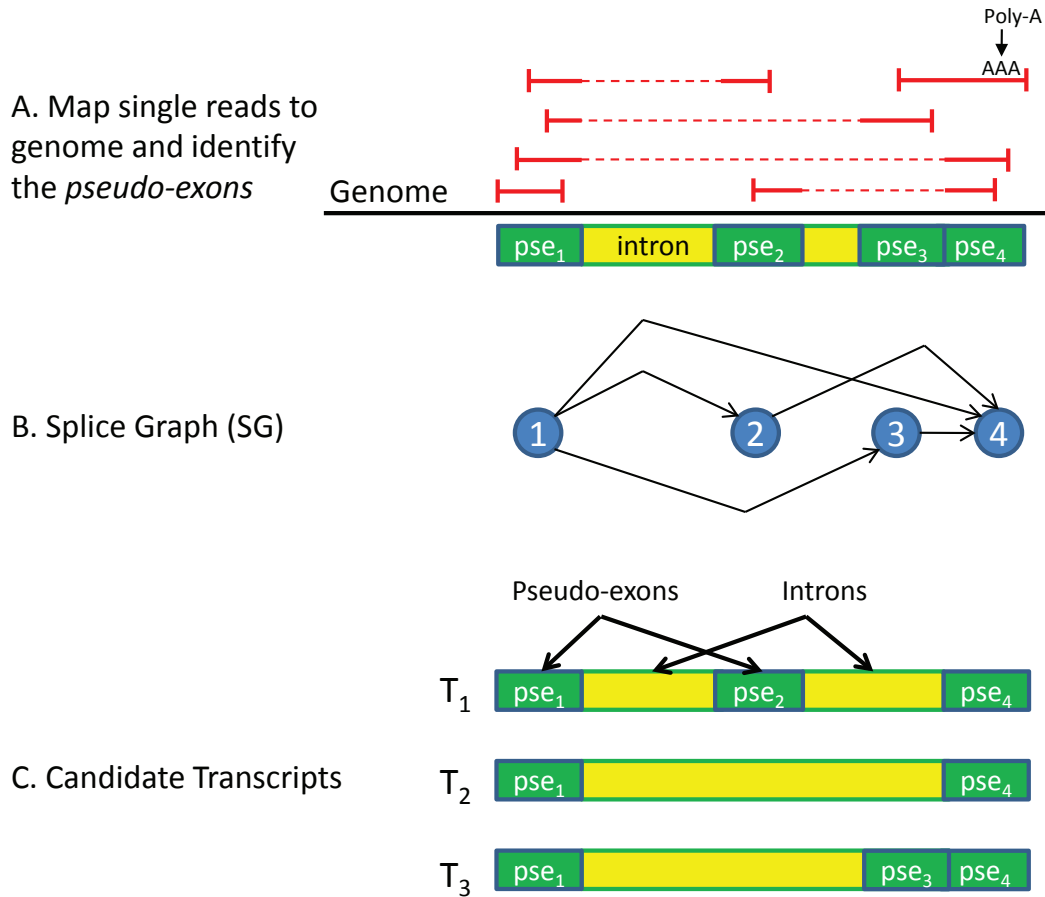


Figure 3.7 Model Description. A - Pseudo-exons. Pseudo-exons(green boxes) : regions of a gene between consecutive transcriptional or splicing events; B - Splice graph. The red horizontal lines represent single-end reads. Reads interrupted by dashed lines are spliced reads. Each vertex of the splice graph 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 [62].

**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[52] 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:

- $x(r \rightarrow t)$     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):

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

Subject to:

- (1) For any  $r$ , at least one transcript  $t$  is selected:  $y(t) \geq x(r \rightarrow t), \forall r, \forall t$
- (2) Read  $r$  can be mapped only to one transcript:  $\sum_{t \in T(r)} x(r \rightarrow t) = x(r), \forall r$
- (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 [63] 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  $G = \{T \cup R, E\}$  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 [43]). 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:

$$D = \frac{\sum_j |o_j - e_j|}{|R|}, \quad (3.2)$$

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:

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

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.

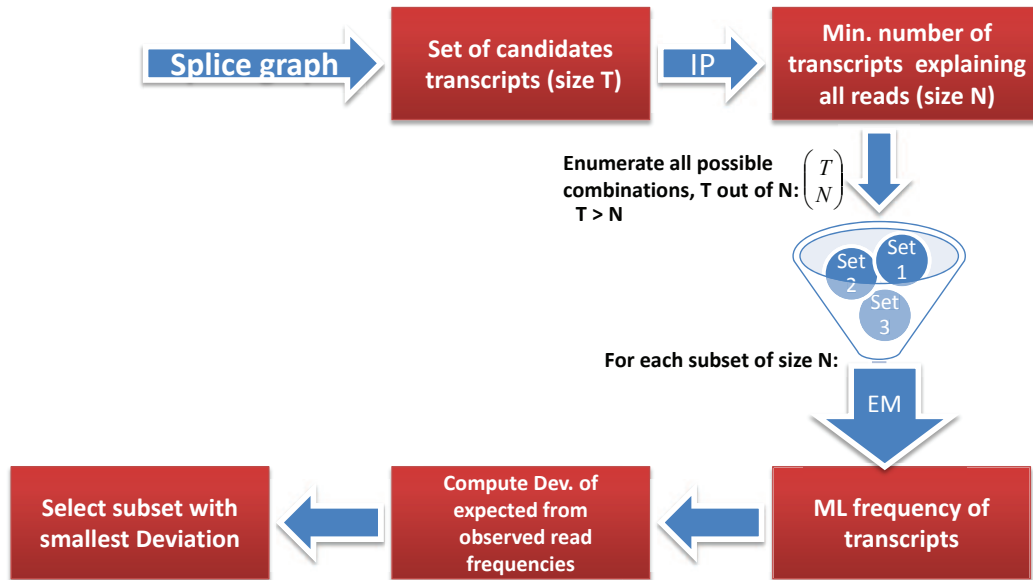


Figure 3.8 Flowchart for MLIP. Input : Splice graph. Output: subset of candidate transcripts with the smallest deviation between observed and expected read frequencies.

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.

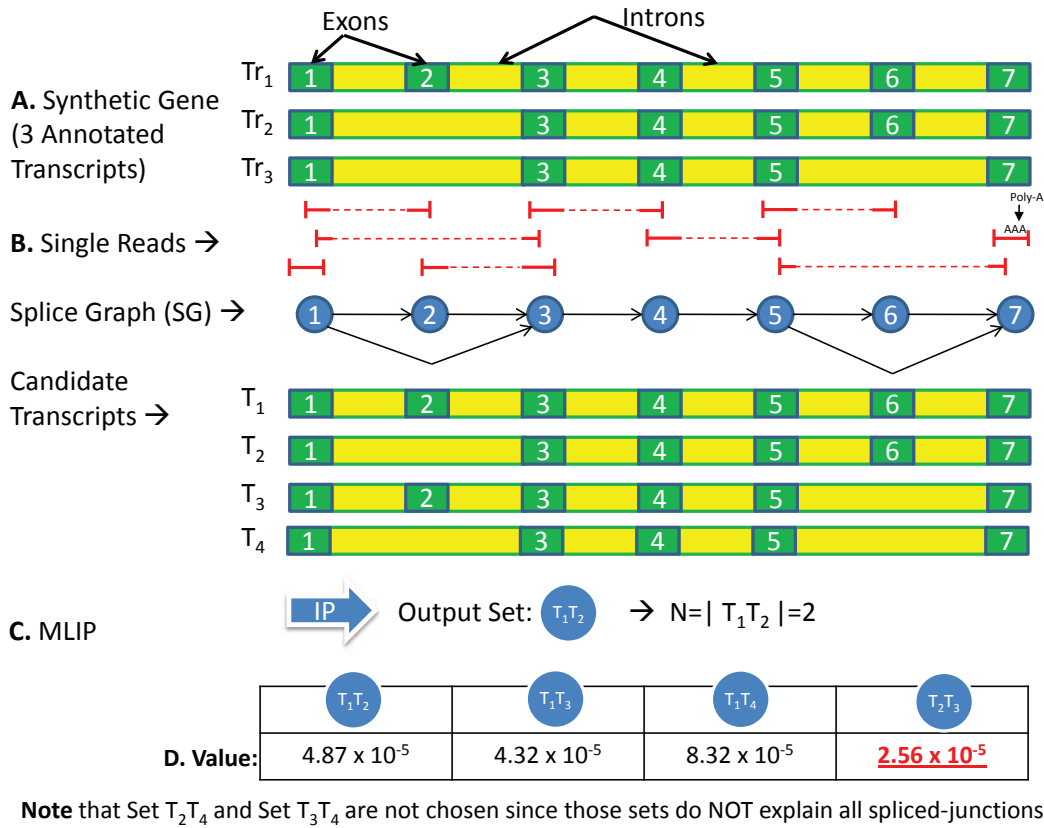


Figure 3.9 A. Synthetic gene with 3 transcripts and 7 different exons. B. Mapped reads are used to construct the splice graph from which we generate  $T$  possible candidate transcripts. C. MLIP. Run  $IP$  approach to obtain  $N$  minimum number of transcripts that explain all reads. We enumerate  $N$  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.

### 3.3.5 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  $100\times$  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.

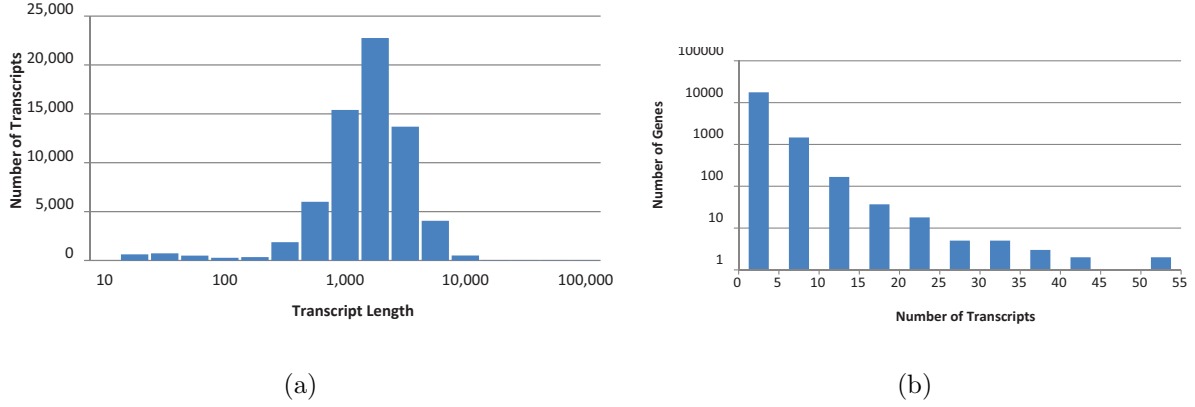


Figure 3.10 Distribution of transcript lengths (a) and gene cluster sizes (b) in the UCSC dataset

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 [62].

**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 [60].

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 [60] 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 [62]) 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.

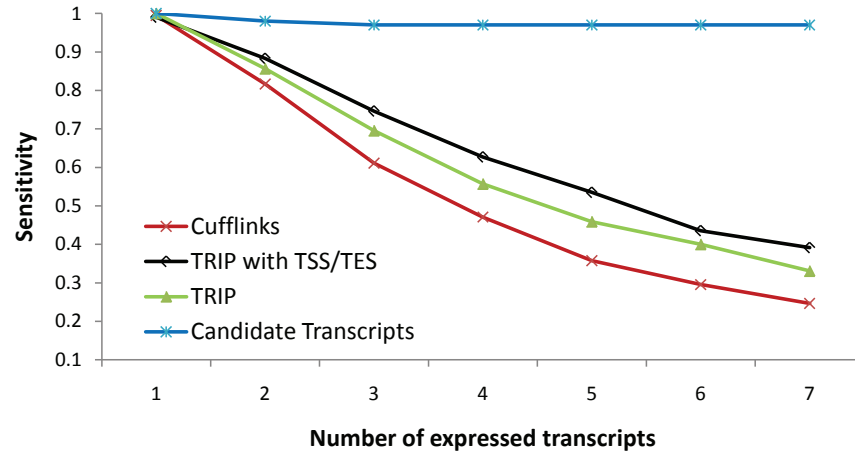
Fig. 3.11(b) shows that Cufflinks has an advantage over TRIP in the portion of correctly

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  $TRIP - L$ .

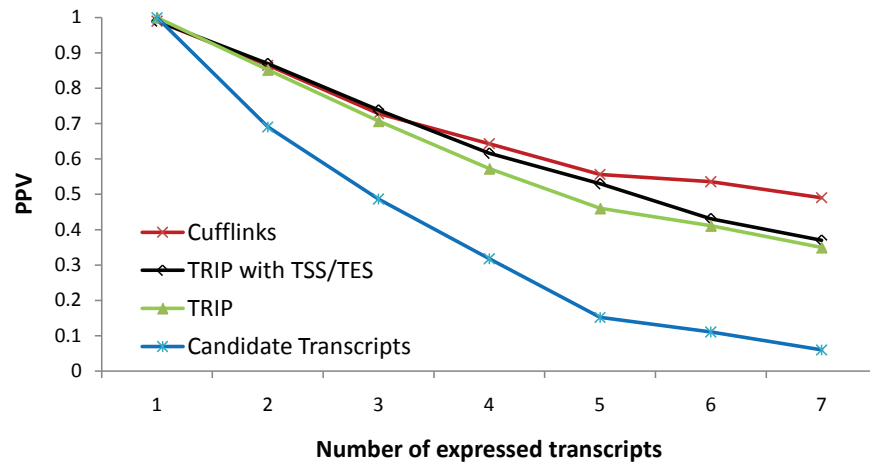
*Influence of Sequencing Parameters.* Although high-throughput technologies allow users to make trade-offs between read length and the number of generated reads, very little has been done to determine optimal parameters for fragment length. Additionally, novel Next Generation Sequencing (NGS) technologies 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.

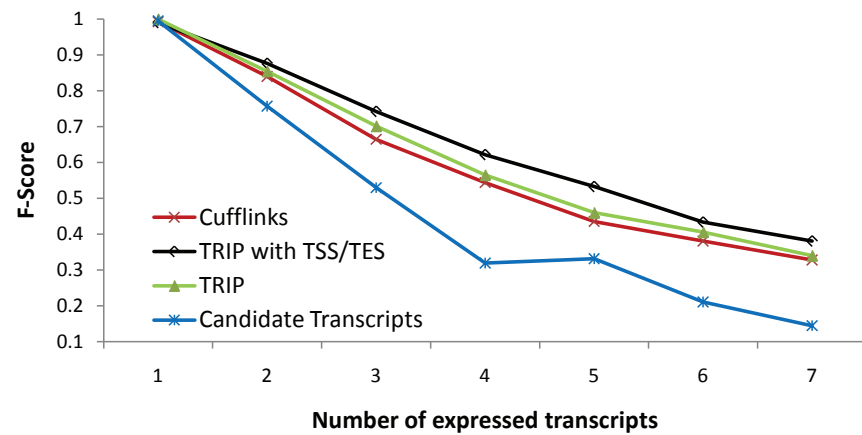
*Results on Real RNA-Seq Data.* We tested TRIP on real RNA-Seq data that we sequenced from a CD1 mouse retina RNA samples. We selected a specific gene that has 33 annotated transcripts in Ensembl. The gene was picked and validated experimentally due to interest in its biological function. We plan to 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)



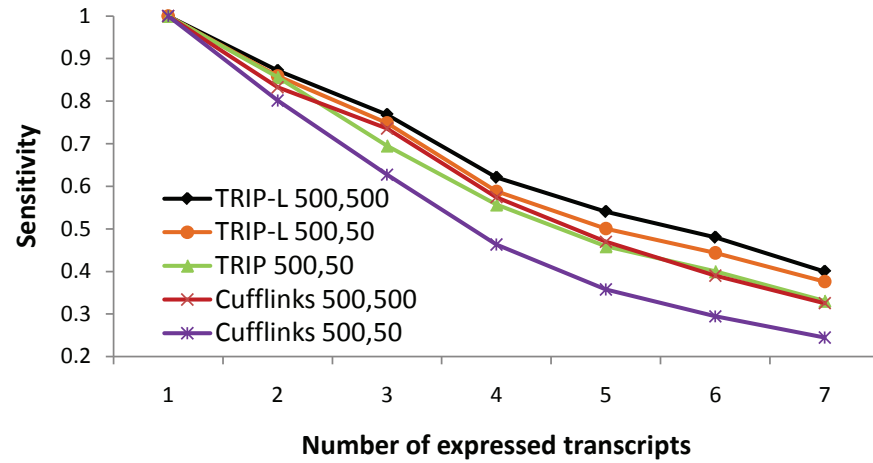
(b)



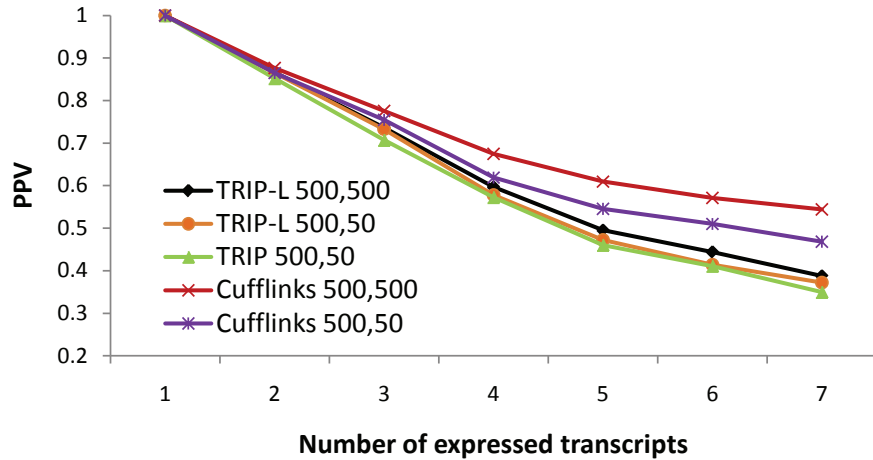
(c)

Figure 3.11 Comparison between methods for groups of genes with  $n$  transcripts ( $n=1, \dots, 7$ ) 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.

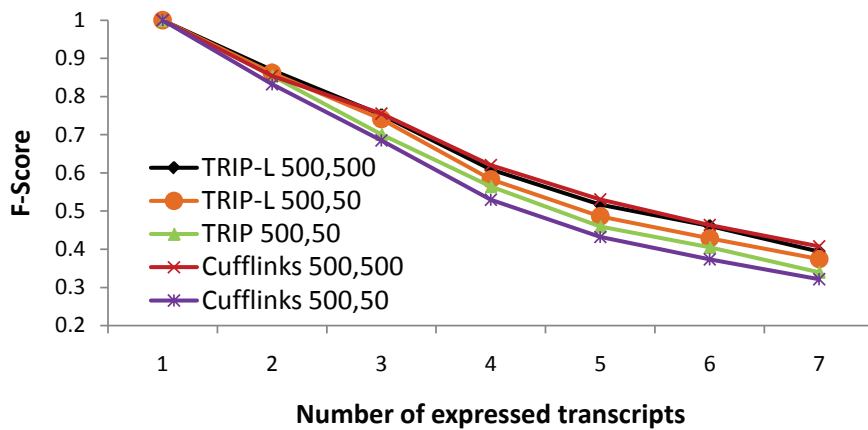




(a)



(b)



(c)

Figure 3.12 Comparison between methods for groups of genes with  $n$  transcripts ( $n=1, \dots, 7$ ) on simulated dataset with different sequencing parameters and distribution assumptions: (a) Sensitivity (b) Positive Predictive Value (PPV) and (c) F-Score.

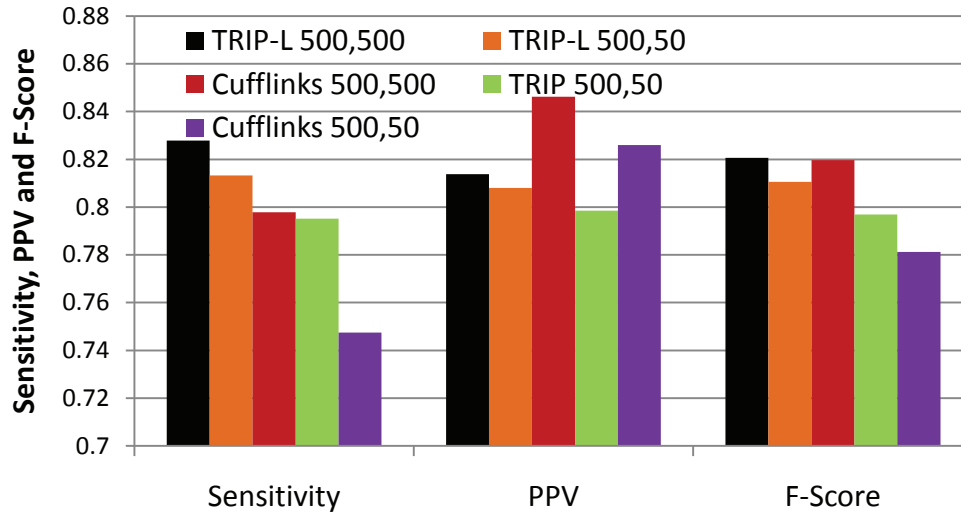


Figure 3.13 Overall Sensitivity, PPV and F-Score on simulated dataset with different sequencing parameters and distribution assumptions.

**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 [60] are used for comparison. We explore the influence of read length, fragment length, and coverage on reconstruction accuracy.

Table 3.2 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 3.2 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 3.3 compares sensitivity, precision and F-score of Cufflinks, IsoLasso, and MLIP for different combinations of read and fragment lengths: (50bp,250bp), (100bp,250bp),

Table 3.3 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 3.4 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

Table 3.4 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

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

### 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 [64] 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.

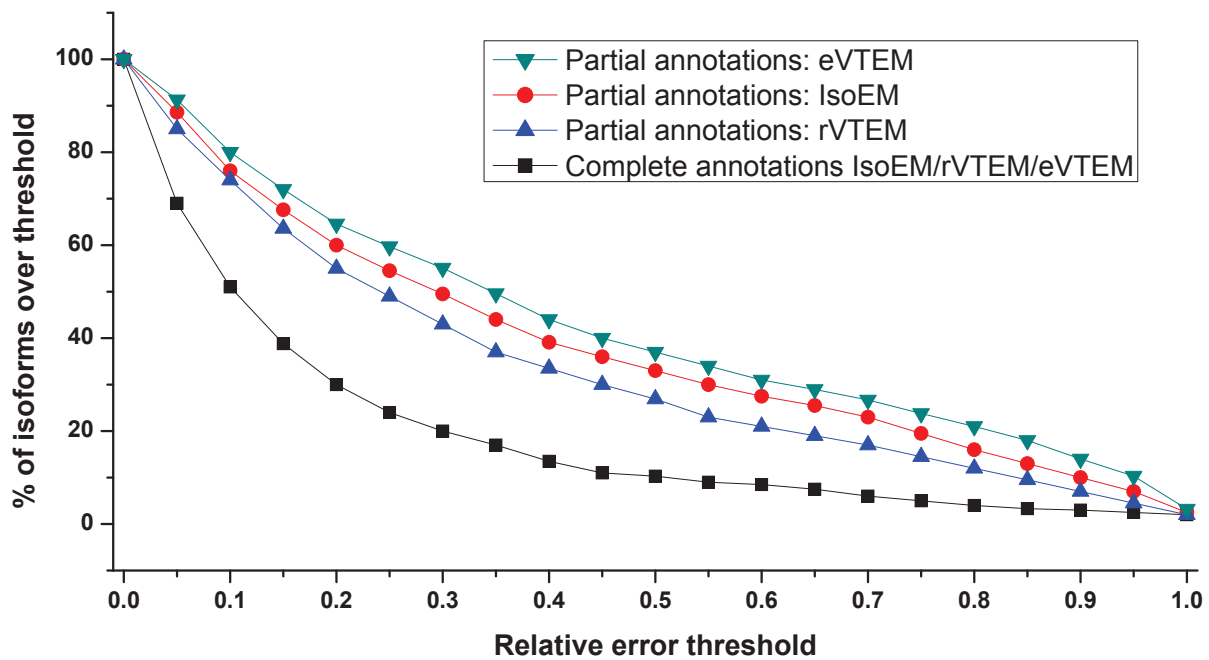
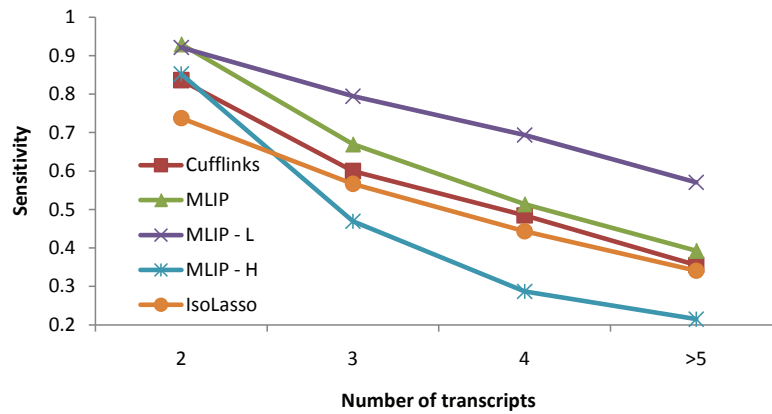
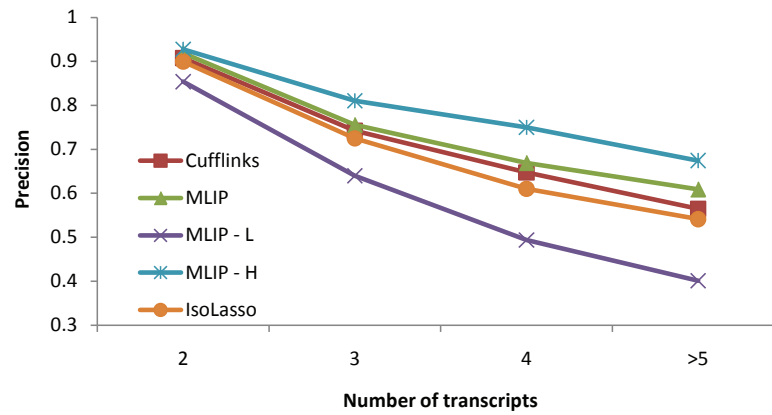


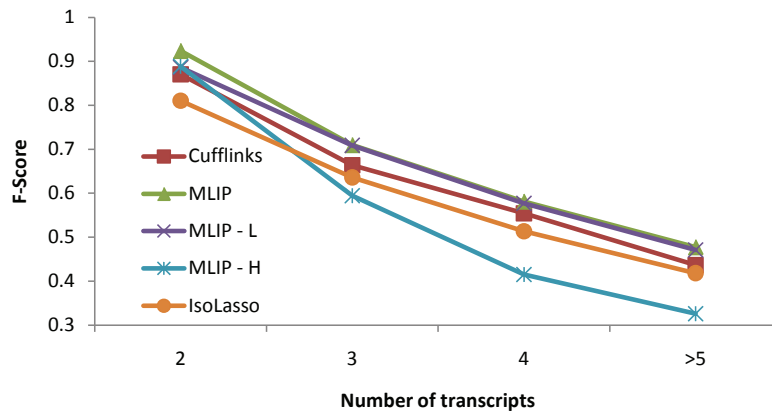
Figure 3.3 Error fraction at different thresholds for isoform expression levels inferred from 30 millions reads of length 25bp simulated assuming geometric isoform expression. Black line corresponds to IsoEM/VTEM with the complete panel, red line is IsoEM with the incomplete panel, blue line is rVTEM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



(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 ( $n=1, \dots, \geq 5$ ) on simulated dataset with mean fragment length 250bp, standard deviation 25bp and read length of 100bp.



## PART 4

### ***DE NOVO* ASSEMBLY AND ANNOTATION OF REAL DATA SETS.**

Functional genomic studies of the molecular mechanisms requires solid genome and transcriptome annotations. Poor or missing annotations are common for many model organisms that could be useful for development and understanding of many biological reasons as well as pharmaceutical research and drug development. In this part we present assembly and annotation of *Bugula neritina* transcriptome (a colonial animal), and *Tallapoosa Darter* genome (a species-rich radiation freshwater shes in North America).

#### **4.1 Assembly of Illumina RNA-Seq Reads and Contig Annotation for *Bugula neritina***

##### 4.1.1 Background

Studying the interactions between eukaryotic organisms and microbial pathogens provides insight into potential treatments for devastating diseases. Researchers, however, are increasingly recognizing the importance of beneficial microbes to the health and ecology of their hosts, and that understanding the interactions between partners in mutualistic symbiosis may also contribute to knowledge of pathogenesis.

Mutualistic symbiosis is a beneficial interaction between two partners, which usually results in enhanced nutrition or defense for one or both partners [65,66]. In one type of defensive symbiosis, the symbiont provides protection to the host by synthesizing small molecules with bioactivity against pathogens, parasites, and predators [67,68]. These small molecules are generally toxic to the hosts adversary by affecting its cellular processes. However, knowledge about the interaction of the host itself with these symbiont-produced bioactive compounds is limited. The marine bryozoan, *Bugula neritina*, has an uncultured bacterial symbiont that produces the bioactive compounds, the bryostatins, which have

activity against cancer, Alzheimers and other neurological diseases, and HIV [69]. Bryostatins are activators of protein kinase C (*PKC*), which is an eukaryotic signaling protein in a variety of cellular processes, which accounts for its diverse pharmacological activity. In this study, we investigated the response and adaptation of the host bryozoan, *B. neritina*, to symbiont-produced bryostatins through a variety of approaches.

Despite the abundance of microbial symbiont-produced compounds [70] and their activity in eukaryotic cellular processes, very few studies have investigated host adaptation or response to these compounds.

#### 4.1.2 Methods

**Collection of *Bugula neritina* larvae and antibiotic treatment.** *Bugula neritina* colonies can be found in three different parts of the Atlantic coast. Arborescent colonies of *B. neritina* growing on floating docks in Beaufort, NC, USA were collected in November 2010 and maintained overnight in the dark in flowing seawater tables in wet laboratory facilities at UNC-Chapel Hills Institute of Marine Sciences in Morehead City, NC, USA. In the morning, the colonies were placed into large glass jars filled with seawater and exposed to sunlight to stimulate larval release. The released larvae swam to the top of the jar and were collected with a wide tip glass pipette into a collection vial. The larvae ( 100 larvae) were pipetted into six-well polystyrene plates (n= 6 replicate plates) containing filter-sterilized seawater with either an antibiotic, gentamicin (treatment) or seawater with a small volume of distilled water (control). The larvae in the plates were allowed to settle and metamorphose.

Adult colonies of *B. neritina* were also collected from Radio Island Marina, Beaufort, NC, USA and Morehead City Yacht Basin, Morehead City, NC, USA in March 2012, as well as from Oyster public docks (Oyster, VA, USA), and Indian River Inlet, DE, USA in June 2012.

**Sequence analysis of PKCs in *Bugula* species.** Total RNA was extracted from environmental adult symbiotic and aposymbiotic Type S *B. neritina* colonies, it was purified and treated to remove any contaminating DNA molecules. The purified total RNA was processed according to standard operating procedure for preparation of mRNA library for sequencing (TruSeq RNA Sample Preparation Kit, Illumina, San Diego, CA, USA). Briefly, RNA quality and quantity was assessed using the Agilent 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA). Poly-A containing mRNA molecules were purified using poly-T oligo-attached magnetic beads. The purified mRNA were fragmented (120-200 bp) using 558 divalent cations at 94 °C. First strand cDNA was synthesized using reverse transcriptase (SuperScript II, Invitrogen, Carlsbad, CA, USA). The RNA template was digested with RNase H, and the second strand of cDNA was synthesized using DNA polymerase I. The adapter-ligated cDNA fragments were purified and selectively enriched by PCR using a primer cocktail complementary to the ends of the adapters. The adapter-ligated cDNA library was hybridized to the surface of an Illumina flow cell and sequenced on an Illumina sequencing platform (Illumina HiSeq 2500, San Diego, CA, USA) at the Integrated Genomics Facility, Georgia Regents University Cancer Center, Augusta, GA, USA.

The paired-end reads were assembled *de novo* using Trinity software (version r2013-02-25), and the assembled contigs were annotated by 570 performing blastx searches (Translated Query-Protein Subject BLAST 2.2.26+) against the Swiss-Prot database. Sequences identified as PKCs were further analyzed by MotifScan (ExPASy, [http : //myhits.isb – sib.ch/cgi – bin/motif\\_scan](http://myhits.isb-sib.ch/cgi-bin/motif_scan)) to identify relevant domains.

#### 4.1.3 Assembly and annotation of *B. neritina* transcriptome sequences

The *Bugula* RNA-Seq Illumina reads analyzed were paired-end reads of length 50bp with 200bp mean fragment length. The reads were assembled into contigs by Trinity. We BLASTed the Trinity contigs on Swissprot database and got 12067 matches, 59.37% ORFs hits, 63.35% Contigs hits and 7,846 Proteins hits. Using IsoDE, we were able to identify 1485 differential expressed genes between two different conditions, namely the *Bugula* from

Eastern coast which has the Symbiont bacteria against the Bugula from the Northern coast missing the bacteria. Finally we found some Bugula orthologs of the Protein kinase C (PKC).

A summary description of the location and type of all six samples of Illumina paired-end reads is presented below:

- L1. Shallow with symbiont (NC)
- L2. Shallow without symbiont (VA)
- L3. Northern with symbiont (NC)
- L4. Northern without symbiont (DE)
- L5. Shallow symbiotic, ovicell-bearing tissue (NC)
- L6. Shallow symbiotic, ovicell-free tissue (NC)

The chart presented in figure 4.1 was obtained by inputting sample 1 (L1) and sample 2 (L2) reads into the online tool metagenomics RAST server [2]: <http://metagenomics.anl.gov/>. Metagenomics problems for assembling illumine reads occur due to reads contamination. Note that around half of those reads comes from the bacteria. Samples L3 and L4 include more errors and therefore we have excluded them from analysis.

In addition to those six reads samples we have used:

- 968 contigs assembled and filtered using the standard default assembly in the 454 software by Selah Clinical Genomic Center at Innovista, Columbia, South Carolina ([www.engencore.sc.edu](http://www.engencore.sc.edu)).

And three sets of NCBI 454 *Bugula neritina* reads that we will assemble into contigs (using Newbler [71]):

- 24 mRNA (from which 14 are complete genes ESTs)
- 3360 Sanger ESTs Source: <http://sra.dnaxexus.com/runs/SRR034781>

Newbler is a software package for de novo DNA sequence assembly. It is designed specifically for assembling sequence data generated by the 454 platforms.

Next we merged the resulting contigs with the 968 contings using the Minimus2 assembler from the AMOS package [72]. Merging the assembled contigs resulted in a much

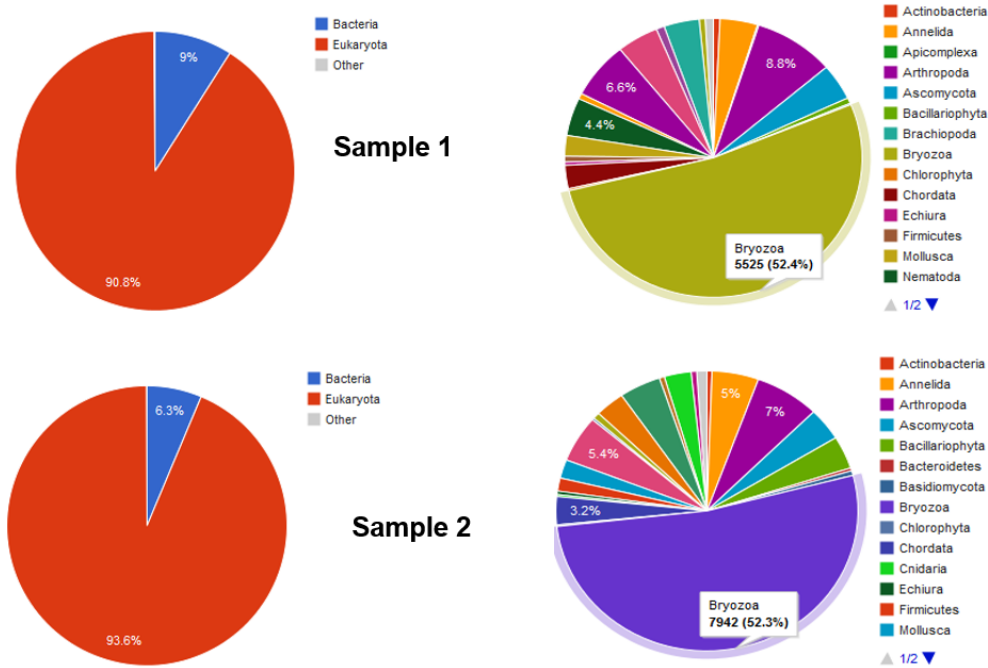


Figure 4.1 Screenshot from Metagenomics [2]

better assembly. Finally we run the scaffolding. In addition we use Illumina reads to fill the gaps (after scaffolds).

Table 4.1 presents the number of contigs that are shared between all shallow species and the other northern ones.

#### 4.1.4 *Bugula neritina* Flows

- Newbler: Run on 454 reads (N=139,131, avg. length=347.5bp, 48Mb total)
  - Number of contigs: 7,582
  - Number of best ORFs: 3,495
  - Number of Hits: 2,206
  - Number of Proteins: 1,556
- Trinity: Run on all Illumina reads (all samples combined N=221,818,850 2x50bp pairs, 22Gb total)
  - Number of contigs: 166,951 (after filtering with  $RSEM\_isopct\_cutoff = 1.00$ )

Table 4.1 Sharring Shallow

```

Trinity-Shallow
Trinity-L1
Trinity-L2
Trinity-L5
Trinity-L6
565 L1
100 L1 L2
39 L1 L2 L5
67 L1 L2 L5 L6
42 L1 L2 L6
81 L1 L5
55 L1 L5 L6
98 L1 L6
437 L2
68 L2 L5
36 L2 L5 L6
63 L2 L6
363 L5
92 L5 L6
434 L6
1309 Shallow
398 Shallow L1
207 Shallow L1 L2
315 Shallow L1 L2 L5
5741 Shallow L1 L2 L5 L6
271 Shallow L1 L2 L6
130 Shallow L1 L5
300 Shallow L1 L5 L6
133 Shallow L1 L6
493 Shallow L2
165 Shallow L2 L5
286 Shallow L2 L5 L6
128 Shallow L2 L6
332 Shallow L5
134 Shallow L5 L6
208 Shallow L6

```

- Number of best ORFs: 76,769
- Number of Hits: 37,026
- Number of Proteins: 12,748
- Minimus: Merge filtered Trinity contigs, Newbler contigs, and NCBI ESTs:
  - Number of contigs: 133,470
  - Number of best ORFs: 52,766
  - Number of hits: 24,130
  - Number of proteins: 12,336
- Binning.

Input: Illumina for K samples (required) + 454 + ESTs (if available)

1) Contig assembly

- a) Assemble Illumina reads (COMBINED from all samples) using Trinity
- b) (optional) assemble 454 reads using Newbler (if 454 available, combine reads from all samples if needed)
- c) (optional) combine contigs from 1) and 2) with ESTs (if available) using minimus
- 2) Compute coverage (fpkms) for all contigs from step 3, for each sample separately, using Illumina reads, this gives a K-dimensional vector of fpkms for each contig
- 3) Coverage- and PE based clustering of contigs
  - based on euclidean distance of K-vectors of fpkm
  - correlation of K-vectors of fpkm
  - PE with one read in one contig and the other in second contig
- 4) Assign read pairs to contig clusters (if one read maps to a contig in a cluster the pair gets assigned to the cluster)
- 5) Assemble independently read pairs assigned to each contig cluster (plus Newbler contigs and ESTs, if any) using accurate assembler (this gives new set of contigs).
- 6) Assemble read pairs that are not assigned to any cluster and add to current set of contigs
- 7) Repeat steps 2-6 until no more contig changes
- 8) Independently scaffold each contig cluster using:
  - a) assigned PEs and SILP algorithm, or
  - b) comparative scaffolding (if related genome is available)

Output:

- contig scaffolds
- final contig fpkm K-vectors

#### 4.1.5 Analysis of results of each flow

Table 4.2 presents a comparison of several assemblies. Newbler merged reads into longer contigs, while Oases [73] produced the overall shortest assembly. The Oases assembly gives

about half the protein hits that Trinity gives, most likely due to the overly stringent coverage filter. The results obtained by assembling each Illumina lane separately with Trinity.

Table 4.2 BlastX Results

Assembly	Newbler	Oases	Oases	Minimus	Trinity Shallow L1L2L5L6	Trinity Shallow L1L2L5L6	Trinity Shallow	Trinity All Filtered	Trinity- All	Trinity- L1	Trinity- L2	Trinity- L3	Trinity- L4	Trinity- L5	Trinity- L6
		min. coverage 50	defaults	filtered Trinity contigs + Newbler contigs + NCBI ESTs			L1,2,5,6 no filtering	RSEM_i sopct_c utoff = 1.00	no filtering	no filtering	no filtering	no filtering	no filtering	no filtering	no filtering
# Contigs	7,582	45,311	290,046	133,470	19,048	19,048	126,916	166,951	207,507	58,819	57,268	60,343	58,607	57,121	51,889
# best ORFs	3,495	25,964	117,102	52,766	20,325	20,325	69,937	76,769	103,976	32,872	33,416	40,489	33,755	32,019	33,966
Contigs w/ ORFs	46.10%	57.30%	40.37%	39.53%	106.70%	106.70%	55.10%	45.98%	50.11%	55.89%	58.35%	67.10%	57.60%	56.05%	65.46%
Protein database	swiss-p	swiss-p	swiss-p	swiss-prot	swiss-p	nr	swiss-p	swiss-p	swiss-p	swiss-p	swiss-p	swiss-p	swiss-p	swiss-p	swiss-p
# BLASTX hits	2,206	14,033	61,182	24,130	12,067	13,883	37,963	37,026	50,828	18,906	19,068	17,150	19,194	18,575	19,343
ORFs w/ hits	63.12%	54.05%	52.25%	45.73%	59.37%	68.31%	54.28%	48.23%	48.88%	57.51%	57.06%	42.36%	56.86%	58.01%	56.95%
Contigs w/ hits	29.10%	30.97%	21.09%	18.08%	63.35%	72.88%	29.91%	22.18%	24.49%	32.14%	33.30%	28.42%	32.75%	32.52%	37.28%
# Proteins w/ hits	1,556	6,820	12,846	12,336	7,846	9,972	10,437	12,748	12,578	8,397	8,316	8,661	8,439	8,084	7,961
BLASTX hits / Protein	1.42	2.06	4.76	1.96	1.54	1.39	3.64	2.90	4.04	2.25	2.29	1.98	2.27	2.30	2.43

Hit distribution

90-100%	196	1,996	3,169	3,848	3,151	4,612	3,654	3,948	3,872	2,922	3,042	2,892	2,988	2,963	2,804
80-90%	112	892	1,584	1,680	1,303	1,639	1,552	1,697	1,657	1,213	1,216	1,237	1,229	1,210	1,176
70-80%	95	593	1,280	1,184	817	937	1,034	1,215	1,197	798	807	792	780	792	750
60-70%	107	524	1,162	1,005	595	694	828	1,029	1,027	625	629	661	664	641	617
50-70%	114	471	1,149	974	482	588	760	1,001	999	603	564	645	592	541	553
40-50%	166	555	1,208	979	480	498	742	1,016	1,018	611	568	641	647	546	577
30-40%	234	596	1,298	1,055	440	467	720	1,140	1,130	615	580	625	594	539	562
20-30%	274	650	1,209	970	382	324	713	1,030	1,018	609	545	659	544	515	535
10-20%	258	543	786	641	196	213	434	672	660	401	365	509	401	337	387
0-10%	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

A large number of protein hits were recovered from all assemblies:

4637 All L1 L2 L3 L4 L5 L6

As expected, there is also a large number of proteins that were found in the combined assembly (called "All") but not found in any of the individual lane assemblies: 2170 All

However, each lane has a fairly large number of protein hits that were not recovered in the combined assembly (some protein hits even appear in all individual samples but not in the combined one), here are the numbers for the most abundant combinations that do not include presence in the combined assembly: 582 L3

489 L4



427 L3 L4

361 L1

306 L2

255 L6

241 L5

61 L1 L2 L3 L4 L5 L6

56 L1 L2

...

This may be due to BLASTX picking different best hits, and we should repeat differential analysis based on expression levels once we stabilize on a list of transcripts.

In addition to Swiss-Prot database, we have also performed BlastX searches against the "non-redundant protein sequences" (nr) database, because initial tests against this database proteins from marine invertebrates in the top hits. For this experiment we have used an e-value cutoff of 1e-20 and save only the top hit for each ORF.

#### 4.1.6 Sequence analysis of genes and C1b domains from *Bugula* species

After assembly and annotation of *B. neritina* transcriptome sequences, 5 contigs with homology to PKC isoenzymes were identified (Table 4.3) using blastx. Two contigs were homologous to cPKCs, two to nPKCs, and one to aPKCs.

Bioinformatic analysis of the PKC homologs revealed that the two cPKCs from *B. neritina* are -types (57.0% similarity between the two). One of the nPKCs is a -type, and the other is an -type. The aPKC appears to be an -type. Expression of these PKCs was confirmed in independently collected *B. neritina* cDNA using primers specific for each of the isoenzymes.

#### 4.1.7 Results

The results from this study suggest that this symbiotic association may be more than just defense: the symbiont, symbiont-derived bryostatins, or both, potentially affect *B.*

Table 4.3 PKC homologs identified by *Bugula neritina* transcriptome sequencing.

Contig	DNA Length	Amino acid Length	Highest hit (by blastp)	e-value	% identity	PKC isoform	GenBank Accession
4634	2130	710	Protein kinase C $\alpha$ type ( <i>Macaca mulatta</i> )	0.0	59%	Conventional $\alpha$	NP_001247662
16020	2031	677	Protein kinase C $\alpha$ type-like isoform X1 ( <i>Haplochromis burtoni</i> )	0.0	61%	Conventional $\alpha$	XP_005942808
12712	2091	697	Protein kinase C $\delta$ type-like ( <i>Saccoglossus kowalevskii</i> )	0.0	55%	Novel $\delta$	XP_002740313
16336	2148	716	Calcium-independent protein kinase C ( <i>Aplysia californica</i> )	0.0	63%	Novel $\epsilon$	NP_001191401
6484	1710	570	atypical protein kinase C ( <i>Lymnaea stagnalis</i> )	0.0	72%	Atypical $\iota$	BAK09601

*neritina* reproduction. We hypothesize that the bryozoan host has evolved to the presence of bryostatins such that the activation of the host PKC by the bryostatins triggers cellular mechanisms for reproduction. This interaction facilitates maintenance of the symbiosis by transmission of the symbiont to subsequent generations of the host and ensures host fitness by passing down the bryostatin-producing symbiont for protection against predation. This study extends our understanding of host-symbiont interactions that are important for the establishment and maintenance of diverse mutualistic partnerships. The difference in adaptive interaction of bryostatins with the host PKC compared to non-host PKC could impact pharmaceutical research and drug development of the bryostatins, and unlock new ways of increasing its efficiency for the treatment of a variety of human diseases.

Both the symbiotic and symbiont-reduced (via antibiotic treatment) *B. neritina* colonies were healthy and grew at statistically similar rates, indicating that the symbiont does not contribute significantly to host nutrition. The fecundity of the symbiont-reduced colonies, however, was significantly decreased as indicated by fewer reproductive structures in the colonies. Western blot analysis of bryostatin-activated conventional PKCs demonstrated a different banding pattern for the control colonies, suggesting that the presence of bryostatins

associated with the symbiont affected the native PKCs, whereas no such differences were noted for bryostatin-independent PKCs. Similar results were also observed for the PKCs in symbiotic and naturally-occurring aposymbiotic colonies. In addition, bryostatins affected fecundity and PKC expression in the model invertebrate, *Caenorhabditis elegans*. Analysis of transcriptome sequencing data revealed the presence of at least 5 PKC isoforms expressed in *B. neritina*.

#### 4.1.8 Conclusions and Future work

The reduction in host fecundity upon loss of the symbiont suggests that host reproduction has evolved to be dependent on the symbiont, the bryostatins, or both, to enhance host fitness by increasing the frequency of symbiont-infected, defended host larvae. Our results indicate that the presence of bryostatins modulates PKC activity in symbiotic *B. neritina* and bryostatin-exposed *C. elegans*. These findings lead us to hypothesize that the symbiont-produced bryostatins are an important cue for reproduction in *B. neritina* via PKC activation.

Future work includes metabolic pathways from proteins, differential expression of contigsscaffolds, and identification of symbiont transcripts

#### **Acknowledgements:**

This work was done in collaboration with Meril Mathew (the leading contributor) under the direction of Dr. Nicole Lopanik from (Department of Biology, Georgia State University, Atlanta, GA). Part of this work was submitted to BMC Biology.

## 4.2 Assembly and Annotation of the *Etheostoma tallapoosae* Genome

The family Percidae contains over 200 species, most of which are within the subfamily Etheostominae. This subfamily (the darters) represents a species rich radiation of freshwater fishes in North America. Evolutionary relationships between the various darter species have been deduced from morphological, mitochondrial DNA sequence and limited nuclear DNA sequence comparisons. However, a thorough understanding of the evolution of the darter species will require comparisons at the whole genome level.

As a first step, the genome of the Tallapoosa darter (*Etheostoma tallapoosae*) has been sequenced utilizing two Illumina MiSeq 250-PE runs generating 52 million reads. This provided an average 12 fold coverage of the estimated 1 billion nucleotide genome. The sequences were assembled with Minia into contigs and these were assembled into scaffolds with SSPACE.

A BLAST server has been set up to allow for the identification of Tallapoosa darter scaffolds homologous to sequences of interest. The scaffolds were also imported into an instance of WebApollo along with gene evidence tracks generated by fgenesh. A set of scripts were developed to facilitate the formatting and import of these tracks and scaffold sequences into WebApollo that will make it simple for labs to set up WebApollo instances for their own genome data without extensive computer system experience. A web site has been developed that gives access to both the BLAST and WebApollo servers to the public to spur interest in darter genomics and to enable annotation of the Tallapoosa darter genome by a community of darter researchers.

### 4.2.1 Introduction

So far, the study of darter evolution has utilized morphological, behavioral and limited DNA sequence analysis. While much of darter phylogeny has been elucidated from these studies, there are still many unresolved questions. For example, to what extent do related species share alleles due to incomplete lineage sorting or hybridization during evolution.

What are the actual adaptive genetic changes that define darter species? To what extent, if any, do allopatrically distributed and genetically differentiated populations of the same species show adaptive genetic differentiation?

A complete understanding of darter evolution must utilize the analysis of complete genomes. While this approach was not financially feasible in the past, the cost of genomic analysis is about to cross a threshold where sequencing of darter genomes of individual species and, soon, populations within species will become affordable.

As a starting point, it will be necessary to have a fully annotated reference darter genome sequence to which the genomic sequences of other darter species can be compared. As a first step in this direction I have recently obtained the genomic sequence of the Tallapoosa darter (*E. tallapoosae*).

#### 4.2.2 Sequencing

This sequence was obtained as a result of two 250 nucleotide PE runs on an Illumina MiSeq. A total of 13 billion nucleotides of sequence was obtained from 52 million such 250 nucleotide sequence reads. This represents, on average, about a 12 fold coverage of the darter genome. Figure 4.2 shows that alignment of reads to previously cloned genomic fragments shows that coverage ranges from 2 to 3 fold to as high as 28 fold.

#### 4.2.3 Assembly

The 250-PE sequences were assembled into contigs utilizing the **Minia** assembler. This assembler was chosen because of its low memory requirements. The sequences were assembled with most of the combinations of k-mer = 31 to k-mer = 80 settings and minimum abundance = 2 or 3 settings. The best assembly in terms of total number of nucleotides assembled and the maximum contig length was achieved with the settings k-mer = 73 and minimum abundance = 2:

- Total number of **contigs** = 539616
- Sum (bp) = 660984269

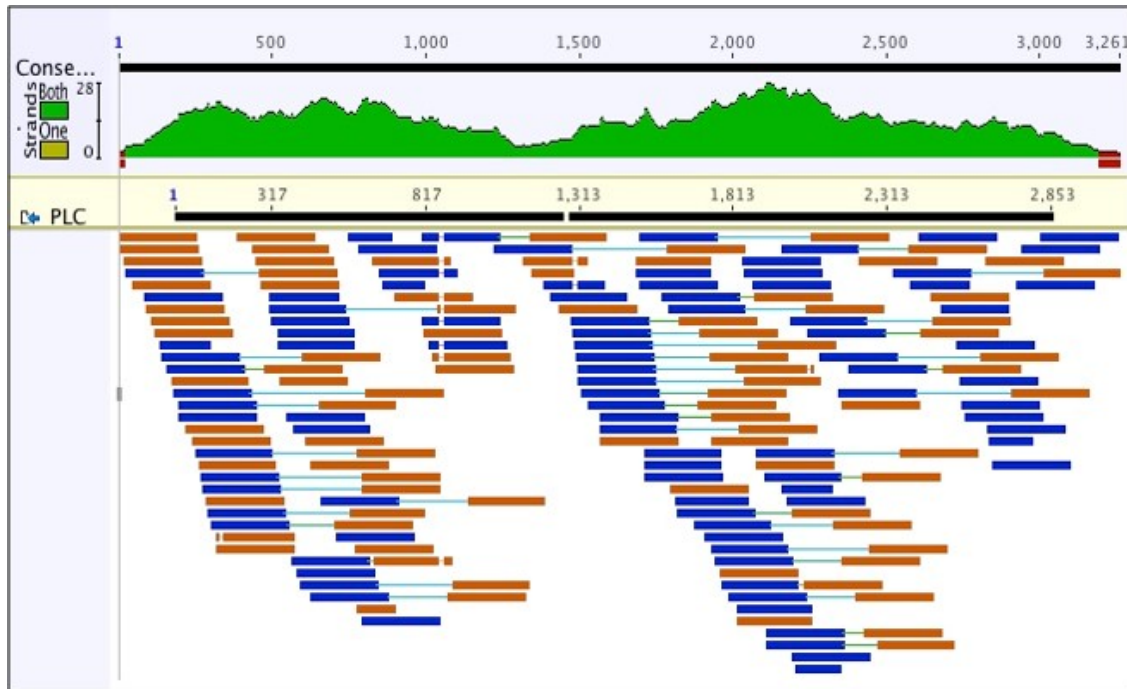


Figure 4.2 Alignment of reads to previously cloned genomic fragments

- Total number of N's = 0
- Sum (bp) no N's = 660984269
- Max contig size = 35431
- Min contig size = 222
- Average contig size = 1224
- N50 = 2197

Because of the paired end nature of the reads, it was possible to further assemble some of these contigs into scaffolds with **SSPACE**. The following results were obtained:

- Total number of **scaffolds** = 470492
- Sum (bp) = 660664090
- Max scaffold size = 57949
- Min scaffold size = 222
- Average scaffold size = 1404
- N50 = 2913

#### 4.2.4 Utility of Assembly for Annotation

In general, the lengths of the scaffolds are relatively short. While a subsequent phase of this genomic sequencing effort will address the issue of scaffold length, can this current version of the Tallapoosa darter genome assembly be utilized to begin an annotation of the genome? Obviously, those scaffolds that are above 5,000 nucleotides in length likely contain a gene or a significant part of a gene.

To check accuracy of assembly, the scaffolds were aligned to previously cloned Tallapoosa darter genomic fragments. In all cases the scaffolds aligned precisely to those genomic sequences.

To further check accuracy of assembly and the utility of scaffolds for annotation, the scaffolds were searched by blastn with several full length *Perca flavescens* mRNA sequences (closest related species to darters).

The examples below show two instances where genes were identified within the scaffolds. In the first example, depicted in figure 4.3(a), the Urate Oxidase gene was found to be contained within one scaffold. In the second example, shown in figures 4.3(b) and 4.3(c), we can see that the neprilysin (NEP1) gene actually spanned many scaffolds. These scaffolds were identified by a high degree of homology to different portions of the NEP1 mRNA sequence. These scaffolds were then concatenated in the order corresponding to NEP1 mRNA homology.

While the current assembly of the Tallapoosa darter genome based on a 12 fold coverage of PE250 reads produced scaffolds that are relatively short, it appears that the assembly is of sufficiently high quality to facilitate the start of darter genome annotation. WebApollo was chosen as the tool to carry out the annotation process of the Tallapoosa darter genome assembly.

#### 4.2.5 Setting up WebApollo

Initial attempts to annotate some of the Tallapoosa darter scaffolds were carried out utilizing the red line workflow on the DNA Subway website ([dnasubway.iplantcollaborative.org](http://dnasubway.iplantcollaborative.org)).

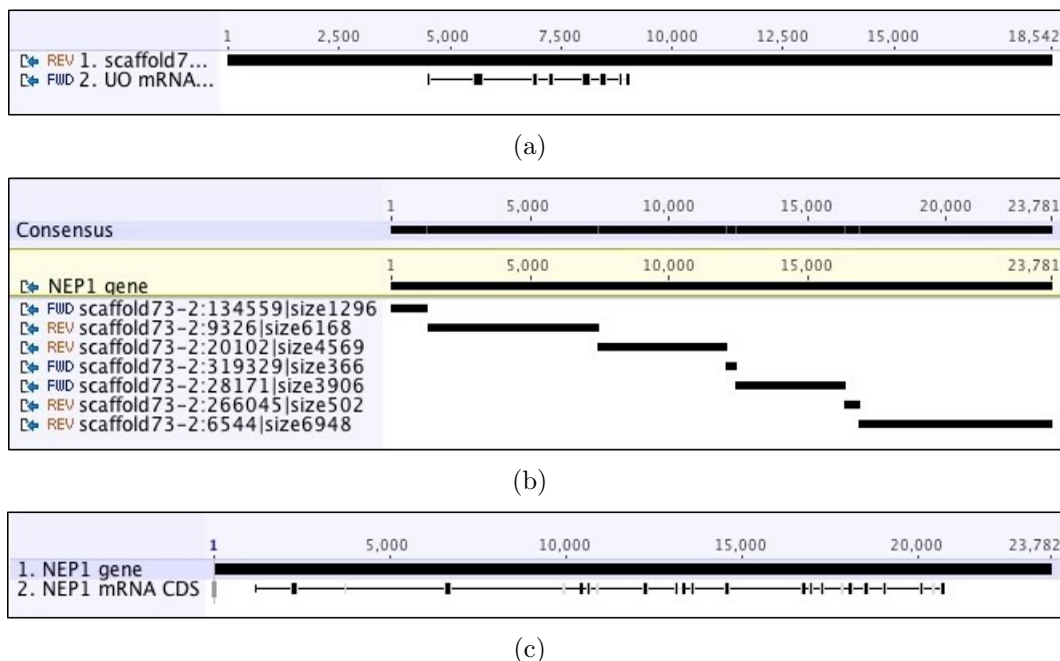


Figure 4.3 Examples of instances where genes were identified within the scaffolds: (a) Urate Oxidase contained within one scaffold. (b) - (c) Neprilysin (NEP1) gene spanning several scaffolds

The annotation workflow that proved highly successful was to begin with fgenesh derived gene models in Apollo, determine with blastp against the GenBank nr database if the gene model codes for a known protein and if a homolog exists, use the homologous protein as input for fgenesh+ determination of the exon/intron structure of the gene within the scaffold. The gene model was then adjusted according to the fgenesh+ derived model. It was decided, therefore, to enable the implementation of this workflow in WebApollo.

WebApollo is a fairly complex server side application to set up. However, a virtual machine implementation of WebApollo has been made available that is preconfigured and was easily incorporated into a VirtualBox running on a MacMini server. This makes it relatively easy to implement WebApollo by research groups lacking server administration expertise.

Once the WebApollo instance was installed, the longer Tallapoosa darter scaffolds were imported into WebApollo along with fgenesh derived gene models as evidence tracks. The



WebApollo virtual machine includes a script (*setup\_webapollo.sh*) that makes it simple for individuals with little computer system experience to create a database of scaffolds and to then add individual scaffolds with evidence tracks one at a time. A modified version of this script was utilized in setting up the Tallapoosa darter WebApollo instance along with additional scripts that were written to make the necessary file format conversions and enable an unattended import of all the desired scaffolds and evidence tracks into WebApollo.

The diagram presented in figure 4.4 shows the workflow that was implemented.

It is anticipated that as other research groups sequence the genomes of other darter species that these research groups will want to set up their own instances of WebApollo. Since many such research groups will likely not have the necessary server administration and unix expertise, a number of scripts were written that make it possible for individuals with very minimal unix experience to import scaffolds and fgenesh generated evidence tracks into WebApollo. These scripts are:

- *fgenesh – splitter.sh*
- *fgenesh – converter.sh*
- *add\_to\_webapollo.sh*

The purpose and use of these scripts is summarized in the previous workflow diagram. Of course, the use of these scripts and the associated workflow is not limited to setting up WebApollo instances of just darter genomes. These may also be of utility to other groups setting up WebApollo instances for annotation of genomes of other species.

#### 4.2.6 Tallapoosa Darter Genome Annotation with WebApollo

To begin annotation of a scaffold, a scaffold is selected from a list (figure 4.5).

Once the scaffold opens in the viewer, the fgenesh derived *ab initio* gene model(s) is/are displayed in an evidence track (see figure 4.6).

The gene model is slid up to the user area and the predicted amino acid sequence is obtained.

If a blastp search of GenBank shows a homologous protein (figure 4.8(a)), that protein

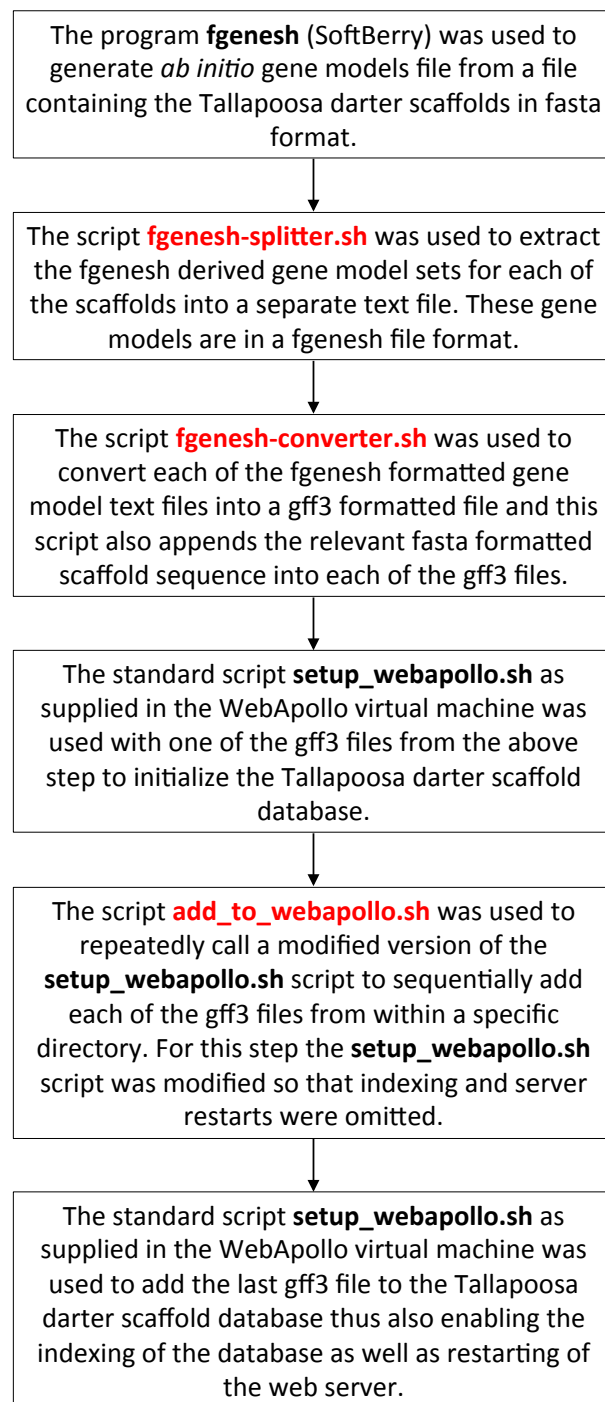
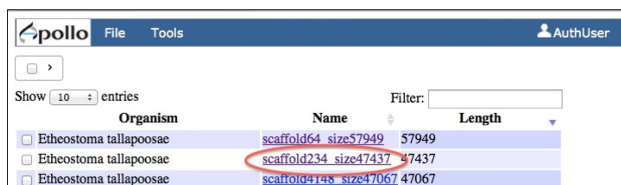


Figure 4.4 WebApollo Work Flow



Organism	Name	Length
Etheostoma tallapoosae	scaffold64_size57949	57949
Etheostoma tallapoosae	scaffold234_size47437	47437
Etheostoma tallapoosae	scaffold4148_size47067	47067

Figure 4.5 Scaffold selection

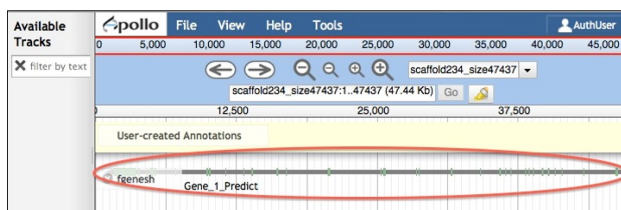


Figure 4.6 Gene model

sequence along with the scaffold DNA sequence is subject to fgenesh+ (SoftBerry) gene prediction analysis (figure 4.8(b)) and the gene model in WebApollo is adjusted accordingly and with additional manual adjustments as necessary (figure 4.8(c)).

The Tallapoosa darter scaffolds can be searched by BLAST and annotated in WebApollo at [www.dartergenomics.org](http://www.dartergenomics.org).

### Acknowledgements

Thanks to our collaborator Dr. Leos Kral from University of West Georgia, the leading contributor of this research. The scripts were written by me assisted by my colleague Charly Blanche Tamate. This work was presented at PAG and it was partially supported by NIFA award 2011-67016-30331.

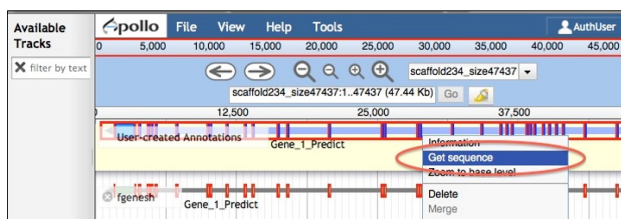
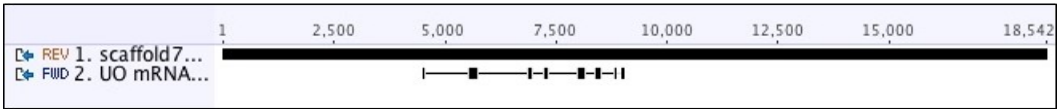
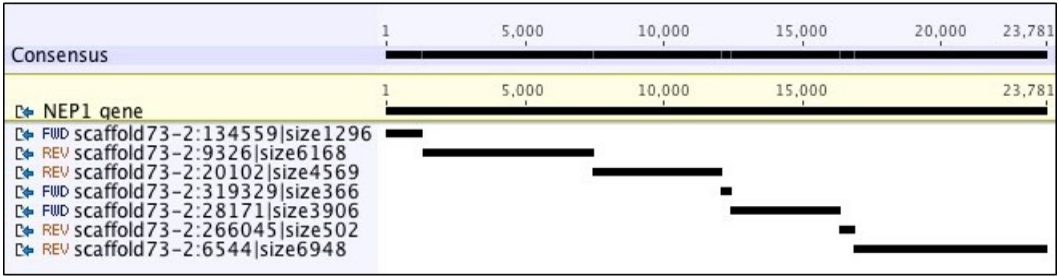


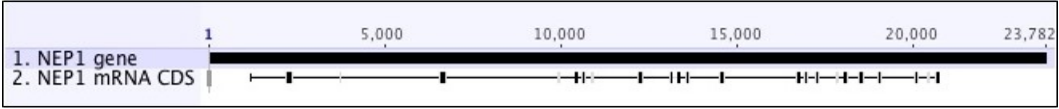
Figure 4.7 Amino Acid



(a)



(b)



(c)

Figure 4.8 Annotation with WebApollo: (a) Homologous protein. (b) Gene prediction analysis (c) Gene adjustment

## PART 5

### SOFTWARE PACKAGES

#### 5.1 Transcriptome Quantification

##### 5.1.1 SimReg

SimReg Source code: <http://alan.cs.gsu.edu/NGS/?q=adrian/simreg>

#### 5.2 Transcriptome Reconstruction

##### 5.2.1 MaLTA

The open source C++ implementation of MaLTA is freely available for download.

<http://alan.cs.gsu.edu/NGS/?q=malta>

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

#### 5.3 Genome Assembly and Annotation

##### 5.3.1 Etheostoma tallapoosae Genome

The Tallapoosa darter scaffolds can be searched by BLAST and annotated in WebApollo at [www.dartergenomics.org](http://www.dartergenomics.org).

## PART 6

### 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 advantage 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] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and David Haussler, “The human genome browser at ucsc,” *Genome Research*, vol. 12, no. 6, pp. 996–1006, 2002. [Online]. Available: <http://genome.cshlp.org/content/12/6/996.abstract>
- [2] F. Meyer, D. Paarmann, M. D’Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke *et al.*, “The metagenomics rast server—a public resource for the automatic phylogenetic and functional analysis of metagenomes,” *BMC bioinformatics*, vol. 9, no. 1, p. 386, 2008.
- [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. Khrebtkova, 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 genotyping 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. Jeddell, 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/doi/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] 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>
- [24] 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>
- [25] 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>
- [26] P. A. Pevzner, “1-Tuple DNA sequencing: computer analysis.” *J Biomol Struct Dyn*, vol. 7, no. 1, pp. 63–73, Aug. 1989.
- [27] 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>
- [28] J. Feng, W. Li, and T. Jiang, “Inference of isoforms from short sequence reads,” in *Proc. RECOMB*, 2010, pp. 138–157.

- [29] 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.
- [30] 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>
- [31] 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>
- [32] 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.
- [33] V. Lacroix, M. Sammeth, R. Guigo, and A. Bergeron, “Exact transcriptome reconstruction from short sequence reads,” in *Proc. WABI*, 2008, pp. 50–63.
- [34] T. Steijger, J. F. Abril, P. G. Engström, F. Kokocinski, T. J. H. The RGASP Consortium, R. Guig, J. Harrow, and P. Bertone, “Assessment of transcript reconstruction methods for RNA-Seq,” *Nature Methods*, vol. 10, pp. 1177–1184, 2013.
- [35] B. Li and C. Dewey, “Rsem: accurate transcript quantification from rna-seq data with or without a reference genome,” *BMC bioinformatics*, vol. 12, no. 1, p. 323, 2011.
- [36] W. Li and T. Jiang, “Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads,” *Bioinformatics*, vol. 28, no. 22, pp. 2914–2921, 2012.
- [37] 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.
- [38] A. I. Tomescu, A. Kuosmanen, R. Rizzi, and V. Mkinen, “A novel min-cost flow method for estimating transcript expression with rna-seq,” in *Proc. RECOMB-seq 2013*, 2013.
- [39] W. Li, J. Feng, and T. Jiang, “IsoLasso: A LASSO Regression Approach to RNA-Seq Based Transcriptome Assembly,” *Journal of Computational Biology*, vol. 18, pp. 1693–1707, 2011.
- [40] A. I. Tomescu, A. Kuosmanen, R. Rizzi, and V. Mäkinen, “A novel min-cost flow method for estimating transcript expression with rna-seq,” *BMC Bioinformatics*, vol. 14, no. S-5, p. S15, 2013.
- [41] 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>
- [42] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of Royal Statistical Society*, vol. 58, pp. 267–288, 1996.
- [43] 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.
- [44] 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>
- [45] 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>
- [46] 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>
- [47] 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>
- [48] 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>
- [49] 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>
- [50] 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>
- [51] 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>
- [52] 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>

- [53] F. E. Angly, D. Willner, F. Rohwer, P. Hugenholtz, and G. W. Tyson, “Grinder: a versatile amplicon and shotgun sequence simulator,” *Nucleic Acids Research*, vol. 40, no. 12, p. e94, 2012. [Online]. Available: <http://nar.oxfordjournals.org/content/40/12/e94.abstract>
- [54] 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>
- [55] 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.
- [56] 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>
- [57] UCSC Genome Database, <http://genome.ucsc.edu>.
- [58] CCDS Genome Database, <http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi>.
- [59] I. Astrovskaya, B. Tork, S. Mangul, K. Westbrook, 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>
- [60] 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.

- [61] 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>
- [62] 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.
- [63] IBM, “Inc: IBM ILOG CPLEX 12.1.” <http://www.ibm.com/software/integration/optimization/cplex/>, 2009.
- [64] 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.
- [65] N. A. Moran, “Symbiosis,” *Current Biology*, vol. 16, no. 20, pp. R866–R871, 2006. [Online]. Available: [http://www.cell.com/current-biology/abstract/S0960-9822\(06\)02212-3](http://www.cell.com/current-biology/abstract/S0960-9822(06)02212-3)
- [66] M. McFall-Ngai, M. G. Hadfield, T. C. G. Bosch, H. V. Carey, T. Domazet-Loo, A. E. Douglas, N. Dubilier, G. Eberl, T. Fukami, S. F. Gilbert, U. Hentschel, N. King, S. Kjelleberg, A. H. Knoll, N. Kremer, S. K. Mazmanian, J. L. Metcalf, K. Nealson, N. E. Pierce, J. F. Rawls, A. Reid, E. G. Ruby,

- M. Rumpho, J. G. Sanders, D. Tautz, and J. J. Wernegreen, "Animals in a bacterial world, a new imperative for the life sciences," *Proceedings of the National Academy of Sciences*, vol. 110, no. 9, pp. 3229–3236, 2013. [Online]. Available: <http://www.pnas.org/content/110/9/3229.abstract>
- [67] E. R. Haine, "Symbiont-mediated protection," *Proceedings of the Royal Society B: Biological Sciences*, vol. 275, no. 1633, pp. 353–361, 2008.
- [68] N. B. Lopanik, "Chemical defensive symbioses in the marine environment," *Functional Ecology*, vol. 28, no. 2, pp. 328–340, 2014. [Online]. Available: <http://dx.doi.org/10.1111/1365-2435.12160>
- [69] A. E. Trindade-Silva, G. E. Lim-Fong, K. H. Sharp, and M. G. Haygood, "Bryostatins: biological context and biotechnological prospects," *Current opinion in biotechnology*, vol. 21, no. 6, pp. 834–842, 2010.
- [70] J. Piel, "Metabolites from symbiotic bacteria," *Natural product reports*, vol. 26, no. 3, pp. 338–362, 2009.
- [71] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen *et al.*, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, no. 7057, pp. 376–380, 2005.
- [72] D. D. Sommer, A. L. Delcher, S. L. Salzberg, and M. Pop, "Minimus: a fast, lightweight genome assembler," *BMC bioinformatics*, vol. 8, no. 1, p. 64, 2007.
- [73] M. Schulz and D. Zerbino, "Oases-de novo transcriptome assembler for very short reads," *Published online: <http://www.ebi.ac.uk/zerbino/oases>*, 2010.