Identification and Functional Annotation of Alternatively Spliced Isoforms

by

Ridvan Eksi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2018

Doctoral Committee:

Assistant Professor Yuanfang Guan, Chair
Professor Matthias Kretzler
Assistant Research Scientist Rajasree Menon
Associate Professor Kayvan Najarian
Professor Gilbert S. Omenn

Ridvan Eksi

ridvan@umich.edu

ORCID ID: 0000-0003-2757-0057

© Ridvan Eksi 2018
All Rights Reserved

ACKNOWLEDGEMENTS

The completion of this thesis would not have been possible without the help and guidance from the brilliant people around me. It was a privilege and great honor for me to have met and interacted with these incredible individuals, to whom I would like to express my sincerest gratitude and appreciation.

I would first like to thank my advisor Dr. Yuanfang Guan for her guidance during my graduate studies. Her remarkable wisdom, passion, and enthusiasm for science were instrumental in my development as a scientist. She always has an open door for her students, and is dedicated to the success of her students. I would also like to thank the members of my thesis committee; Dr. Matthias Kretzler, Dr. Gilbert S. Omenn, Dr. Rajasree Menon and Dr. Kayvan Najarian. Their discussion and feedback have been invaluable for my thesis research.

I would like to thank all the past and current members of Guan Lab for their tremendous help. They were always there for me whenever I needed help, especially during earlier years of my graduate school. I would like to thank my friends, here in Michigan and around the world, for encouraging me and supporting me throughout my graduate career and my life.

I am grateful to Turkish Ministry of National Education for financially supporting me for the majority of my doctoral studies.

Above all, I would like to thank my family for their continued love and support, patience and understanding, and for always standing by me and believing in me. Their unconditional love has kept me going during all these years.

TABLE OF CONTENTS

LIST OF FIGURES

Figures

Supplementary Figures

LIST OF TABLES

Tables

Supplementary Tables

ABSTRACT

Alternative splicing is a key mechanism for increasing the complexity of transcriptome and proteome in eukaryotic cells. A large portion of multi-exon genes in humans undergo alternative splicing, and this can have significant functional consequences as the proteins translated from alternatively spliced mRNA might have different amino acid sequences and structures. The study of alternative splicing events has been accelerated by the next-generation sequencing technology. However, reconstruction of transcripts from short-read RNA sequencing is not sufficiently accurate. Recent progress in single-molecule long-read sequencing has provided researchers alternative ways to help solve this problem. With the help of both short and long RNA sequencing technologies, tens of thousands of splice isoforms have been catalogued in humans and other species, but relatively few of the protein products of splice isoforms have been characterized functionally, structurally and biochemically. The scope of this dissertation includes using short and long RNA sequencing reads together for the purpose of transcript reconstruction, and using high-throughput RNA-sequencing data and gene ontology functional annotations on gene level to predict functions for alternatively spliced isoforms in mouse and human.

In the first chapter, I give an introduction of alternative splicing and discuss the existing studies where next generation sequencing is used for transcript identification. Then, I define the isoform function prediction problem, and explain how it differs from better known gene function prediction problem. In the second chapter of this dissertation, I describe our study where the overall transcriptome of kidney is studied using both long reads from PacBio platform and RNA-

seq short reads from Illumina platform. We used short reads to validate full-length transcripts found by long PacBio reads, and generated two high quality sets of transcript isoforms that are expressed in glomerular and tubulointerstitial compartments. In the third chapter, I describe our generic framework, where we implemented and evaluated several related algorithms for isoform function prediction for mouse isoforms. We tested these algorithms through both computational evaluation and experimental validation of the predicted 'responsible' isoform(s) and the predicted disparate functions of the isoforms of *Cdkn2a* and of *Anxa6*. Our algorithm is the first effort to predict and differentiate isoform functions through large-scale genomic data integration. In the fourth chapter, I present the extension of isoform function prediction study to the protein coding isoforms in human. We used a similar multiple instance learning (MIL)-based approach for predicting the function of protein coding splice variants in human. We evaluated our predictions using literature evidence of ADAM15, LMNA/C, and DMXL2 genes. And in the fifth and final chapter, I give a summary of previous chapters and outline the future directions for alternatively spliced isoform reconstruction and function prediction studies.

CHAPTER 1

Introduction

1.1 Introduction

Alternative splicing is a process that enables a messenger RNA (mRNA) to synthesize different protein variants. It happens by rearranging the pattern of intron and exon elements that are joined by splicing to modify the mRNA sequence. Alternative splicing is a key mechanism for increasing the complexity of transcriptome and proteome in eukaryotic cells. It is estimated that more than 95% of multi-exon genes in the human undergo alternative splicing [1, 2]. Alternative splicing can have significant functional consequences as the proteins translated from alternatively spliced mRNA might have different amino acid sequences and structures. Functional consequences of alternative splicing include but not limited to alterations in mRNA decay [3] and changes in the translation process [4].

The study of alternative splicing events has been accelerated by the next-generation sequencing technology. RNA sequencing (RNA-seq) is a technology that enables identification of novel genes and splice variants and estimation of transcript abundances [5-7]. Sequencing reads are produced by RNA-seq from the expressed transcripts. Correctly assigning these reads to transcripts and assembling them into full-length expressed transcripts are two major computational challenges and crucial steps for transcript quantification and differential expression analysis. Thus, they have an important role in understanding tissue-specific splicing and the regulation of gene expressions [8].

Transcript assembly remains an open and challenging problem, especially for complex transcriptomes with high diversity of splice variants. Several studies have shown that reconstruction of transcripts from short-read RNA sequencing is not sufficiently accurate [9, 10]. Recent progress in single-molecule long-read sequencing has provided researchers alternative ways to help solve this problem. Pacific Biosciences (PacBio) has introduced the sequencing technique that allows the sequencing of transcripts without fragmentation or PCR amplification. The sequencing length is enough to cover size distribution of most transcripts in eukaryotes. Thus, PacBio's full-length or nearly full-length transcripts eliminate the need of assembly for the downstream analysis. However, transcript assembly or quantification is merely the first steps of gaining insights into complex transcriptomes.

With the help of rapid progression of RNA sequencing technologies, tens of thousands of splice isoforms have been catalogued in humans and other species. Relatively few of the protein products of splice isoforms have been characterized functionally, structurally and biochemically. During the past few decades, significant progress has been made for gene function prediction problem. A major intellectual limitation of the gene function prediction paradigm is that it considers a gene as a single entity without differentiating the functional diversity of alternatively spliced isoforms. Attempts to capture differential functions of alternatively spliced isoforms are limited to low-throughput experimental approaches. We expect that computational methods such as the ones that will be explained in the later chapters developed to differentiate isoform functions through integrating functional genomic data will assist a deeper, high-resolution understanding of protein functions.

This dissertation focuses on identification and functional annotation of alternatively spliced isoforms. In this chapter, I will introduce the alternative splicing and how RNA

sequencing is used to reconstruct transcripts with different splicing events. I will talk about differences of using short vs. long reads in terms of transcript reconstruction. Then, I will introduce the isoform function prediction problem and how it differs from the previous gene function prediction studies. Finally, I will give an overview of the following chapters.

1.2 Background

1.2.1 Alternative splicing

Transcription is the first step of gene expression. During transcription, a particular segment of DNA is copied into mRNA which is catalyzed by RNA polymerase enzyme. However, in eukaryotes, before the mRNA is translated into proteins, non-coding portions of the sequence, which are called introns, are removed and protein-coding parts, which are called exons, are joined by RNA splicing process to produce a mature mRNA.

RNA splicing was first discovered in viruses [11, 12] and later in eukaryotes [13]. Shortly after, alternative patterns of pre-mRNA splicing which generates different mature mRNAs containing different combinations of exons from a single precursor mRNA are discovered. This process by which exons or portions of exons or introns within a pre-mRNA transcript are differentially joined or skipped is called alternative splicing. Alternative splicing results in multiple protein products that are encoded by a single gene and it generates a great amount of protein diversity in eukaryotes.

Alternative splicing mechanism has been studied extensively [14-16]. It is regulated by *cis*-acting factors within pre-mRNAs and *trans*-acting factors as well as by the secondary structure of the pre-mRNA transcript. These factors together form the "splicing code" and determine the cell type-specific splicing [17]. The 5′ splice site, the 3′ splice site, exonic or intronic splicing enhancers and silencers are the essential *cis*-acting factors [18, 19]. *Trans*-acting

factors regulate alternative splicing by associating with *cis*-acting elements [19, 20]. There are five basic types of alternative splicing events, namely exon skipping, mutually exclusive exons, alternative 5’ donor sites, alternative 3’ acceptor sites, and intron retention. In the first type, a complete exon is spliced out or retained. In the second type, only one of two consecutive exons is retained in a mutually exclusive manner. In the third type, an alternative 5' donor site is used which changes the 3' boundary of the upstream exon. In the fourth type, an alternative 3' acceptor site is used, which changes the 5' boundary of the downstream exon. And finally, in the fifth type, a sequence may be spliced out or simply retained in the final mRNA. This is different from exon skipping because the retained sequence is not flanked by introns. In addition to these five primary modes of alternative splicing, there are two other mechanisms by which distinct mRNAs may be generated from the same gene; alternative promoter usage and alternative polyadenylation sites. With the former mode, transcripts with different 5’-most exons are generated, whereas with the latter mode transcripts with different 3’ end points are generated.

Alternative splicing may have significant functional consequences because it can modify protein-protein interactions [21], protein sequence and consequently the domains in the final protein product [22]. Moreover, it has been shown that alternative splicing may alter mRNA decay [3] and the translation process [4].

1.2.2 RNA sequencing and transcript assembly

A comprehensive way to measure transcriptome composition and to discover new exons or genes is through high-throughput sequencing of cDNA which is referred as RNA-seq. It is a well-established technology that enables measurements of expression abundances as well as identification of novel genes and splice variants. Reads that are produced by RNA-seq are

sampled from the expressed transcripts. Assembling these reads so that the full-length expressed transcripts can be accurately reconstructed is a major computational challenge.

Existing transcript assembly strategies can be organized into two categories. First category is the de novo transcript assembly which is the direct assembly of sequenced reads into transcripts without mapping to a reference genome. Several software packages have been shown to generate contig sets reconstructing most of the expressed transcripts correctly. TransABySS[23], Trinity[24], and Oases[25] are the three of most widely used de novo assemblers. De novo methods are generally used for non-model species where a complete reference genome is not available and for cancer samples where reference genome is significantly diverged.

The second category of transcript assembly methods is the reference-based methods, which require aligning the short reads to a reference genome. Reference-based methods generally start with using the alignments generated by one of the RNA-seq aligners (TopHat2 [26], STAR [27], SpliceMap [28]) to create a splice graph for each locus, and then reconstruct the transcripts by decomposing the graph. Cufflinks [29], Scripture [30], IsoLasso [31], CIDANE [32] are some of the widely used reference-based transcript assemblers. Reference- based methods generally have better accuracy when compared to de novo assemblers when a high-quality reference genome is available. These transcript assemblers get less accurate as the transcriptome under study is getting more complex. Their accuracy is low when multiple splice forms, sequencing and alignment errors are present.

1.2.3 PacBio RNA sequencing as an alternative to assembly

As explained in the previous section, development of the next-generation sequencing technology has accelerated the study of alternative splicing events. It provided new opportunities

to investigate the complexity of the transcriptomes in mammals. However, because of the limited length of RNA-seq sequences from Illumina, it is still difficult to reliably detect the alternative splicing events using only short reads. It is especially challenging to detect intron retention and alternative polyadenylation events. The retained introns that are detected by short reads cannot be reliably distinguished from the contamination of DNA [1, 33, 34]. In the case of alternative polyadenylation, it is difficult to differentiate real polyadenylation sites from internal priming events [35-40].

Another limitation of short reads arises for the case of detecting the correct combination of distant alternative splicing events on a transcript. Although, some earlier studies reported some cases of correlated inclusive or mutually exclusive splicing events [41-43], a more recent study claims that combination of alternative splicing events is largely independent along the transcript [44]. Thus, it remains challenging to investigate the combination of distant alternative splicing events only using short reads. The distance between two alternative splicing events is generally larger than the insert size of the fragments that short reads are produced from. RNA-seq with short reads could only be used to investigate alternative splicing events that are nearby, within the insert size interval. In short, some novel alternative splicing events such as intron retention and alternative polyadenylation events could not be reliably detected by only using short reads. Moreover, it is difficult to position distant alternative splicing events along the transcript due to limited insert size of Illumina libraries. Considering the fact that these novel alternative splicing events as well as the novel combinations of these events on a transcript may lead to generation of thousands of novel isoforms, one can argue that transcript assembly studies using only short reads from Illumina RNA-seq will miss a considerable number of transcripts present in the sample.

Pacific Biosciences (PacBio) has introduced the sequencing technique that is based on single-molecule sequencing chemistry with real-time detection. It allows the sequencing of long sections of genomic DNAs or transcripts without fragmentation or PCR amplification. The sequencing length of the PacBio platform can go up to 10 kb, which is enough to cover size distribution of most transcripts in eukaryotes. Thus, PacBio's full-length or nearly full-length transcripts eliminate the need of assembly for the downstream analysis. PacBio's long RNA sequences can reliably differentiate intron retention and alternative polyadenylation events, as well as it can reliably position the distant alternative splicing events on the transcript.

A drawback of the PacBio platform is its high sequence error rate. However, these errors show no context-specific bias, thus randomly distributed across the reads. Recently improved read-length and base-calling algorithms of the PacBio's software platform and the use of circular molecules have mitigated the high error rate problem. When read length exceeds the length of the cDNA template, each base pair is covered on both strands multiple times and these low-quality base calls are combined to derive high-quality reads.

In the recent years, PacBio long-read transcriptome sequencing platform has been successfully applied to human and other species for the purpose of detecting novel alternative splicing events and novel transcripts that are otherwise difficult to distinguish with short reads. Many novel alternative splicing events and novel transcripts have been detected in various organisms including but not limited to circovirus [45], sugarcane [46], barley [47], chicken [48] and rabbit [49]. Zhang et al. used long read RNA sequencing to study isoform evolution in primates. They claim that they substantially expanded the repertoire of alternative RNA processing events in primates, and found that intron retention and alternative polyadenylation are more prevalent in primates than previously estimated [44]. O'Grady et al. reported a new

workflow where they integrated three sequencing technologies to overcome limitations to transcript structure resolution. Their approach integrates PacBio long read sequencing with deepCAGE data to identify and validate transcript 5′ ends and Illumina short-read RNA-Seq data to identify and validate splice junctions and 3′ ends [50].

1.2.4 Isoform function prediction paradigm

Determining the functions of proteins is fundamental for understanding the molecular basis of diverse genetic diseases and is one of the central goals of genetics [51-57]. Traditionally, functional annotation studies have been centered on genes. Knock-out or knock-down experiments are generally performed for functional evaluation, and the resulting functional associations are resolved at the gene level and accumulated in databases such as Gene Ontology (GO) [58] and KEGG pathways [59]. Moreover, functional genomic data are usually resolved at the gene level. For example, the expression levels from a microarray study are generally assigned to genes, and most often gene-gene interactions are recorded in physical interaction datasets. With the exponential growth of such functional genomic data during the past decades, many computational studies have emerged for predicting gene functions [60-64]. These studies utilized many different machine learning methods including but not limited to Bayesian network-based methods [65], kernel-based classifiers [61], and decision trees [66].

A major limitation of the traditional gene function prediction paradigm is that it considers a gene as a single entity without differentiating the functional diversity of alternatively spliced isoforms. However, many studies support the fact that isoforms that are produced by alternative splicing may carry out different or sometimes opposing biological functions. For example, alternative splicing of the B cell lymphoma-x gene (Bcl-x) produces two variants, anti-apoptotic Bcl-xL and pro-apoptotic Bcl-xS [67]. Bcl-xS is missing two Bcl-2 family motifs, BH1 and

BH2, which may account for functional differences. Another example of such case is the carboxypeptidase E (CPE) gene in the Wnt signal transduction pathway [68]. This pathway has a key role in many diseases including cancers and particularly colorectal cancer. It has been shown that the alternative splice variant form of CPE activates the Wnt signal, but the full-length canonical CPE is an inhibitor of Wnt/B-catenin signaling [68]. Moreover, abnormal splicing events can lead to disease by creating isoforms with non-functional structures or by changing isoform expression levels [18, 69-71]. P53β is an isoform of the human tumor protein p53 (TP53) gene, it can increase p53 target gene expression and it is apoptotic, whereas the isoform Δ133p53 inhibits p53-mediated apoptosis [72]. These specific examples represent only a small portion of the alternatively spliced isoforms with different functions.

The computational approaches to gene function prediction problem are typically starting point for experimental validation of functions. However, experimentally validating functions for isoforms is difficult for several reasons. Isoform-specific antibodies are generally not available to be used in experiments like Western blot. Moreover, isoform-specific oligonucleotides are difficult to design because many isoforms of the same gene differ only by a short sequence fragment. This makes it difficult to perform isoform-specific qPCR and RNA interference experiments. Although there are several small scale experimental functional studies focusing on individual isoforms [67, 73, 74], computational prediction is often the only option for genome-wide scale functional studies for isoforms because of the experimental limitations. Our study that is explained in Chapter 3 represents the first effort to predict and differentiate functions for isoforms through large-scale genomic data integration. Later, a similar method was used to predict isoform functions in humans by integrating RNA-seq data into co-expression networks and propagating isoform-level labels based on the same multiple instance learning concept [75].

They assess the accuracy of their model with cross validation of single-isoforms genes which can introduce bias in the performance estimate, because multi-isoform genes were not included. They also gave five isoforms of *TP53* as an example, for which the known apoptosis regulation function as well as the regulation direction were predicted correctly [75]. More recently, Luo et al. proposed a novel approach to differentiate functions of protein coding isoforms by integrating sparse simplex projection, which is a nonconvex sparsity-inducing regularizer, with the multiple instance learning framework [76].

1.3 Dissertation overview

Alternative splicing is a key mechanism for increasing the complexity of transcriptome and proteome in eukaryotic cells. A large portion of multi-exon genes in humans undergo alternative splicing, and this can have significant functional consequences. The study of alternative splicing events has been accelerated by the next-generation sequencing technology. However, reconstruction of transcripts from short-read RNA sequencing is not sufficiently accurate [9, 10]. Moreover, relatively few of the protein products of splice isoforms have been characterized functionally. The goal of this dissertation is to address abovementioned issues for identification and functional annotation of alternatively spliced isoforms.

In Chapter 2, we describe our study where the overall transcriptome of glomerular and tubulointerstitial compartments of a kidney is studied using both long reads from PacBio platform and short reads from Illumina platform. We used short reads to validate full-length transcripts found by long PacBio reads, and generated two high quality sets of transcript isoforms that are expressed in glomerular and tubulointerstitial compartments. Then, we compared the confirmed transcript isoforms to the set of known transcript isoforms in GENCODE and provided the final list of expressed transcript isoforms along with their

annotation status to be used by researchers in downstream kidney transcriptome studies. Integrating data from two different sequencing technologies allowed us to identify and validate nearly 14k known transcripts in tubulointerstitial and glomerular compartments. In addition to that, we identified and validated nearly 12k and 8k multi-exon potential novel transcripts from each compartment respectively.

In Chapter 3, we present our generic framework that interrogates public RNA-seq data at the transcript level to differentiate functions for alternatively spliced isoforms. For a specific function, our algorithm identifies the 'responsible' isoform(s) of a gene and generates classifying models at the isoform level instead of at the gene level. Through cross-validation, we demonstrated that our algorithm is effective in assigning functions to genes, especially the ones with multiple isoforms, and robust to gene expression levels and removal of homologous gene pairs. We identified genes in the mouse whose isoforms are predicted to have distinct functionalities and experimentally validated the 'responsible' isoforms using data from mammary tissue. With protein structure modeling and experimental evidence, we further validated the predicted isoform functional differences for the genes *Cdkn2a* and *Anxa6*. Our generic framework is the first to predict and differentiate functions for alternatively spliced isoforms, instead of genes, using genomic data. It is extendable to any base machine learner and other species with alternatively spliced isoforms, and shifts the current gene-centered function prediction to isoform-level predictions.

In Chapter 4, we describe the function prediction study for human protein coding isoforms. We used a multiple instance learning based approach for predicting the function of protein coding splice variants. We used transcript-level expression values and gene-level functional associations from the Gene Ontology database. A support vector machine (SVM)-

based 5-fold cross-validation was applied for computational evaluation. Comparatively, genes with multiple protein coding isoforms performed better than single protein coding isoform genes, and performance also improved when more examples were available to train the models. We demonstrated our predictions using literature evidence of ADAM15, LMNA/C, and DMXL2 genes.

1.4 Bibliography

1. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470-6. doi: 10.1038/nature07509. PubMed PMID: 18978772; PubMed Central PMCID: PMCPMC2593745.

2. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40(12):1413-5. doi: 10.1038/ng.259. PubMed PMID: 18978789.

3. Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A. 2003;100(1):189-92. doi: 10.1073/pnas.0136770100. PubMed PMID: 12502788; PubMed Central PMCID: PMCPMC140922.

4. Sanford JR, Gray NK, Beckmann K, Caceres JF. A novel role for shuttling SR proteins in mRNA translation. Genes Dev. 2004;18(7):755-68. doi: 10.1101/gad.286404. PubMed PMID: 15082528; PubMed Central PMCID: PMCPMC387416.

5. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621-8. doi: 10.1038/nmeth.1226. PubMed PMID: 18516045.

6. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008;133(3):523-36. doi: 10.1016/j.cell.2008.03.029. PubMed PMID: 18423832; PubMed Central PMCID: PMCPMC2723732.

7. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63. doi: 10.1038/nrg2484. PubMed PMID: 19015660; PubMed Central PMCID: PMCPMC2949280.

8. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464(7289):768-72. doi: 10.1038/nature08872. PubMed PMID: 20220758; PubMed Central PMCID: PMCPMC3089435.

9. Tilgner H, Raha D, Habegger L, Mohiuddin M, Gerstein M, Snyder M. Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. G3 (Bethesda). 2013;3(3):387-97. doi: 10.1534/g3.112.004812. PubMed PMID: 23450794; PubMed Central PMCID: PMCPMC3583448.

10. Steijger T, Abril JF, Engstrom PG, Kokocinski F, Consortium R, Hubbard TJ, et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013;10(12):1177-84. doi: 10.1038/nmeth.2714. PubMed PMID: 24185837; PubMed Central PMCID: PMCPMC3851240.

11. Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A. 1977;74(8):3171-5. PubMed PMID: 269380; PubMed Central PMCID: PMCPMC431482.

12. Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell. 1977;12(1):1-8. PubMed PMID: 902310.

13. Berget SM. Exon recognition in vertebrate splicing. J Biol Chem. 1995;270(6):2411-4. PubMed PMID: 7852296.

14. Hiller M, Zhang Z, Backofen R, Stamm S. Pre-mRNA secondary structures influence exon recognition. PLoS Genet. 2007;3(11):e204. doi: 10.1371/journal.pgen.0030204. PubMed PMID: 18020710; PubMed Central PMCID: PMCPMC2077896.

15. Yu Y, Maroney PA, Denker JA, Zhang XH, Dybkov O, Luhrmann R, et al. Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. Cell. 2008;135(7):1224-36. doi: 10.1016/j.cell.2008.10.046. PubMed PMID: 19109894; PubMed Central PMCID: PMCPMC2645801.

16. Feng Y, Chen M, Manley JL. Phosphorylation switches the general splicing repressor SRp38 to a sequence-specific activator. Nat Struct Mol Biol. 2008;15(10):1040-8. doi: 10.1038/nsmb.1485. PubMed PMID: 18794844; PubMed Central PMCID: PMCPMC2668916.

17. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. Nature. 2010;465(7294):53-9. doi: 10.1038/nature09000. PubMed PMID: 20445623.

18. Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet. 2007;8(10):749-61. doi: 10.1038/nrg2164. PubMed PMID: 17726481.

19. Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem. 2003;72:291-336. doi: 10.1146/annurev.biochem.72.121801.161720. PubMed PMID: 12626338.

20. Zhou Z, Fu XD. Regulation of splicing by SR proteins and SR protein-specific kinases. Chromosoma. 2013;122(3):191-207. doi: 10.1007/s00412-013-0407-z. PubMed PMID: 23525660; PubMed Central PMCID: PMCPMC3660409.

21. Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, et al. Tissue-specific alternative splicing remodels protein-protein interaction networks. Mol Cell. 2012;46(6):884-92. doi: 10.1016/j.molcel.2012.05.037. PubMed PMID: 22749401.

22. Light S, Elofsson A. The impact of splicing on protein domain architecture. Curr Opin Struct Biol. 2013;23(3):451-8. doi: 10.1016/j.sbi.2013.02.013. PubMed PMID: 23562110.

23. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010;7(11):909-12. doi: 10.1038/nmeth.1517. PubMed PMID: 20935650.

24. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644-52. doi: 10.1038/nbt.1883. PubMed PMID: 21572440; PubMed Central PMCID: PMCPMC3571712.

25. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28(8):1086-92. doi: 10.1093/bioinformatics/bts094. PubMed PMID: 22368243; PubMed Central PMCID: PMCPMC3324515.

26. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36. doi: 10.1186/gb-2013-14-4-r36. PubMed PMID: 23618408; PubMed Central PMCID: PMCPMC4053844.

27. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. doi: 10.1093/bioinformatics/bts635. PubMed PMID: 23104886; PubMed Central PMCID: PMCPMC3530905.

28. Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res. 2010;38(14):4570-8. doi: 10.1093/nar/gkq211. PubMed PMID: 20371516; PubMed Central PMCID: PMCPMC2919714.

29. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511-5. doi: 10.1038/nbt.1621. PubMed PMID: 20436464; PubMed Central PMCID: PMCPMC3146043.

30. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010;28(5):503-10. doi: 10.1038/nbt.1633. PubMed PMID: 20436462; PubMed Central PMCID: PMCPMC2868100.

31. Li W, Feng J, Jiang T. IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. J Comput Biol. 2011;18(11):1693-707. doi: 10.1089/cmb.2011.0171. PubMed PMID: 21951053; PubMed Central PMCID: PMCPMC3216102.

32. Canzar S, Andreotti S, Weese D, Reinert K, Klau GW. CIDANE: comprehensive isoform discovery and abundance estimation. Genome Biol. 2016;17:16. doi: 10.1186/s13059-015-0865-0. PubMed PMID: 26831908; PubMed Central PMCID: PMCPMC4734886.

33. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, et al. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. 2014;24(11):1774-86. doi: 10.1101/gr.177790.114. PubMed PMID: WOS:000344442000006.

34. Li Y, Rao X, Mattox WW, Amos CI, Liu B. RNA-Seq Analysis of Differential Splice Junction Usage and Intron Retentions by DEXSeq. PLoS One. 2015;10(9):e0136653. doi: 10.1371/journal.pone.0136653. PubMed PMID: 26327458; PubMed Central PMCID: PMCPMC4556662.

35. Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. Patterns of variant polyadenylation signal usage in human genes. Genome Res. 2000;10(7):1001-10. PubMed PMID: 10899149; PubMed Central PMCID: PMCPMC310884.

36. Nam DK, Lee S, Zhou G, Cao X, Wang C, Clark T, et al. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. Proc Natl Acad Sci U S A. 2002;99(9):6152-6. doi: 10.1073/pnas.092140899. PubMed PMID: 11972056; PubMed Central PMCID: PMCPMC122918.

37. Fu Y, Sun Y, Li Y, Li J, Rao X, Chen C, et al. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Res. 2011;21(5):741-7. doi: 10.1101/gr.115295.110. PubMed PMID: 21474764; PubMed Central PMCID: PMCPMC3083091.

38. Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA. 2011;17(4):761-72. doi: 10.1261/rna.2581711. PubMed PMID: 21343387; PubMed Central PMCID: PMCPMC3062186.

39. Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, et al. A quantitative atlas of polyadenylation in five mammals. Genome Res. 2012;22(6):1173-83. doi: 10.1101/gr.132563.111. PubMed PMID: 22454233; PubMed Central PMCID: PMCPMC3371698.

40. Wilkening S, Pelechano V, Jarvelin AI, Tekkedil MM, Anders S, Benes V, et al. An efficient method for genome-wide polyadenylation site mapping and RNA quantification. Nucleic Acids Res. 2013;41(5):e65. doi: 10.1093/nar/gks1249. PubMed PMID: 23295673; PubMed Central PMCID: PMCPMC3597643.

41. Ubby I, Bussani E, Colonna A, Stacul G, Locatelli M, Scudieri P, et al. TMEM16A alternative splicing coordination in breast cancer. Mol Cancer. 2013;12:75. doi: 10.1186/1476-4598-12-75. PubMed PMID: 23866066; PubMed Central PMCID: PMCPMC3728142.

42. Schreiner D, Nguyen TM, Russo G, Heber S, Patrignani A, Ahrne E, et al. Targeted combinatorial alternative splicing generates brain region-specific repertoires of neurexins. Neuron. 2014;84(2):386-98. doi: 10.1016/j.neuron.2014.09.011. PubMed PMID: 25284007.

43. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nat Biotechnol. 2015;33(7):736-42. doi: 10.1038/nbt.3242. PubMed PMID: 25985263; PubMed Central PMCID: PMCPMC4832928.

44. Zhang SJ, Wang C, Yan S, Fu A, Luan X, Li Y, et al. Isoform Evolution in Primates through Independent Combination of Alternative RNA Processing Events. Mol Biol Evol. 2017;34(10):2453-68. doi: 10.1093/molbev/msx212. PubMed PMID: 28957512.

45. Moldovan N, Balazs Z, Tombacz D, Csabai Z, Szucs A, Snyder M, et al. Multi-platform analysis reveals a complex transcriptome architecture of a circovirus. Virus Res. 2017;237:37-46. doi: 10.1016/j.virusres.2017.05.010. PubMed PMID: 28549855.

46. Hoang NV, Furtado A, Mason PJ, Marquardt A, Kasirajan L, Thirugnanasambandam PP, et al. A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. BMC Genomics. 2017;18(1):395. doi: 10.1186/s12864-017-3757-8. PubMed PMID: 28532419; PubMed Central PMCID: PMCPMC5440902.

47. Dai F, Wang X, Zhang XQ, Chen Z, Nevo E, Jin G, et al. Assembly and analysis of a qingke reference genome demonstrate its close genetic relation to modern cultivated barley. Plant Biotechnol J. 2017. doi: 10.1111/pbi.12826. PubMed PMID: 28871634.

48. Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics. 2017;18(1):323. doi: 10.1186/s12864-017-3691-9. PubMed PMID: 28438136; PubMed Central PMCID: PMCPMC5404281.

49. Chen SY, Deng F, Jia X, Li C, Lai SJ. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. Sci Rep. 2017;7(1):7648. doi: 10.1038/s41598-017-08138-z. PubMed PMID: 28794490; PubMed Central PMCID: PMCPMC5550469.

50. O'Grady T, Wang X, Honer Zu Bentrup K, Baddoo M, Concha M, Flemington EK. Global transcript structure resolution of high gene density genomes through multi-platform data integration. Nucleic Acids Res. 2016;44(18):e145. doi: 10.1093/nar/gkw629. PubMed PMID: 27407110; PubMed Central PMCID: PMCPMC5062983.

51. Schmitz R, Young RM, Ceribelli M, Jhavar S, Xiao W, Zhang M, et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. Nature. 2012;490(7418):116-20. doi: 10.1038/nature11378. PubMed PMID: 22885699; PubMed Central PMCID: PMCPMC3609867.

52. Guan Y, Ackert-Bicknell CL, Kell B, Troyanskaya OG, Hibbs MA. Functional genomics complements quantitative genetics in identifying disease-gene associations. PLoS Comput Biol. 2010;6(11):e1000991. doi: 10.1371/journal.pcbi.1000991. PubMed PMID: 21085640; PubMed Central PMCID: PMCPMC2978695.

53. Chen KF, Crowther DC. Functional genomics in Drosophila models of human disease. Brief Funct Genomics. 2012;11(5):405-15. doi: 10.1093/bfgp/els038. PubMed PMID: 22914042.

54. Liang H, Cheung LW, Li J, Ju Z, Yu S, Stemke-Hale K, et al. Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer. Genome Res. 2012;22(11):2120-9. doi: 10.1101/gr.137596.112. PubMed PMID: 23028188; PubMed Central PMCID: PMCPMC3483541.

55. Nelson AC, Pillay N, Henderson S, Presneau N, Tirabosco R, Halai D, et al. An integrated functional genomics approach identifies the regulatory network directed by brachyury (T) in chordoma. J Pathol. 2012;228(3):274-85. doi: 10.1002/path.4082. PubMed PMID: 22847733.

56. Zhang X, Cowper-Sal lari R, Bailey SD, Moore JH, Lupien M. Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. Genome Res. 2012;22(8):1437-46. doi: 10.1101/gr.135665.111. PubMed PMID: 22665440; PubMed Central PMCID: PMCPMC3409257.

57. Consortium EP. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011;9(4):e1001046. doi: 10.1371/journal.pbio.1001046. PubMed PMID: 21526222; PubMed Central PMCID: PMCPMC3079585.

58. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9. doi: 10.1038/75556. PubMed PMID: 10802651; PubMed Central PMCID: PMCPMC3037419.

59. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32(Database issue):D277-80. doi: 10.1093/nar/gkh063. PubMed PMID: 14681412; PubMed Central PMCID: PMCPMC308797.

60. Hu P, Bader G, Wigle DA, Emili A. Computational prediction of cancer-gene function. Nat Rev Cancer. 2007;7(1):23-34. doi: 10.1038/nrc2036. PubMed PMID: 17167517.

61. Guan Y, Myers CL, Hess DC, Barutcuoglu Z, Caudy AA, Troyanskaya OG. Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biol. 2008;9 Suppl 1:S3. doi: 10.1186/gb-2008-9-s1-s3. PubMed PMID: 18613947; PubMed Central PMCID: PMCPMC2447537.

62. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics. 2003;19 Suppl 1:i197-204. PubMed PMID: 12855458.

63. Wu H, Su Z, Mao F, Olman V, Xu Y. Prediction of functional modules based on comparative genome analysis and Gene Ontology application. Nucleic Acids Res. 2005;33(9):2822-37. doi: 10.1093/nar/gki573. PubMed PMID: 15901854; PubMed Central PMCID: PMCPMC1130488.

64. Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, et al. The functional landscape of mouse gene expression. J Biol. 2004;3(5):21. doi: 10.1186/jbiol16. PubMed PMID: 15588312; PubMed Central PMCID: PMCPMC549719.

65. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci U S A. 2003;100(14):8348-53. doi: 10.1073/pnas.0832373100. PubMed PMID: 12826619; PubMed Central PMCID: PMCPMC166232.

66. Schietgat L, Vens C, Struyf J, Blockeel H, Kocev D, Dzeroski S. Predicting gene function using hierarchical multi-label decision tree ensembles. BMC Bioinformatics. 2010;11:2. doi: 10.1186/1471-2105-11-2. PubMed PMID: 20044933; PubMed Central PMCID: PMCPMC2824675.

67. Revil T, Toutant J, Shkreta L, Garneau D, Cloutier P, Chabot B. Protein kinase C-dependent control of Bcl-x alternative splicing. Mol Cell Biol. 2007;27(24):8431-41. doi: 10.1128/MCB.00565-07. PubMed PMID: 17923691; PubMed Central PMCID: PMCPMC2169420.

68. Skalka N, Caspi M, Caspi E, Loh YP, Rosin-Arbesfeld R. Carboxypeptidase E: a negative regulator of the canonical Wnt signaling pathway. Oncogene. 2013;32(23):2836-47. doi: 10.1038/onc.2012.308. PubMed PMID: 22824791; PubMed Central PMCID: PMCPMC3676431.

69. Srebrow A, Kornblihtt AR. The connection between splicing and cancer. J Cell Sci. 2006;119(Pt 13):2635-41. doi: 10.1242/jcs.03053. PubMed PMID: 16787944.

70. Menon R, Omenn GS. Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers. Cancer Res. 2010;70(9):3440-9. doi: 10.1158/0008-5472.CAN-09-2631. PubMed PMID: 20388783; PubMed Central PMCID: PMCPMC2866518.

71. Fackenthal JD, Godley LA. Aberrant RNA splicing and its functional consequences in cancer cells. Dis Model Mech. 2008;1(1):37-42. doi: 10.1242/dmm.000331. PubMed PMID: 19048051; PubMed Central PMCID: PMCPMC2561970.

72. Bourdon JC, Fernandes K, Murray-Zmijewski F, Liu G, Diot A, Xirodimas DP, et al. p53 isoforms can regulate p53 transcriptional activity. Genes Dev. 2005;19(18):2122-37. doi: 10.1101/gad.1339905. PubMed PMID: 16131611; PubMed Central PMCID: PMCPMC1221884.

73. Omenn GS, Menon R, Zhang Y. Innovations in proteomic profiling of cancers: alternative splice variants as a new class of cancer biomarker candidates and bridging of proteomics with structural biology. J Proteomics. 2013;90:28-37. doi: 10.1016/j.jprot.2013.04.007. PubMed PMID: 23603631; PubMed Central PMCID: PMCPMC3841011.

74. Menon R, Omenn GS. Identification of alternatively spliced transcripts using a proteomic informatics approach. Methods Mol Biol. 2011;696:319-26. doi: 10.1007/978-1-60761-987-1_20. PubMed PMID: 21063957; PubMed Central PMCID: PMCPMC4123872.

75. Li W, Kang S, Liu CC, Zhang S, Shi Y, Liu Y, et al. High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. Nucleic Acids Res. 2014;42(6):e39. doi: 10.1093/nar/gkt1362. PubMed PMID: 24369432; PubMed Central PMCID: PMCPMC3973446.

76. Luo T, Zhang W, Qiu S, Yang Y, Yi D, Wang G, et al., editors. Functional Annotation of Human Protein Coding Isoforms via Non-convex Multi-Instance Learning. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2017: ACM.

CHAPTER 2

A Catalogue of Alternatively Spliced Isoforms in Kidney[*]

2.1 Abstract

Background: Whole transcriptome studies are essential to understand the complexity of genetic regulation. However, short-read sequencing platforms cannot reliably differentiate between transcript isoforms. The Pacific Biosciences (PacBio) RS II platform with Iso-Seq protocol is capable of reading longer stretches of sequences, thus can sequence full-length transcripts and reliably distinguish between transcript isoforms of the same gene.

Methods: In this study, tumor nephrectomy samples performed at Michigan Medicine were sequenced on PacBio RSII by Iso-Seq protocol using SMRT technology. Full length transcripts are identified from glomerular and tubulointerstitial compartments separately. These transcripts are then validated using short reads from Illumina RNA-seq data from matching samples and samples from a different cohort. Then, validated transcripts are compared against the latest GENCODE transcript list and novel transcripts are identified.

Results: Through validation we identified 13536 and 13993 annotated transcripts, 15615 and 17268 potentially novel transcripts in glomerular and tubulointerstitial compartments respectively, among them novel transcripts belonging to several kidney related genes. As an example, we presented seven novel transcripts for NPHS2 gene.

[*] Chapter 2 is submitted for publication as **Ridvan Eksi**, Rajasree Menon, Bradley Godfrey, Christopher L. O'Connor, Markus Bitzer ,Matthias Kretzler, Yuanfang Guan. (2017) "A Catalogue of Alternatively Spliced Isoforms for Glomerular and Tubulointerstitial Compartments in Kidney Through Integrating of Long-reads and Short-reads" (*In review*)

Conclusion: PacBio Iso-seq sequencing with its advantage of full-length transcript reads can expand the repertoire of the expressed transcripts in the kidney.

2.2 Introduction

Alternative splicing (AS) process which happens during transformation of a pre-mRNA transcript to a mature transcript greatly increases the information content of genomes by producing multiple transcripts from a single gene. Previous studies on 'splicing code' have found that *cis*- and *trans*-acting factors have a role in regulating the AS process as well as the secondary structure of the pre-mRNA transcript. These factors together form the "splicing code" and determine the cell type-specific splicing [1]. It is estimated that 95% of multi-exonic human genes undergo AS, producing >100,000 distinct transcripts from ~20,000 protein-coding genes [2]. In addition to creating multiple forms of mRNA from a single gene, AS can influence gene expression by altering the mRNA stability and translation through nonsense-mediated decay and miRNA regulation [3].

High throughput short-read RNA sequencing has been a powerful tool for the study of gene expression levels and individual splice junctions [4, 5]. However, several studies have shown that reconstruction and quantification of transcript isoforms from short-read RNA sequencing is not sufficiently accurate [6, 7]. Recent progress in single-molecule long-read sequencing has provided powerful new tools to researchers to help solve this problem. Pacific Biosciences ('PacBio', http://www.pacificbiosciences.com/), has introduced the sequencing technique that is based on single-molecule sequencing chemistry with real-time detection (SMRT). It allows the sequencing of long sections of genomic DNAs or transcripts without fragmentation or PCR amplification. The sequencing length of the PacBio RS II platform can go up to 10 kb, which should be enough to cover size distribution of most transcripts in eukaryotes.

Thus, PacBio's full-length or nearly full-length transcripts eliminate the need of assembly for the downstream analysis. A limitation of the PacBio platform is its high sequence error rate. However, these errors show no context-specific bias, thus randomly distributed across the reads. Recently improved read-length and base-calling algorithms of the PacBio's SMRT analysis platform and the use of circular molecules have mitigated the high error rate problem. When read length exceeds the length of the cDNA template, each base pair is covered on both strands multiple times and these low-quality base calls are aggregated to derive high-quality, single-molecule Reads of Inserts (ROIs).

PacBio long-read transcriptome sequencing platform (Iso-Seq) has been successfully applied to human and other species, and it is shown that use of Iso-seq has a significant advantage over short-read RNA-Seq methods for identifying novel isoforms, detecting AS events and gene fusion events [8-11]. However, we did not come across a study where kidney transcriptome is exclusively studied with the Iso-Seq platform. Majority of kidney transcriptome studies use the generic transcriptome databases such as Ensembl, GENCODE or NCBI which are incomplete and may miss some of the important kidney specific transcripts. Moreover, when the complete set of annotated transcripts are used; some reads from expressed annotated transcripts may be assigned to very similar non-expressed annotated transcripts, resulting in miscalculated FPKM values, and less power in differential expression analysis. It is reported that primary non-inflammatory glomerular diseases that include minimal change disease, focal and segmental glomerulosclerosis and membranous nephropathy are rare diseases that cause serious morbidity and high mortality. They account for approximately 15% of prevalent ESRD cases (2008) at an annual cost in the USA of more than $3 billion [12, 13]. A major obstacle to a successful therapeutic intervention is our limited knowledge of disease mechanisms. Especially, the role of

transcript isoforms, including splice and length variants is not well known and not well studied. Hence, there is a need for a complete and reliable database of transcript isoforms that are robustly expressed in kidney.

In this work, the overall transcriptome of glomerular and tubulointerstitial compartments from a kidney is studied using both long reads from PacBio Iso-Seq platform and RNA-seq short reads from Illumina platform. We used short reads from Illumina to validate full-length transcripts found by PacBio, and generated two high quality sets of transcript isoforms that are expressed in glomerular and tubulointerstitial compartments. Then, we compared the confirmed transcript isoforms to the known set of known transcript isoforms in GENCODE and provided the final list of expressed transcript isoforms along with their annotation status to be used by researchers in downstream kidney transcriptome studies.

2.3 Methods

2.3.1 RNA Extraction

Human kidney cortex tissue was obtained from a nephrectomy. Tissue was immediately placed in RNALater at 4°C for 12-24 hours and stored at -20°C thereafter. Microdissection of glomerular and tubulointerstitial compartments were performed on 5 patient nephrectomy samples as previously described [14, 15]. For PacBio library, RNA was pooled in equal proportions for a total of 500nG for each compartment and processed for NGS library preparation.

2.3.2 RNA Library Preparation /Sequencing

For Illumina RNA-seq runs, RNA was assessed for quality using the TapeStation (Agilent, Santa Clara, CA) using manufacturer's recommended protocols. Samples with RINs (RNA Integrity Numbers) of 8 or greater were prepared using the Illumina TruSeq mRNA Sample Prep v2 kit (Catalog #s RS-122-2001, RS-122-2002) (Illumina, San Diego, CA) using manufacturer's recommended protocols. Where 0.1-3ug of total RNA was converted to mRNA using a poly(A) purification. The mRNA is then fragmented and copied into first strand cDNA using reverse transcriptase and random primers. The 3 prime ends of the cDNA are then adenylated and adapters are ligated. One of the adapters that are ligated has a 6 nucleotide barcode that will be unique for each sample which allowed us to sequence more than one sample in each lane of a HiSeq flow cell (Illumina). The products are purified and enriched by PCR to create the final cDNA library. Final libraries were checked for quality and quantity by TapeStation (Agilent) and qPCR using Kapa's library quantification kit for Illumina Sequencing platforms (catalog # KK4835) (Kapa Biosystems, Wilmington MA) using manufacturer's recommended protocols. They were clustered on the cBot (Illumina) and sequenced 4 samples per lane on a 50 cycle paired end for tumor nephrectomy samples, and 1 sample per lane on a 100 cycle paired end run for ERCB samples on a HiSeq 2000 (Illumina) in High Output mode using version 3 reagents according to manufacturer's recommended protocols.

PacBio sequencing library preparation was done according to the manufacturer's recommendation for Isoform Sequencing using the Clontech SMARTer PCR cDNA synthesis kit and BluePippin Size-Selection System. cDNA SMRTbell templates were fractionated into 1 kb – 2 kb, 2 kb – 3 kb, 3 kb – 6 kb, and 5 kb – 10 kb. Sixteen SMRT cells are used in total: two for each size fractions of both glomerular and tubulointerstitial compartments. Sequencing was

performed on a Pacific Biosciences PacBio RSII by University of Michigan DNA Sequencing Core.

2.3.3 Sequence Generation and Alignments

Pacific Bioscience SMRT raw reads were initially processed using the Pacific Biosciences' SMRT analysis software version 2.3.0. The polymerase reads were partitioned into sub reads. Read of Inserts (ROI) were generated using the default number of polymerase full passes. The Iso-Seq *classify* tool was then used to separate the ROIs into full length non-chimeric and non-full length reads. Full-length reads were defined as containing 5' and 3' cDNA primers and poly(A) tails. Then, the Iso-Seq *cluster* tool was used to cluster all the full-length reads derived from the same transcript to get the consensus full-length transcripts (CFLs). CFLs that are unpolished by Quiver are used in the rest of the analysis, because it has been reported that Quiver polishing sometimes obscures the introns [10].

SMRT CFLs were aligned and mapped with GMAP [16] release 2015-07-23 to the human genome(hg19 assembly). Reads mapping to a single location were kept (argument –n 1).

Illumina RNA-Seq reads were aligned and mapped using STAR [17] version 2.5 to the human genome (hg19 assembly). STAR was ran in 2-pass mode with suggested parameters under "ENCODE options" heading in the STAR manual.

As next phase of the analysis, pipeline for identification of transcription start sites, splice junctions and transcription end sites was adapted from the TRIMD pipeline by O'Grady et al. [10]. Single-exon CFLs are excluded from the following validation steps, as most of them are potential intronic fragments resulting from pre-processed mRNAs. Single-exon CFLs are added back to analysis before collapse step.

2.3.4 Identification of Transcription Start Sites (TSS)

CFL 5’ ends clusters were generated with CFL 5′ ends mapping within 8 bp of each other. Only CFL isoforms whose 5’ ends did not contain mismatches are used. A single TSS is determined for each of the clusters by calculating weighted (based on number of SMRT reads for each start coordinate) averages of the start coordinates of CFLs within the cluster. These consensus transcription start sites are considered validated if there is an annotated transcription start site within 10bp vicinity [10]. Annotated transcription start sites are extracted from GENCODE comprehensive annotation set (version 24).

2.3.5 Identification of Splice Junctions

Splice junctions from Iso-Seq CFLs were identified using GMAP, and splice junctions for Illumina reads were identified with STAR. Splice junctions from Iso-Seq CFLs are required to have at least 1 full-length reads spanning it to be identified. A splice junction from an Iso-Seq CFL is marked as validated if there are at least 3 short reads spanning it or if the junction is already annotated. Annotated junctions are extracted from GENCODE comprehensive annotation set (version 24).

2.3.6 Identification of Transcription End Sites (TES)

Iso-Seq CFL 3′ ends that align within 8 bp of each other on the genome are considered a single candidate transcription end site [10]. The CFL consensus transcription end sites were determined by calculating weighted averages of the end coordinates. Weights are determined by the number of PacBio consensus sequence reads ending at each coordinate. Only putative PacBio 3’ end sites that are supported by at least three SMRT reads are kept.

Illumina reads that has poly(A) tails were extracted from SAM alignment files. These putative reads with poly(A) tails are the reads that have FLAG code as being first-of-pair, and either end with a run of at least five As, at least two of which are mismatched on plus strand or

that start with a run of at least five Ts, at least two of which are mismatched on minus strand. The alignment position base next to mismatch location is considered a candidate transcription end site. TES that are within 8 bp of each other considered single candidate TES. The consensus TES coordinate was determined using weighted average of putative 3′ ends based on number of short reads supporting each TES coordinate.

Transcription end sites are marked as validated either if there is an Illumina TES on the same strand within four bases upstream or ten bases downstream or if there is an annotated TES in the 10bp vicinity [10].

2.3.7 CFL validation and comparison to known annotations

Overall flow of our study is shown in Figure 2-1. Transcription start sites, each splice junction and transcription end sites of each glomerular CFLs are compared to coordinates extracted from short-reads from ERCB glomerular samples and annotated transcripts. For tubulointerstitial compartment, transcript features are compared to short-read RNA-seq data from matching tumor nephrectomy samples and ERCB tubulointerstitial samples separately. Iso-Seq CFL validation was done through validating every splice junction present in the CFL. An Iso-Seq CFL is considered validated if every splice junction is validated based on the criteria explained above. Two different short-read RNA-seq dataset for tubulointerstitial compartment is combined for this step. Single-exon CFLs are added to the set of validated multi-exon CFLs, pending further investigation based on their relative location to a known transcript. The set of validated multi-exon CFLs and single-exon CFLs are collapsed with the collapse_isoforms_by_sam.py script in the tofu package provided by PacBio which is developmental version of the official Iso-Seq protocol. Then, collapsed isoforms are compared to

GENCODE comprehensive set of annotations (version 24) with cuffcompare tool from Tuxedo suite of tools. [18-21]

2.4 Results

2.4.1 Long-read Sequencing of Kidney Transcriptome

In this study, we examined human kidney cortex tissue obtained from patients undergoing nephrectomy to apply this method which requires larger amounts of RNA as would be available from kidney biopsy tissue samples. Library preparation and long-read sequencing were done using the PacBio's Iso-seq protocol, and initial data analyses are done using PacBio's SMRT Portal. Each of the glomerular and tubulointerstitial compartments had eight sequencing cells, generating 132240 and 125047 SMRT consensus "full-length" isoforms (CFLs) respectively. These CFLs are supported by 206415 and 232845 SMRT Reads of Inserts (ROIs) that were determined to be full length based on presence of 5' and 3' cDNA primer sequences, and poly(A) tail at 3' end. Glomerular CFLs ranged in length from 300 to 27890 bases, with a mean of 2446 bases and a median of 2319 bases. Tubulointerstitial CFLs ranged in length from 300 to 22765 bases, with a mean of 2862 bases and a median of 2613 bases. Size fractionation of library before sequencing reduced the bias toward shorter transcripts (Supplementary Figure 2-1). These CFLs are mapped to human genome (hg19 assembly) using GMAP [16]. Approximately 98% of the CFLs from both compartments are mapped uniquely to the genome.

2.4.2 Transcriptome Resolution through Integration of Pacific Biosciences SMRT Reads and Illumina Short Reads

Our goal is to identify glomerular and tubulointerstitial transcripts that are supported by reads from two different sequencing technologies. Even though PacBio platform produces long

enough reads to span the full length of most human transcripts, they have their limitations in attributing the final transcript structures. Most notable anomaly present in PacBio reads is the prevalence of CFLs with 5' ends that map further downstream of annotated start sites. These varying 5' ends locations suggest that these CFLs are likely to be truncated and do not represent full-length transcripts. This phenomenon was also observed for 3' ends of the CFLs, with lower frequency (Figure 2-2). CFLs with truncated 3' ends are less likely to be found, because full-length read detection looks for poly(A) tail signal in addition to the presence of 3' end primers, which is usually more accurate. We also found some CFLs with splicing at non-canonical splice junctions, which appears to be result of deletions introduced during library preparation or sequencing. We have used short-reads from Illumina platform to eliminate CFLs with splice junctions that are due to technical artifacts.

2.4.3 Transcription Start Site Validation in Kidney Transcriptome

Using the criteria explained in the Methods sections, we identified 42,896 and 38,831 putative transcription start sites from PacBio tubulointerstitial and glomerular multi-exon CFLs respectively. Since, TSS positions cannot be reliably inferred from Illumina short reads, our validation criteria relies only on annotated TSS. The fact that less than 10% of the putative TSS sites from PacBio CFLs are validated (Table 2-1) is due to very high proportion of PacBio CFLs with truncated 5' ends. These truncations are likely stemming from strand invasion during cDNA synthesis, and truncated 5' ends are more prevalent in longer transcripts, because reverse transcription process gets less efficient for longer transcripts. We showed this phenomenon on an example gene in Figure 2-2. As can be seen in section B of the Figure 2-2, gene AIF5 has multiple CFLs with varying start positions. During transcript collapsing by collapse_isoforms_by_sam.py script, the CFLs that are identical except their start positions are

merged together and start coordinate of CFL with longest 5’ end is taken as the true 5’ end for merged transcript. It should be noted that even the CFL with longest 5’end could still be truncated, and there is no way to distinguish a truncated CFL from a CFL with an alternative TSS using long or short reads. Thus, we decided not to eliminate any CFLs based on their TSS position, but readers should exercise caution with the transcript structures provided, as exact TSS position might be on the upstream of provided coordinates.

2.4.4 Splice Junction Validation in Kidney Transcriptome

Through mapping PacBio multi-exon CFLs to the genome, we identified 185,185 and 171,742 splice junctions in tubulointerstitial CFLs and glomerular CFLs respectively. Each of these splice junctions has at least one full-length SMRT read spanning them. Validated set of splice junctions is the union of annotated junctions, and splice junctions that have at least three short reads spanning them from Illumina RNA-seq data. For tubulointerstitial, by using the shallower tumor nephrectomy Illumina RNA-seq data, we were able to validate around 64% of all PacBio tubulointerstitial splice junctions. Using the deeper ERCB RNAseq data allowed us to validate additional 5719 splice junction, all of which are novel. For glomerular, around 65% of PacBio junctions are validated (Table 2-1).

2.4.5 Transcription End Site Verification in Kidney Transcriptome

A total of 24,625 and 26,260 putative transcription end sites were identified in PacBio tubulointerstitial and glomerular multi-exon CFLs through the process explained in Methods section. Then, short reads containing poly(A) reads were extracted from Illumina RNA-seq data, and poly(A) site coordinates are extracted from those reads. Putative transcription end sites from PacBio are considered validated if they are in the vicinity of either an annotated transcription end site, or a poly(A) site extracted from short reads. Even though it is not as prevalent as 5’ ends, we

have a high variation in the transcript end sites extracted from PacBio CFLs (Figure 2-2), and this variation is not fully captured by short-reads as we have limited number of reads with poly(A) tails. It has been reported that translational efficiency is regulated by the length of the 3' untranslated region, and transcripts with varying 3’ UTR lengths might be regulated differently [22]. In the scope of this study, we decided to provide CFLs present in the sample with their original 3’ end locations to give readers a choice to process CFLs with different 3’ UTR lengths based on their study objective, and we did not filter out CFLs based on their TES.

2.4.6 CFL Validation and Collapsing into Final Structures

Since exact TSS and TES coordinates do not agree very well with annotated coordinates because of the reasons explained earlier, these transcript features were not used in whole CFL validation. Our validation criteria require the CFL to have all of its junctions validated either by short read support or by annotation. With this criteria 53,540 tubulointerstitial multi-exon CFLs, and 47,785 glomerular multi-exon CFLs are marked as validated. But, some of these CFLs are redundant, because they differ only by their truncated 5’ ends, and they need to be collapsed. Collapse script by PacBio honors 3' ends very tightly. This is because poly(A) tail is used as 3' signal, and no truncations expected on 3’ end. Hence, any two isoforms that differ on the 3' end by more than 100 bp (a defined threshold by PacBio) is considered a different isoform. The collapse script honors 5' ends much less. This is because PacBio’s regular cDNA protocol kit does not do cap trap and that results in 5’ end with varying lengths. If two isoforms differ only by their 5' ends, meaning one isoforms has 0, 1, or more 5' exons than the other but all remaining exons agree, then the shorter isoform is considered identical to the longer one, and it is collapsed to the longer isoforms. After collapsing, 45,778 and 58,378 isoforms are formed from tubulointerstitial and glomerular tissue.

2.4.7 Comparison of Validated Isoforms to Annotated Transcripts

List of collapsed CFLs are compared to an annotated set of transcripts from GENCODE version 24. Comparison is done with cuffcompare tool in Cufflinks which classifies each input transcript into twelve distinct classes based on their overlap with an annotated transcript. The seven most prevalent classes and the numbers of collapsed transcripts from tubulointerstitial and glomerular compartments belonging to each of these seven classes are shown in Table 2-2. "Complete match of intron chain with an annotated transcript", and "Contained within a reference transcript" classes together comprise the set of expressed annotated transcripts in the sample. The set of transcripts that belong to "A transfrag falling entirely within a reference intron" class was discarded. These are single exon transcripts which are most likely by-products of intron decay process [8]. Another single exon transcript class is "Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron" class, which was also discarded. These are possible pre-mRNA fragments that are pulled down due to inefficient poly(A) selection step.

The class of "Potentially novel isoform" includes transcripts that share at least one junction with an annotated transcript, but have junctions that do not occur in any annotated transcript, or in the case that junctions occur in annotated transcript, they are present in a novel combination. Again, readers should be cautious about the exact location of TSS in these transcripts, as they might be truncated.

The class of "Intergenic transcripts" includes transcripts that map to an intergenic location. Most of these intergenic transcripts are single exon transcripts, more likely regulatory non-coding mRNAs.

The two sets of validated and collapsed isoform sequences with their corresponding annotation class are provided in the Supplementary Files.

2.4.8 Novel Isoforms of NPHS2 through New Combinations of Exon Skipping

NPHS2 (podocin) is a protein coding gene located on chromosome 1. It encodes a protein that has a role in the regulation of glomerular permeability and is enriched in glomerular compartment. Single point mutations in this gene have been shown to cause steroid-resistant nephrotic syndrome [23]. This gene has 8 exons, and has two protein coding splice variants according to GENCODE database. NCBI's RefSeq lists three other predicted protein coding splice variants. Our validated set of glomerular isoforms has 11 different splice variants for this gene (Figure 2-3). Two of these match exactly to the variants in GENCODE, and other two matches to the two of the predicted splice variants in RefSeq. Another four splice variants have the same first and last exon, but have a different combinations of exons making them novel. Remaining three novel splice variants have alternative end sites. Among the seven novel splice variants, there is one novel junction which is supported by multiple short reads.

2.4.9 Novel Intergenic Transcripts

Among the final set of isoforms, there are 4208 and 9501 intergenic transcripts from tubulointerstitial and glomerular compartments. Majority of these transcripts have single exon, and don't have any junctions to be validated. There are 76 and 55 multi-exon intergenic transcripts in tubulointerstitial and glomerular. All of the junctions in these transcripts are supported by multiple short reads, as they passed the validation criteria. One of those multi-exon intergenic glomerular transcripts is shown in Figure 2-4. They are located on chromosome 12, and consist of four exons. Two novel transcripts differ by the length of their last two exons. To gain further insights into the functional nature of these transcripts, we assessed their coding

potential with CPAT (Coding Potential Assessment Tool) [24]. Even though there are short open reading frames present in the transcripts, both of the transcripts were predicted to be non-coding by CPAT. According to UCSC Genome Browser, there are evidences of transcription at this locus as it shows multiple Human ESTs and Human mRNAs from GenBank.

2.4.10 Comparison of Expressed Set of Annotated Transcripts in Glomerular and Tubulointerstitial Compartments

Glomerular and tubulointerstitial compartments each are expected to have distinct transcriptome profiles. In this section, we compared the set of expressed annotated transcript from each compartment. As shown in Table 2-2 B, 13,993 and 13,536 annotated transcripts are expressed in these compartments, for which 8198 are common transcripts (Figure 2-5 A). We performed KEGG pathway enrichment on the transcripts that are uniquely expressed in each compartment. Full list of enriched pathways is on Supplementary Table 2-1. Figure 2-5 B shows top 3 enriched pathway for each compartment. Glomerular-only expressed transcripts are enriched for Non-alcoholic fatty liver disease (NAFLD). Many of the genes in NAFLD pathway has been shown to play important roles in kidney. Su et al. stated the role of IL6 and IL6R in several renal diseases, such as IgA nephropathy, lupus nephritis, diabetic nephropathy, acute kidney injury, and chronic kidney disease [25]. Another pathway highly enriched in glomerular transcripts is the RAP1 signaling pathway. It has been suggested that small changes in RAP1 signaling pathways critically affects podocytes [26].

2.5 Discussion

Integrating data from two different sequencing technologies allowed us to identify and validate nearly 14k known transcripts in tubulointerstitial and glomerular compartments of a kidney. In addition to that, we identified and validated nearly 12k and 8k multi-exon potential

novel transcripts from each respectively. Using short reads for validation of splice junctions allowed us to eliminate PacBio transcripts with erroneous junctions. On the other hand, none of the sequencing technology is able to reliably and accurately locate TSSs for transcripts. PacBio tend to produce transcript with 5'ends truncated. Cartolano et al. proposed the implementation of the TeloPrime Full Length cDNA Amplification kit to the Pacific Biosciences Iso-Seq technology in order to enrich for genuine full-length transcripts in the cDNA libraries [27]. In another study, deepCAGE (Cap Analysis of Gene Expression) data is used as a mean to correct for exact 5' end coordinate of transcripts with success [10]. In this study, we imputed the ambiguous start positions by assuming the 5′-most Iso-Seq finding as the true 5′ end which may still result in a shorter than actual transcript.

The functions of these newly discovered novel alternatively spliced transcripts are unknown, but it is likely that alternative splicing creates variants with distinct functions. Protein coding variants with new exon skipping events can translate into proteins with a loss or gain of functional domains, which would result in a protein with a different function. The novel single-exon intergenic transcripts which are likely non-coding may play a regulatory role. Future functional studies for these novel transcripts are crucial for assessing their functional importance.

2.6 Figures

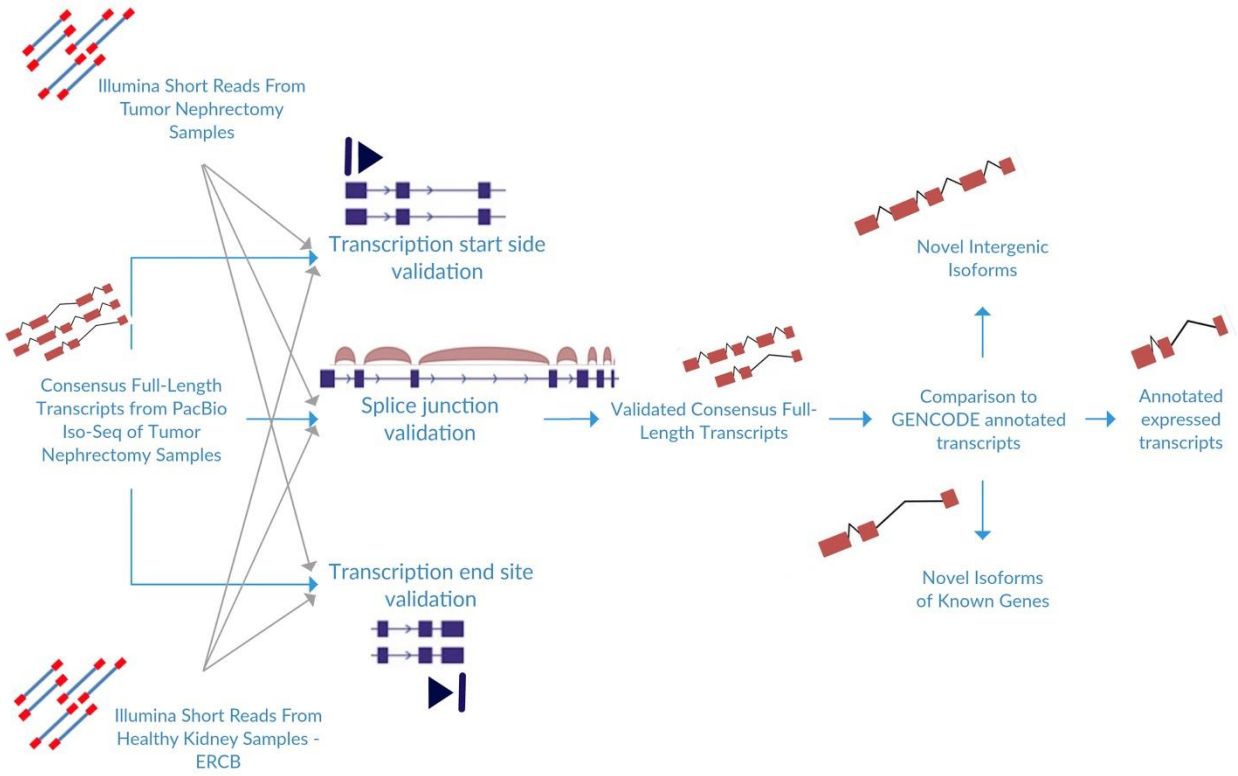


Figure 2-1 Overall flowchart of the study.

Three features of every multi-exon Consensus Full-Length transcripts found in PacBio reads are validated through short reads from two different RNA-seq datasets. Then, whole CFL is validated by validating every junction in the CFL. Validated CFLs are collapsed and compared to GENCODE annotation.

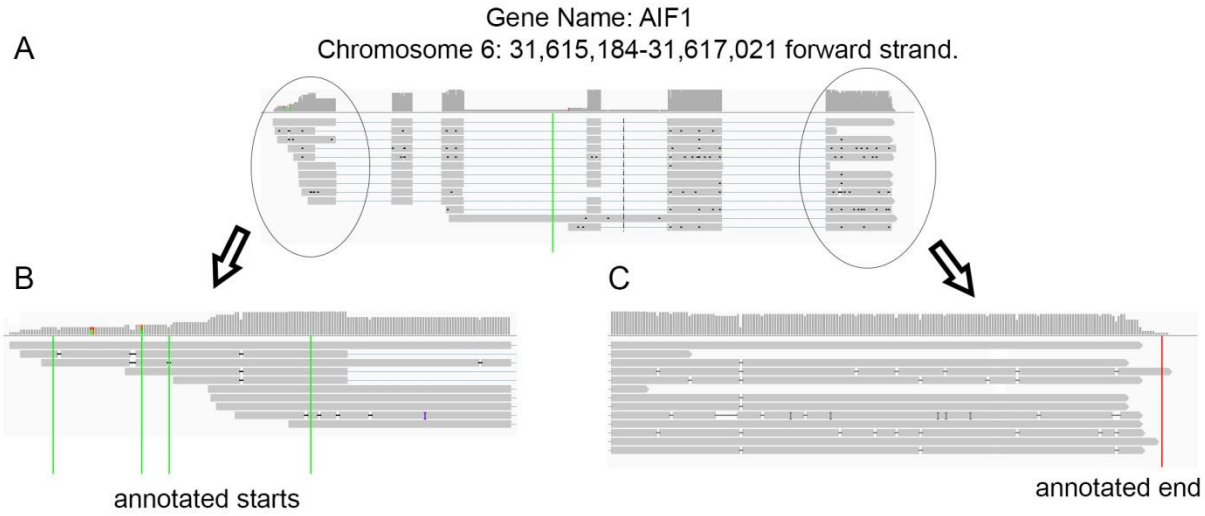


Figure 2-2 Variation of transcription start and end positions of CFLs.

(A) Thirteen multi-exon CFLs from AIF1 locus. **(B)** First exon in more detail. Vertical (green) lines show the location of annotated transcription start sites. Thirteen CFLs have a total of 13 different TSS, only 6 out of 13 TSS are within 10 bp of an annotated TSS **(C)** Last exon in more detail. Vertical (red) line shows the only annotated TES. Thirteen CFLs have a total of 7 different TES, only 4 out of 7 TES is within 10 bp annotated TES.

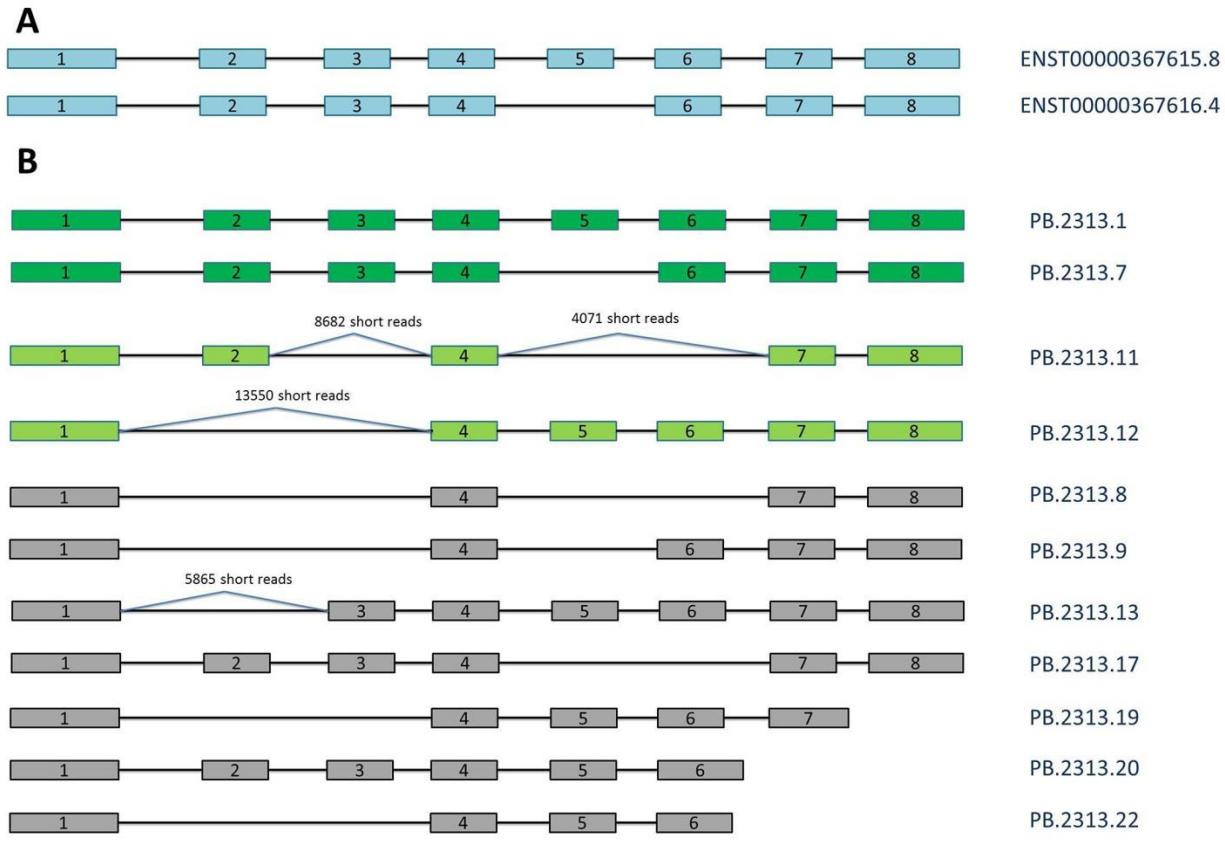


Figure 2-3 Gene model for NPHS2

Panel A shows the transcript structures of two annotated transcripts of NPHS2 in GENCODE. Second shorter annotated transcript is missing exon number 5. Panel B shows the transcripts found in PacBio glomerular sample and validated by our method. First two isoforms (dark green) match exactly to the annotated transcripts in GENCODE. Next two isoforms (light green) match to two predicted transcript variants in NCBI's RefSeq annotation (XM_017002298.1 and XM_005245483.3) Remaining seven isoforms (gray) isoforms are potential novel transcript variants of NPHS2. Numbers of uniquely mapped reads to each of the novel junctions are noted above junctions.

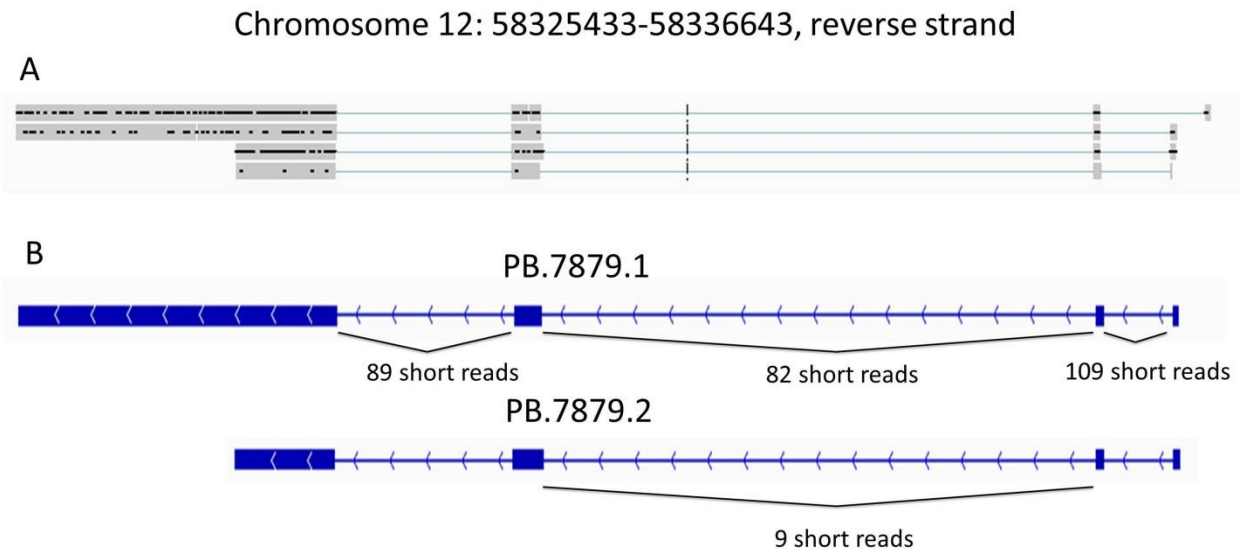
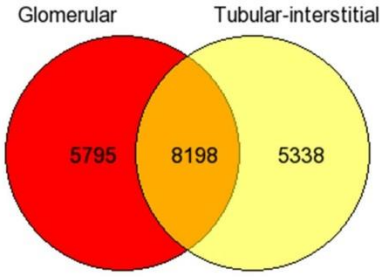


Figure 2-4 Two novel intergenic transcripts located on chromosome 12 that are expressed in glomerular compartment.

(A) Four CFLs that mapped to the locus. **(B)** Final transcript models that are collapsed from validated CFLs. First CFL is not validated because the rightmost junction doesn't have short-read support. The remaining three CFLs are validated and collapsed into two transcripts. They differ in their third and fourth exons. Number of short reads uniquely mapped to each unique junction is shown below the junction

A



B

	KEGG Pathway	p-value
Glomerular	Rap1 signaling pathway	2.9E-6
	Ribosome	8.3E-6
	Non-alcoholic fatty liver disease (NAFLD)	1.1E-4
Tubular-interstitial	Metabolic pathways	4.8E-7
	Ubiquitin mediated proteolysis	9.9E-5
	Aldosterone-regulated sodium reabsorption	2.7E-4

Figure 2-5 **(A)** Venn diagram of expressed annotated transcripts from glomerular and tubulointerstitial compartments. **(B)** Enriched KEGG pathways and corresponding p-values for the set of transcripts expressed only in glomerular and tubulointerstitial compartments.

2.7 Tables

Table 2-1 Validation of transcript features.

This table gives number of validated features for each of the transcript feature examined. First column is for validation of features of PacBio Tubulo multi-exon CFLs by short reads from Illumina TN data. Second column is for validation of features of PacBio Tubulo multi-exon CFLs by Illumina ERCB tubulo data. Third column is for validation of features of PacBio Glomerular multi-exon CFLs by Illumina ERCB glomerular data. Last column gives number of multi-exon CFLs for which every splice junction is validated. For Tubulo, two validated sets from each of Illumina datasets are merged.

* Transcript start sites are validated by annotation only. Short reads are used for splice junction and transcription end site validation.

	PacBio Tubular-interstitial with Illumina Tumor Nephrectomy (validated / total)			**PacBio Tubular-interstitial with Illumina ERCB Tubular-interstitial (validated / total)**			**PacBio Glomerular with Illumina ERCB Glomerular (validated / total)**		
Transcription start sites*	3534 / 42,896						3732 / 38,831		
	by short reads only (e.g. novel)	by annotation only	by short reads or annotation	by short reads only (e.g. novel)	by annotation only	by short reads or annotation	by short reads only (e.g. novel)	by annotation only	by short reads or annotation
Splice junctions	3105	7329	**119,932 / 185,185**	8824	516	**125,651 / 185,185**	5583	1194	**111,506 /171,742**
Transcription end sites	8	5278	**5318 / 24,625**	104	4883	**5425 / 24,625**	39	4850	**5071 / 26,260**
Multi-exon CFLs with all junctions validated	53,540 / 75,573						47,785 / 71,492		

Table 2-2 Comparison of validated-collapsed isoforms to GENCODE annotation.

cuffcompare tool is used to compare the set of validated isoforms to the set of annotated transcripts from GENCODE. cuffcompare classifies each match into 12 classes. In (A), we listed seven classes with the most number of isoforms. The first two rows together comprise the list of all annotated transcripts. The third and fourth rows combined comprise the list of potentially novel isoforms. These totals are shown in (B).

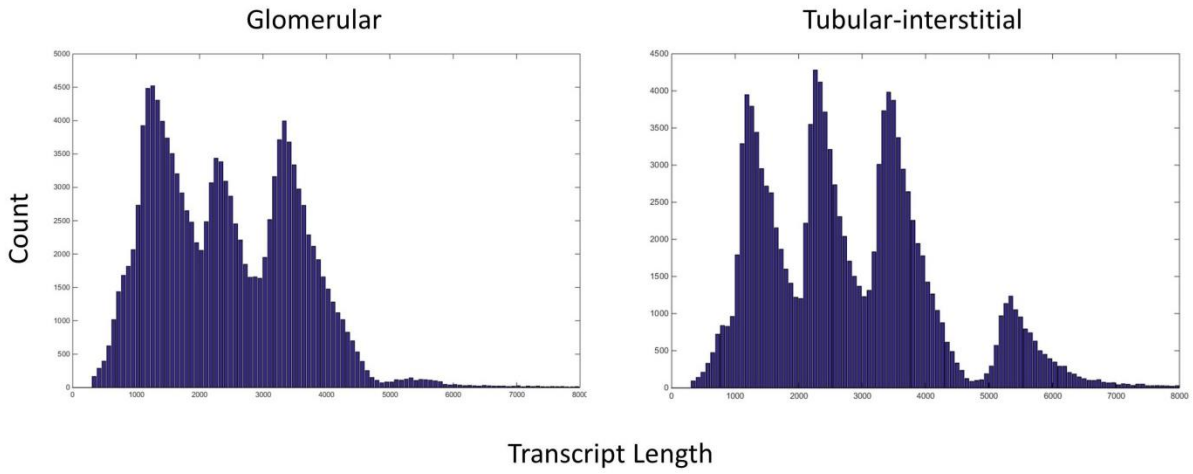
(A)

Type of Match	Validated Tubulo Isoforms	Validated Glomerular Isoforms
Complete match of intron chain with an annotated isoform	10407	10882
Contained within a reference isoform	3852	3857
Potentially novel isoform	11407	7767
Intergenic transcript	4208	9501
A transfrag falling entirely within a reference intron	11407	16627
Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron	3729	4956
Exonic overlap with reference on the opposite strand	2310	3420

(B)

Total annotated transcripts	13536	13993
Total potentially novel transcripts	15615	17268

2.8 Supplementary Files



Supplementary Figure 2-1 Consensus full-length transcripts length distribution

Supplementary Table 2-1 Full list of enriched pathways for glomerular-only and tubularinterstitial-only expressed genes

Glomerular-only expressed genes	
KEGG Pathway	**P-Value**
Rap1 signaling pathway	2.90E-06
Ribosome	8.30E-06
Non-alcoholic fatty liver disease (NAFLD)	1.10E-04
Thyroid hormone signaling pathway	1.40E-04
Transcriptional misregulation in cancer	1.90E-04
Neurotrophin signaling pathway	2.50E-04
Epstein-Barr virus infection	3.70E-04
TNF signaling pathway	5.10E-04
Wnt signaling pathway	5.50E-04
Pathways in cancer	6.80E-04
Adherens junction	8.80E-04
Platelet activation	9.10E-04
Bacterial invasion of epithelial cells	9.50E-04
Colorectal cancer	1.10E-03
Ubiquitin mediated proteolysis	1.70E-03
Osteoclast differentiation	2.00E-03
Alzheimer's disease	2.10E-03
Herpes simplex infection	2.10E-03
Proteoglycans in cancer	2.70E-03
Endocytosis	2.90E-03
HIF-1 signaling pathway	3.50E-03
Parkinson's disease	3.70E-03
Sphingolipid signaling pathway	3.70E-03
Spliceosome	5.10E-03
mTOR signaling pathway	5.30E-03
Vascular smooth muscle contraction	5.70E-03
Viral carcinogenesis	8.00E-03
Chemokine signaling pathway	8.30E-03
Leukocyte transendothelial migration	8.70E-03
Oxidative phosphorylation	8.70E-03
Viral myocarditis	9.30E-03
MAPK signaling pathway	1.10E-02

Tubular-interstitial-only expressed genes	
KEGG Pathway	**P-Value**
Metabolic pathways	4.80E-07
Ubiquitin mediated proteolysis	9.90E-05
Aldosterone-regulated sodium reabsorption	2.70E-04
GnRH signaling pathway	7.60E-04
Gastric acid secretion	7.90E-04
Axon guidance	1.10E-03
Sphingolipid signaling pathway	1.20E-03
Neurotrophin signaling pathway	1.20E-03
Renal cell carcinoma	1.60E-03
Oxytocin signaling pathway	1.90E-03
Insulin resistance	2.30E-03
Choline metabolism in cancer	2.60E-03
Central carbon metabolism in cancer	2.90E-03
Inflammatory mediator regulation of TRP channels	3.00E-03
Aldosterone synthesis and secretion	4.50E-03

Validated Transcript in Glomerular (in fasta format)

Validated Transcript in Tubulointerstitial (in fasta format)

2.9 Bibliography

1. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. Nature. 2010;465(7294):53-9. doi: 10.1038/nature09000.

2. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40(12):1413-5. doi: 10.1038/ng.259.

3. Reddy ASN, Marquez Y, Kalyna M, Barta A. Complexity of the alternative splicing landscape in plants. Plant Cell. 2013;25(10):3657-83. doi: 10.1105/tpc.113.117523.

4. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456(7221):470-6. doi: 10.1038/nature07509.

5. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63. doi: 10.1038/nrg2484.

6. Tilgner H, Raha D, Habegger L, Mohiuddin M, Gerstein M, Snyder M. Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. G3. 2013;3(3):387-97. doi: 10.1534/g3.112.004812.

7. Steijger T, Abril JF, Engström PG, Kokocinski F, Consortium R, Hubbard TJ, et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013;10(12):1177-84. doi: 10.1038/nmeth.2714.

8. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 2013;31(11):1009-14. doi: 10.1038/nbt.2705.

9. Weirather JL, Afshar PT, Clark TA, Tseng E, Powers LS, Underwood JG, et al. Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. Nucleic Acids Res. 2015;43(18):e116. doi: 10.1093/nar/gkv562.

10. O'Grady T, Wang X, Bentrup KHz, Baddoo M, Concha M, Flemington EK. Global transcript structure resolution of high gene density genomes through multi-platform data integration. Nucleic Acids Res. 2016;44(18):e145-e. doi: 10.1093/nar/gkw629.

11. Gonzalez-Ibeas D, Martinez-Garcia PJ, Famula RA, Delfino-Mix A, Stevens KA, Loopstra CA, et al. Assessing the Gene Content of the Megagenome: Sugar Pine (Pinus lambertiana). G3. 2016;6(12):3787-802. doi: 10.1534/g3.116.032805.

12. Maisonneuve P, Agodoa L, Gellert R, Stewart JH, Buccianti G, Lowenfels AB, et al. Distribution of primary renal diseases leading to end-stage renal failure in the United States, Europe, and Australia/New Zealand: results from an international comparative study. Am J Kidney Dis. 2000;35(1):157-65. doi: 10.1016/S0272-6386(00)70316-7.

13. Collins AJ, Foley RN, Chavers B, Gilbertson D, Herzog C, Johansen K, et al. 'United States Renal Data System 2011 Annual Data Report: Atlas of chronic kidney disease & end-stage renal disease in the United States. Am J Kidney Dis. 2012;59(1 Suppl 1):A7, e1-420. doi: 10.1053/j.ajkd.2011.11.015.

14. Ju W, Eichinger F, Bitzer M, Oh J, McWeeney S, Berthier CC, et al. Renal gene and protein expression signatures for prediction of kidney disease progression. Am J Pathol. 2009;174(6):2073-85. doi: 10.2353/ajpath.2009.080888.

15. Kato M, Wang M, Chen Z, Bhatt K, Oh HJ, Lanting L, et al. An endoplasmic reticulum stress-regulated lncRNA hosting a microRNA megacluster induces early features of diabetic nephropathy. Nat Commun. 2016;7:12864. doi: 10.1038/ncomms12864.

16. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21(9):1859-75. doi: 10.1093/bioinformatics/bti310.

17. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. doi: 10.1093/bioinformatics/bts635.

18. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511-5. doi: 10.1038/nbt.1621.

19. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011;12(3):R22. doi: 10.1186/gb-2011-12-3-r22.

20. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011;27(17):2325-9. doi: 10.1093/bioinformatics/btr355.

21. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31(1):46-53. doi: 10.1038/nbt.2450.

22. Tanguay RL, Gallie DR. Translational efficiency is regulated by the length of the 3' untranslated region. Mol Cell Biol. 1996;16(1):146-56.

23. Boute N, Gribouval O, Roselli S, Benessy F, Lee H, Fuchshuber A, et al. NPHS2, encoding the glomerular protein podocin, is mutated in autosomal recessive steroid-resistant nephrotic syndrome. Nat Genet. 2000;24(4):349-54. doi: 10.1038/74166.

24. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013;41(6):e74. doi: 10.1093/nar/gkt006.

25. Su H, Lei C-T, Zhang C. Interleukin-6 Signaling Pathway and Its Role in Kidney Disease: An Update. Front Immunol. 2017;8:405. doi: 10.3389/fimmu.2017.00405.

26. Potla U, Ni J, Vadaparampil J, Yang G, Leventhal JS, Campbell KN, et al. Podocyte-specific RAP1GAP expression contributes to focal segmental glomerulosclerosis–associated glomerular injury. J Clin Invest. 2014;124(4):1757-69. doi: 10.1172/jci67846.

27. Cartolano M, Huettel B, Hartwig B, Reinhardt R, Schneeberger K. cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing. PLoS One. 2016;11(6):e0157779. doi: 10.1371/journal.pone.0157779.

CHAPTER 3

Systematically differentiating functions for alternatively spliced isoforms in mouse through integrating RNA-seq data[†]

3.1 Abstract

Integrating large-scale functional genomic data has significantly accelerated our understanding of gene functions. However, no algorithm has been developed to differentiate functions for isoforms of the same gene using high-throughput genomic data. This is because standard supervised learning requires 'ground-truth' functional annotations, which are lacking at the isoform level. To address this challenge, we developed a generic framework that interrogates public RNA-seq data at the transcript level to differentiate functions for alternatively spliced isoforms. For a specific function, our algorithm identifies the 'responsible' isoform(s) of a gene, the isoform that carries out the function that the gene is annotated to, and generates classifying models at the isoform level instead of at the gene level. Through cross-validation, we demonstrated that our algorithm is effective in assigning functions to genes, especially the ones with multiple isoforms, and robust to gene expression levels and removal of homologous gene pairs. We identified genes in the mouse whose isoforms are predicted to have disparate functionalities and experimentally validated the 'responsible' isoforms using data from mammary tissue. With protein structure modeling and experimental evidence, we further

[†] Chapter 3 is published as **Ridvan Eksi**, Hong-Dong Li, Rajasree Menon, Yuchen Wen, Gilbert S. Omenn, Matthias Kretzler, and Yuanfang Guan. "Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data." PLoS Computational Biology 9, no. 11 (2013): e1003314.

validated the predicted isoform functional differences for the genes *Cdkn2a* and *Anxa6*. Our generic framework is the first to predict and differentiate functions for alternatively spliced isoforms, instead of genes, using genomic data. It is extendable to any base machine learner and other species with alternatively spliced isoforms, and shifts the current gene-centered function prediction to isoform-level predictions.

3.2 Author Summary

In mammalian genomes, a single gene can be alternatively spliced into multiple isoforms which greatly increase the functional diversity of the genome. In the human, more than 95% of multi-exon genes undergo alternative splicing. It is hard to computationally differentiate the functions for the splice isoforms of the same gene, because they are almost always annotated with the same functions and share similar sequences. In this paper, we developed a generic framework to identify the 'responsible' isoform(s) for each function that the gene carries out, and therefore predict functional assignment on the isoform level instead of on the gene level. Within this generic framework, we implemented and evaluated several related algorithms for isoform function prediction. We tested these algorithms through both computational evaluation and experimental validation of the predicted 'responsible' isoform(s) and the predicted disparate functions of the isoforms of *Cdkn2a* and of *Anxa6*. Our algorithm represents the first effort to predict and differentiate isoforms through large-scale genomic data integration.

3.3 Introduction

Determining the functions of proteins is a central goal of genetics, fundamental for understanding the molecular basis of diverse genetic diseases [1-7]. During the past few decades, significant efforts have been made to integrate and develop diverse machine learning algorithms for gene function prediction through large-scale genomic data integration [8-12], such as Support

Vector Machines, Bayesian classifications and Artificial Neural Networks. Despite differences in implementation and performance of the specific algorithms, the essence of these methods for gene function prediction is 'supervised learning', in which a model of features derived from functional genomic data (such as microarray, protein-protein physical interactions, gene-gene genetic interactions) is constructed to delineate a defined set of 'positives' (genes annotated to the function under consideration) and 'negatives' (genes without the function). These algorithms have significantly accelerated our understanding of gene functions.

A major intellectual limitation of the current function prediction paradigm is that it considers a gene as a single entity without differentiating the functional diversity of alternatively spliced isoforms. Alternative splicing is a major source of protein molecular function diversity and regulatory diversity. In humans, 95% of multi-exon genes undergo alternative splicing, generating proteins of potential different functions [13-21]. For example, TRPM3, which encodes a type of cation-selective channels in human, can be alternatively spliced into two variants targeting different ions [22-24]. Other splice variants have been reported with distinctly opposite functions. For example, the splice variants of *BCLX* are anti-apoptotic and pro-apoptotic, respectively [25]. Similarly, *CASP3*-L variant is pro-apoptotic while *CASP3*-S is anti-apoptotic [26]. Differences in function are sometimes reflected on the regulatory level: two alternatively spliced transcripts of *OSR2* have opposite transcriptional activities, activation and repression [26]. Attempts to capture such differential functions are currently limited to low-throughput experimental approaches and protein domain analysis. However, based on the protein domain annotation we downloaded in Dec, 2012, only 34% of the isoform pairs (of the same gene) in the NCBI database have different domains. The majority of the isoforms of the same gene differ in subtle ways which are not reflected by protein domains. We expect that

computational methods developed to differentiate isoform functions through integrating functional genomic data will assist a deeper, high-resolution understanding of gene functions.

The key challenge facing isoform prediction is the lack of a systematic catalog of isoform-level function annotations and large-scale genomic data resolved at the isoform level. The latter is resolved by the unprecedented amount of transcriptomic data generated by next-generation sequencing [27-29]. RNA-seq data available in public databases [30] now surpass the number of microarray data that have been previously used to infer functions at gene levels [31]. Algorithms have been developed to assign isoform-level expression values [27, 28, 32-38]. They provide a resource for isoform-level features that can be used as input data to infer isoform functions.

However, the fundamental challenge remains: a genome-wide set of 'ground-truth' annotations of functions at the isoform level is still lacking. Isoform functions have been computationally inferred through domains, binding regions and individual binding sites [39-45]. In widely used databases such as Gene Ontology [46, 47] and KEGG [48], biological functions are defined at the gene level. Under these circumstances, to predict whether a gene is related to a specific function, supervised learning algorithms will be deployed to derive a model from the genomic data we collect. For example, to predict 'mitochondria biogenesis'-related genes, we need a 'gold standard' set of genes known to be related to mitochondria biogenesis, and find out how these genes differ from other genes in their expression patterns. Without an existing and comprehensive set of annotated isoforms, standard 'supervised learning' approaches are not applicable to predicting different functions for splice isoforms.

We developed a generic framework that improves gene function prediction and provides information for isoform-level functions. Our framework intends to locate the 'responsible' set of

isoform(s) of annotated genes of a specific function, through iteratively correcting the composition of this set to maximize its 'discriminativity' against the negatives. For example, for mitochondria biogenesis, we have a set of genes G^+_1, G^+_2, …, G^+_n annotated to this function (positive genes). Each gene is considered as a bag of alternatively spliced isoforms. We then try to find out which isoform(s) of this positive set of genes can be selected to maximize the difference between them and the negative isoforms, *i.e.*, the $G^{i}_{k}{}^{+}$, where k is the gene index, and i is the isoform index. The isoforms in the $G^{i}_{k}{}^{+}$ set are iteratively updated to maximize the similarity within them. The model derived from the $G^{i}_{k}{}^{+}$ set (instead of G_k^+) is then used to classify whether an isoform has the biological function, mitochondria biogenesis, in this example. From this perspective, our problem is a multiple instance learning task [49-52], in which each example considered positive with respect to certain property contains multiple discrete elements, of which at least one of them must be positive.

We used an iterative algorithm to approximate the solution to the above optimization problem. We found that our algorithm can successfully capture the differential isoform functions as evidenced by prediction accuracy on multi-isoform genes as well as literature and experimental validation on isoforms that are drastically different in their assigned functions. Computationally predicting isoform functions or differentiating functions for isoforms of the same genes in a genome-wide manner using high-throughput genomic data has not been done prior to our work. This framework is also generic. It can be integrated into any basic machine learning algorithm such as logistic regression, random forests or deep learning, and can be readily extended to predict isoform functions in other organisms as well as other properties of isoforms such as phenotypes and interaction networks. Our prediction results are available to the user community online at http://guanlab.ccmb.med.umich.edu/isoPred/.

3.4 Results

We first established the workflow that generates isoform-level function predictions. Then we validated our models with computational cross validation, focusing on performance comparison between multiple-isoform genes and single-isoform genes. The predicted 'functional' isoform(s) are further validated using breast cell proteomic data. Finally, through protein structure modeling and experimental evidence, we validated our predictions for *Cdkn2a* and for *Anxa6*, whose isoforms were predicted to be responsible for different functions.

3.4.1 The isoform function prediction framework

We aim at predicting isoform functions while functional annotations have been based on genes. The model therefore needs to be learned at the isoform level rather than at the gene level (see Supplementary Figure 3-1 for a comparison between a traditional gene function learning problem and our isoform function prediction problem). Our core idea is to model the common patterns of a **subset** of isoforms across genes associated with a particular function, with the requirement that at least one of the isoforms of each positive gene must retain the common feature pattern. This is a multiple instance problem, the aim of which is to identify the hidden labels of the isoforms of the positively annotated genes and use these hidden labels to construct classification models to label additional isoforms.

We used maximum margin-based classification as base-learner in this iterative process [53], due to the success of SVM in the protein function prediction domain [9]. Such a framework could be integrated into any other machine learner, such as deep learning or Bayesian classification.

To elucidate our algorithm, we will use the following definitions throughout the paper:

- A '**bag**' refers to a gene, which consists of multiple isoforms.
- **'Instances'** refer to individual isoforms.
- A **'positive'** bag refers to a gene related to the specific function under study.
- **'Witness(es)'** refer to the isoform(s) of a positive gene related to the specific function under study. A 'positive' bag has at least one witness;

Our algorithm aims to identify a subset of isoforms of the positive genes that maximizes the difference between them and the negative isoforms. Identifying the best combination of isoforms from the positive genes is difficult. The ideal solution requires excessive computational time. We therefore approximated the solution with an iterative algorithm (Figure 3-1). For the training set, in the first iteration, we assign every isoform of a positive gene to be positive. We then establish a classifier, with which we can go back to classify the training set. This classifier will assign some isoforms of the positive gene to be positive (the 'witnesses'), and some to be negative, but at least one isoform of a positive gene must be positive. This subset of 'witnesses' is iteratively updated to maximize the inter-class distance. Using this assignment, a new classifier is established, which could again be used to classify the training set. We iterate this process to update the 'witnesses'. This procedure is repeated until convergence is achieved (Figure 3-1).

Because of the properties and relationships between genes and isoforms, this Multiple Instance Learning (MIL) framework is suitable for our problem. MIL assumes that there are one or more positive instances in a positive bag, and, if we can identify these positive instances in positive bags, we can expect to have a better classifier by excluding remaining instances from the positive class. Biologically, if a gene is annotated to a function, one or more of its isoforms will be selected as witness(es) as the ones that are related to this function and other isoforms will be

marked as non-functional. By excluding these non-functional isoforms, the classifier is expected to be more accurate.

3.4.2 Implementation and testing parameters

To generate alternatively spliced transcript-level features, we collected transcriptomic data from public RNA-seq experiments (Dataset S1). These RNA-seq datasets cover a wide sampling of tissues and different experimental conditions, such as liver, brain, muscle and testis. The ENCODE RNA-seq data covering ovary, mammary gland, stomach, kidney, liver, lung, spleen, colon and heart, are also included in our data. Co-expression patterns in these data can be highly informative for co-functionality. We assigned isoform-level expression values for each of these experiments using the state-of-the-art tools [32, 34, 35] (Figure 3-1). Public RNA-seq datasets came from different experimental protocols and such information is not always recorded in databases in a standardized way. We filtered these datasets based on their quality and coverage (see Methods). This data collection serves as our genomic data input. Essentially, each isoform was described by a vector of values specifying its normalized expression value in various RNA-seq samples. Gene Ontology is arranged in a hierarchy, where complex relationships exist between different terms. To test different variations of MIL and tune the parameters of our algorithms, we focused on a list of biological process terms that have been voted by biologists to be able to describe and cover different biological processes that are experimentally testable [54]. This list included 99 terms, with GO term size 20–300.

Two important parameters (other than the standard SVM parameters) in this algorithm are the proportion of positive isoforms to be labeled as positive in each iteration, and whether we should label the rest of isoforms in the positive bag negative or discard them. Therefore, we tested two basic formulations of the algorithm. The first approach tries to impute all non-witness

instances in positive bags as negative instances and then considers the problem as a supervised learning problem. The second approach tries to identify a single witness from each positive bag which is responsible for the positive label. Then, a classifier is built based on these witnesses only, while other instances are dropped. SVM formulations of these two approaches are, respectively, mi-SVM and MI-SVM [50]. For the first approach, we envisioned that different ratio of instances can be retained as 'witnesses' and tested three different cutoffs (Figure 3-2 A-B).

Ideally, testing of isoform function prediction should use isoform-level gold-standard functional annotation. However, such comprehensive functional annotation does not exist in any database (if they did exist, regular supervised classification would be sufficient to predict isoform functions). Therefore, we first evaluated the performance of our algorithms at the gene level. The probability of each gene to be associated to a function is assigned with the maximum value of all its instances, under the assumption that the eventual gene function is carried out by at least one of its isoforms. For all methods and parameters tested, the algorithm converged within several iterations. Additionally, we found that different thresholds or methods resulted in relatively stable performance on the gene level. mi-SVM with 75% of all negative scores as the cutoff for defining 'witnesses' resulted in the highest AUC of 0.73 (Figure 3-2 C). Therefore, we used this method for inferring isoform functions and all the evaluation and validation below is based on this threshold.

3.4.3 Cross-validation of the function prediction algorithm.

We adopted several lines of validation to test our algorithm, including computational validation of multi- and single-isoform genes, literature evidence for top predicted candidates and experimental validation of top predictions. For all the following evaluations, we presented

1792 biological process terms with 20 to 300 genes annotated to each. This size range was selected based on previous statistical studies showing that GO terms of this size show robust cross-validation behavior [55].

We first compared the performance of our algorithm in capturing gene-level functions to a direct SVM model using the same data input at the gene level. For each GO term, we carried out five-fold cross-validation to evaluate our prediction results. We partitioned the training and test groups by genes instead of isoforms to prevent information leak in the evaluation process. We used both the area under the ROC curve (AUC) and the area under the precision-recall curve (AUPRC) to measure predictive performance (complete evaluation results are included in Dataset S2). Because GO terms of different sizes vary in terms of predictability [9], we divided all GO terms into 5 groups so that each group contains roughly the same number of GO terms according to GO term size. This resulted in groups of [20, 27], [27, 39], [39, 60], [60, 105] and [105, 298]. For each GO term group, we calculated AUC, AUPRC, precision at 1% recall and precision at 10% recall (Figure 3-3). The median AUCs of the 5 GO term groups are 0.66, 0.67, 0.68, 0.69, and 0.71, respectively. Strictly, these AUC values cannot be directly compared to those reported results in the literature because of the difference in the data used. However, it is still meaningful to benchmark, at least roughly, our results against the reported results. The work of Peña-Castillo *et al.* [56] is a benchmark of the mouse gene function prediction performance in 2008, using heterogeneous genomic data including physical interaction, protein domain, phenotypes and expression. Peña-Castillo *et al.* reported a median AUC in predicting novel gene annotations across 72 GO biological function terms is 0.69 with a standard deviation 0.071. Thus, our algorithm achieved satisfactory results using transcriptomic data alone; the additional benefit of differentiating isoform functions will be detailed in the following sections.

For some of the biological processes, interpreting data at the isoform level can dramatically improve the prediction performance over gene level results. These biological processes are those likely to be affected or carried out by certain specific isoforms of the genes. For example, for GO terms GO:0019882 (antigen processing and presentation), GO:0019395 (fatty acid oxidation), GO:0031032 (actomyosin structure organization), GO:0046649 (lymphocyte activation), GO:0002252 (immune effector process), GO:0045058 (T cell selection) (Supplementary Figure 3-2), AUPRC increased from 0.087 to 0.105 (20% improvement, baseline 0.0018), from 0.026 to 0.036 (36% improvement, baseline 0.0021), from 0.047 to 0.060 (28% improvement, baseline 0.0012), from 0.0414 to 0.0593 (43% improvement, baseline 0.0101), from 0.037 to 0.046 (27% improvement, baseline is 0.0061), from 0.017 to 0.039 (137% improvement, baseline is 0.0011) respectively using our iterative algorithm compared to using the gene-level SVM method only. The improvement in performance is likely due to having access to more data and eliminating noisy, non-predictive patterns from positive class which is achieved by the MIL formulation.

3.4.4 Better performance for multi-isoform genes than single-isoform genes

The performance obtained in the previous section is a mix between single-isoform genes and multiple-isoform genes. We hypothesize that, if our framework does bring in discriminative power at the isoform level, multi-isoform genes should be predicted with better accuracy than single-isoform genes for the same set of GO terms evaluated. Single-isoform genes include both real ones and those that are missed in the database [57]. In fact, although it has been estimated that 95% of the multi-exon genes in human have multiple isoforms [13], only 13% of the genes are documented with validated multiple isoforms in NCBI. For these genes, the performance is expected to be poorer since the input features for each gene approximate the average of all its

isoforms and our algorithm is not applicable to these genes in differentiating isoform functions. Therefore, we separately evaluated the performance of single-isoform genes and multiple-isoform genes. Two-fold cross-validation was carried out to evaluate our models to ensure there are sufficient genes in both the training and test sets for multi- and single-isoform genes. To make the comparison feasible, negatives in each group were randomly chosen to ensure that the ratio of positives to negatives is the same for multi- and single-isoform genes for the same GO term. In doing so, the AUC, AUPRC, precision at 1% recall and precision at 10% recall of each GO term are recalculated for both sets separately. These results are again organized into 5 groups, based on the number of positive genes in the test set for each GO term (Figure 3-4).

We found consistently better performance for multi-isoform genes than single-isoform genes for the same GO term evaluated (Figure 3-4; complete evaluation results are included in Dataset S3). For multi-isoform genes, we acquired median AUCs of 0.68, 0.70, 0.70, 0.73 and 0.76 for the five groups of multi-isoform genes, compared to median AUCs of the same groups of GO terms for single-isoform genes, 0.62, 0.63, 0.65, 0.66 and 0.68, which correspond to 57%, 56%, 34%, 47% and 40% improvements against the baseline (0.5), respectively. Similar better performance is seen for AUPRC values. We observed an increase from 0.002 to 0.006 (162% improvement, baseline is 0.0009), from 0.003 to 0.009 (168% improvement, baseline is 0.0015), from 0.005 to 0.012 (152% improvement, baseline is 0.0021), from 0.008 to 0.022 (156% improvement, baseline is 0.0034), from 0.016 to 0.042 (154% improvement, baseline is 0.0072). In fact, 71% of the 1591 GO terms with more than 20 positive genes and less than 300 positive genes have gained better performance for multi-isoform genes. This implies that our approach is effective in capturing the functionality of multi-isoform genes, which is robust regardless of the metrics used to quantify performance. For example, for GO terms GO:0000279 (M phase),

GO:0006952 (defense response), GO:0006936 (muscle contraction), GO:0007507 (heart development), GO:0002252 (immune effector process), GO:0003002 (regionalization), precision at 10% recall increased from 0.121 to 0.571 (baseline 0.0069), from 0.051 to 0.500 (baseline is 0.0103), from 0.052 to 0.364 (baseline is 0.0038), from 0.045 to 0.389 (baseline is 0.0097), from 0.042 to 0.750 (baseline is 0.0048), from 0.074 to 0.320 (baseline is 0.0089), respectively (Supplementary Figure 3-3). Such precision improvement is robust to accuracy measurement across the entire precision-recall spectrum and consistent across a wide sampling of GO terms. For the five groups of GO terms, median of precision at 1% recall increased from 0.005 to 0.014 (baseline is 0.0009), from 0.008 to 0.031 (baseline is 0.0015), from 0.011 to 0.045 (baseline is 0.0021), from 0.016 to 0.083 (baseline is 0.0034), from 0.028 to 0.154 (baseline is 0.0072) (Figure 3-5). This result indicates that our algorithm is more effective in predicting functions for genes with multiple isoforms than single-isoform genes, which is likely caused by the power of our algorithm in differentiating isoform functions for the same genes.

3.4.5 Robustness of predictions with respect to gene expression levels and exclusion of homolog gene pairs

We further considered two factors to validate the robustness of our algorithm. First, because the estimated expression levels are less reliable for genes that have low expression values, we tested whether our algorithm can predict the functions for low-expressing isoforms. We partitioned genes into three groups of equal numbers - the high, the medium and the low groups - based on their expression level averaged across all experimental conditions and evaluated prediction results separately for these groups to see if performance is affected by the overall expression level of genes. We showed that prediction performance is fairly robust to the genes' expression levels (Figure 3-6 **A**). Although the genes in the highest expression group do

show better performance (AUC=0.79), AUCs for the medium (0.69) and low (0.69) groups are comparable to the global accuracy (Figure 3-6 A), indicating that our method is applicable to genes that have relatively low expression levels.

Secondly, we tested whether our good performance comes from the homolog gene pairs of which one member is grouped into the training set and the other member is grouped into the test set. This can potentially cause overfitting by information leak, which is a common phenomenon in all functional prediction algorithms, especially those based on sequence information. In order to test robustness of our method with respect to having homolog pairs in test and training sets, we partitioned all genes into equal size test and training sets by putting all genes in a homolog group as defined in Ensembl [58] together and re-evaluated the performance. We found that the performance is comparable to the case where paralog genes can occur in training and test sets (Figure 3-6 B). This suggests that our method utilizing RNA-seq data integration is also robust to the information leak from homologs.

3.4.6 Validation of predicted 'functional' isoform(s)

To generate the final isoform function prediction, we applied bootstrap bagging to assign scores for each isoform (Figure 3-1, http://guanlab.ccmb.med.umich.edu/isoPred/ for complete prediction results). The median AUC for the bootstrap result across all GO terms is the same as the cross-validation result, indicating the robust performance of bootstrapping. Essentially, genes are sampled with replacement to construct a training set, and the rest of the examples form an "out-of-bag" set. This process is iterated, and the eventual predictions are drawn from the median across all "out-of-bag" sets. Bootstrap bagging is suitable for our isoform function prediction task, where the numbers of positive and negative examples are highly imbalanced [55]. The robustness of bootstrap bagging has been tested for positive example set sizes ranging from less

than 20 to more than 200 [55, 59], which is close to the GO term sizes for which we provide predictions in this study.

Because each GO term has a different background probability for a gene to be associated with it, we calculated the fold change against background for each gene to be associated to a GO term. Indeed, it is pervasive that isoforms of the same gene are assigned with different confidence levels for the same function under consideration (see Supplementary Table 3-1 for a selection of examples and http://guanlab.ccmb.med.umich.edu/isoPred/ for complete prediction results). Because biological functions of transcripts are eventually delivered at the protein level, we hypothesized that the functional isoform(s) of a gene must be expressed at the protein level in the normal physiological condition. We therefore used splice variant protein expression data in normal mammary tissue to validate our predictions. Using our previously developed protocol [60], we identified genes of which only one isoform is strongly expressed (but the other isoforms are not detected). In this gene list, we focused on the ones that have a known specific function (*i.e.*, the function has less than 300 genes annotated to them), but their isoforms are predicted with drastically different confidence levels to carry out this function. In total this resulted in 15 isoform groups. We could then check whether the predicted 'responsible' isoform(s) correspond to the expressed isoforms in normal breast tissues.

We found a strong match between the expressed splice variant and the isoform(s) predicted to be responsible for the known specific function of the gene (Table 3-1 and Dataset S4). All the expressed isoforms are predicted with the highest value in at least one of the known functions, indicating the consistency between the predicted functional isoforms and the expressed isoform. For example, we predicted that NM_172745.3 of Tufm is responsible for its function translation (34 fold over background probability), while the other isoform,

NM_001163713.1, is much less likely to be functional (3 fold over background probability). Indeed, only NM_172745.3 is expressed in normal breast tissue. Among the six alternatively spliced isoforms of Tardbp, we correctly predicted that NM_145556.4 is the one responsible for its function in RNA splicing and stabilization; this variant is the only one identified in the proteomic sample. In fact, of the 9 functions annotated for Tardbp, NM_145556.4 is predicted with the highest value eight times. The majority of the exceptions occur for tissue or developmental-stage-specific GO terms, such as GO:0007507 heart development, GO:0035051 cardiac cell differentiation and GO: 0006936 muscle contraction. The predicted functional isoform is not the one identified in our proteomic sample, most likely because our sample is tissue-specific and normal and these functions might be carried out by other isoforms in other tissues or conditions. However, overall, the predicted functional isoforms are consistent with the ones we identified in our proteomic data, indicating that our algorithm can correctly identify the functional splice variants in normal conditions.

3.4.7 Validation of predicted disparate functions for isoforms of CDKN2a and of ANXA6

Our algorithm is more than just identifying the 'functional' isoform of a gene. The power of our algorithm to predict the functional disparity between isoforms is further illustrated by isoforms that carry out different aspects of gene functions. It is relatively common that only one isoform is predicted to be responsible for one particular function of a gene, as described in the previous section. In this section, we focus on analyzing specific examples in which isoforms of a single gene are assigned with unrelated biological functions.

CDKN2a is the only known example where alternative splicing results in different reading frames (Figure 3-7 A). Two genes (XBP1, GNAS1) produce alternate reading frames, but start with single transcript and therefore do not fall into the realm of alternative splicing. We

predicted that the two isoforms of CDKN2a would carry distinct functions for the gene (Figure 3-7 B). NM_001040654.1 is predicted to be involved in apoptotic nuclear changes with a probability 74 times the background probability, while the probability for NM_009877.2 approximates background. On the other hand, NM_009877.2 is predicted to be involved in positive regulation of the transmembrane receptor protein serine/threonine kinase signaling pathway (3 times background), while NM_001040654.1 is not. Because crystallized protein structures are available for both proteins in the human but not in the mouse, we used I-TASSER, the state-of-the-art protein structure prediction algorithm [61], to model the 3-D protein structures of the two isoforms (Figure 3-7 C-D). Although the translated products of NM_001040654.1 (168 aa) and NM_009877.2 (169 aa) are almost the same lengths, the transcript sequences are in different open reading frames. This resulted in five ankyrin repeats in the NM_001040654.1 (Figure 3-7 C), compared to a cyclin-dependent kinase inhibitor N-terminus domain in NM_009877.2 (Figure 3-7 D). The drastically different 3-D structures support the potential disparate functions predicted by our algorithm.

The function predictions for NM_009877.2 and NM_001040654.1, which we made by mining only large-scale public RNA-seq datasets, are consistent with the two distinct biological roles of the two isoforms. NM_009877.2 is an inhibitor of CDK4 kinase, a member of the Ser/Thr protein kinase family, directly supporting its role in GO:0071900, regulation of protein serine/threonine kinase activity. NM_001040654.1 encodes an alternate open reading frame (ARF) that generates a protein structurally unrelated to NM_009877.2. This protein enhances p53-dependent transactivation and apoptosis [62], supporting its role in apoptotic nuclear changes. Interestingly, although coding for structurally dissimilar proteins, both isoforms share a common functionality in cell cycle G1 control [63]. This shared functionality is correctly

predicted by our algorithm; for regulation of G1/S transition of mitotic cell cycle (GO:2000045): NM_001040654.1 has a probability 3.5 times the background probability, and NM_009877.2 has a probability 2.4 times that of background.

The CDKN2a example involves isoforms of drastically different protein domains. We used the isoforms of ANXA6 (NM_013472.4 and NM_001110211.1) to validate our model in predicting isoforms of very similar structure. The only difference between the two isoforms at the protein sequence level is the presence of six residues in the longer NM_013472.4 ('VAAEIL', 525–530) which are missing in NM_001110211.1. The three-dimensional structures of the translated sequences of these isoforms were published by some authors of this paper [64]. Although the global topology of the I-TASSER models for the two isoforms of ANXA6 is almost identical (with RMSD=0.38 Å and TM-score=0.99), there is an obvious structural variation identified by TM-align [65]. The positions of Thr-535 and Ser-537 in NM_013472.4 compared to NM_001110211.1 make NM_013472.4 more likely to undergo phosphorylation [64]. The fold changes on GO-terms related to phosphorylation by our function prediction algorithm supported the conclusion from the structural comparisons. The fold change for peptidyl-serine phosphorylation for NM_013472.4 was 3.5 compared to 1.9 for NM_001110211.1. Re-searching the mass-spectrometric data with phosphorylation on Serine or Threonine (Phospho (S) and Phospho (T)) as potential residue modification, yielded a peptide 'DQAQEDAQVAAEILEIADTPSGDKTSLETR' (found only in NM_013472.4) with 3281.506 daltons as the mh (calculated peptide mass plus a proton) indicating potential phosphorylation [66]. In contrast, the peptide that matched the spliced region in NM_001110211.1 (residues 'VAAEIL' are missing in this peptide) did not show any phosphorylation. These observations further supported our predictions. In addition, the overall

function enrichment showed the smaller isoform, NM_001110211.1 as involved in biological processes related to cell adhesion and cell migration, whereas the longer form is predicted to be involved in localization. It is important to emphasize the fact that our computational predictions based on RNA-seq data alone were able to pick up the differences between these two isoforms with 99% sequence identity and predict distinct functions for the isoforms. These findings suggest that our approach can solve the pressing need of isoform function differentiation, which would be invaluable for a better understanding of the diversity of functions created by alternative splicing of a limited set of genes.

3.5 Discussion

Gene functions are delivered through alternatively spliced transcript isoforms that encode proteins of different functions. It is highly beneficial that the investigation of functions is carried out at the isoform level. From this point of view, the standard gene function prediction paradigm has a major drawback in that it considers a gene as one single entity without differentiating its isoforms. The availability of transcript-level expression data from RNA-seq provides a rich resource for addressing this drawback. However, algorithmically, any supervised learning algorithm developed for gene function prediction cannot be directly applied to isoform function prediction because of the lack of isoform-level, 'ground-truth' functional annotations.

To address this challenge, we developed an iterative algorithm that predicts functions at the individual isoform level by conceptualizing a gene as a 'bag' of isoforms of potentially different functions. Our key idea is to iteratively extract the common pattern of a subset of isoforms across the positive genes of the function under investigation, aiming at maximizing the coherence within this subset of isoforms and the discriminative power against the other 'negative' genes (genes not related to the specific function under consideration).

Through experimental validation, we demonstrated that our approach in combination with publicly available RNA-seq data is capable of differentiating isoform functions, promising better and deeper understanding of gene functions. RNA-seq data are the richest resource for genome-wide, isoform-level data so far. But the basic concept is extendable to other large-scale datasets providing isoform-level information, such as protein domain data and post-translational regulation datasets. These datasets are not included in this study due to their strong overlap with the 'Gold-Standard' gene ontology annotations, which might lead to a circularity problem in evaluating our algorithms. Furthermore, our study only focuses on the base learner SVM. However, our approach is highly extendable to other modeling methods, such as logistical regression and random forests.

Our study is limited to the incomplete isoform catalog maintained by NCBI, but it can be readily updated whenever the genome annotation of isoforms is updated. Additionally, alternatively spliced isoforms often show tissue-specific expression and functions [23, 27, 67-70]. Our generic algorithm does not yet take the tissue-specific functionality into consideration. We expect that more accurate and biologically meaningful isoform function prediction could be achieved if tissue specificity were taken into account. As a result, our validation carried out in breast tissue is only used to validate the 'generic' functions of the isoforms. Recent studies found that the same principal isoform is often present in different tissues [71-74]. We expect that tissue-specific functions can be validated in corresponding tissues when these tissue-specific predictions can be made. Our study is further limited by the current technology to assign isoform-level expression values, as well as the differential capability between platforms for capturing isoform-specific expression. We used Cufflinks in the Tuxedo suite [32], one of the state-of-the-art algorithms, to estimate the isoform-level read counts and achieved good

performance. However, if more advanced algorithms are developed, our algorithm could directly utilize the estimates from those algorithms and generate isoform function predictions.

Our approach represents a novel and generic strategy to look at gene functions at a higher resolution. Cross-validation, literature, and experimental analysis of proteomic data provided evidence that our algorithm is powerful in differentiating isoform functions. Broadly speaking, the genomic data integration field typically relies on the supervised learning concept, which cannot generate predictions for spliced isoforms, *e.g.*, predicting gene-disease association and gene regulatory networks. We envision that similar concepts will be developed for generating isoform-level models for these prediction tasks.

3.6 Methods

3.6.1 Pre-processing public RNA-seq datasets

We downloaded 811 RNA-seq experiments for the mouse from the NCBI sequence read archive (SRA) database as of May 1, 2012 [30]. These datasets represent different conditions and tissues. Heterogeneity of the datasets allowed us to look at the isoform expression variations across different conditions and tissues. Because datasets are heterogeneous in terms of library preparation procedures and sequencing platform (Dataset S1), we adopted the following processing and filtering pipeline to ensure that all datasets included in the final predictions have sufficient coverage. NCBI build 37.2 reference genome was downloaded from the TopHat homepage and Bowtie2 [75] index files were created by bowtie2-build software. For each RNA-seq dataset, short reads were aligned against the NCBI *Mus musculus* reference genome (Build 37.2) using TopHat v2.0.051 [32, 34]. The reference GTF annotation file from NCBI (build 37.2, downloaded from TopHat homepage) was given to TopHat and the no-novel-juncs option was used. With this option, TopHat creates a database of splice junctions indicated in the supplied

GTF file and maps the previously unmapped reads against the database of these junctions to create an estimate of isoform expression levels. Then we used Cufflinks v2.0.0 [32] to measure the relative abundances of the transcripts using normalized RNA-seq fragment counts [32, 35]. The unit of expression levels is Fragments per Kilobase of exon per Million fragments mapped (FPKM). To ensure data quality and overall coverage, we removed those experiments with less than 10 million reads or with less than 50% reads being successfully mapped to the genome. The above procedure resulted in 365 experiments used in our study. Genes detected in less than half of the experiments were removed. After filtering these poorly-covered genes, there are 19209 genes left with a total of 24274 isoforms. Distribution of the number of isoforms per gene is included in Supplementary Figure 3-4. FPKM values were $\log_2$-transformed; missing values were approximated with a value of "-15".

3.6.2 Assembling gene-level gold standard functional annotations

We constructed gold standard gene functions using the Gene Ontology (GO) database [46, 47]. For each biological process term, we treated the genes which are annotated to that GO term and any of its descendent terms as positives and others as negatives. We maintained all GO evidence codes. The sources of these annotations include (1) hand annotation from primary literature, (2) electronic annotation based on gene name and symbols, (3) annotation from SwissProt keywords and (4) Enzyme Commission (EC) numbers [76]. These annotation sources are reasonably accurate for our analysis.

3.6.3 Mathematical definition and solution of the isoform function prediction problem

As stated in the Results section, the isoform function prediction problem is a Multiple Instance Learning problem. Since the introduction of MIL by [77] for the drug activity prediction, several methods to solve this problem have been proposed in the literature. In [78],

the authors introduced the concept of Diversity Density, whose aim is to find a point in feature space that has a high Diverse Density. This means high density of instances from positive bags and low density of instances from negative bags. Additionally, Ray and Page [79] proposed the method multiple-instance regression. Their algorithm assumed that each bag has a witness instance and treated it as a missing value; then the EM (Expectation-Maximization) method was used to learn the witness instances and do the regression simultaneously. Ramon et al. [80] utilized the Neural Network technique on Multiple Instance Learning and proposed the Multiple Instance Neural Network. Finally, Andrews et al. [50] applied the Support Vector Machines to Multiple Instance Learning.

In this paper, we chose to use a method that utilizes the Support Vector Machine. Two different approaches have been proposed to solve the MIL problem using SVM. The first tries to impute all non-witness instances in positive bags as negative examples and then considers the problem as a supervised learning problem. The second tries to identify a single witness from each positive bag which is responsible for positive label. Then, a classifier is built based on these witnesses only, while other instances are dropped out of the classification process. SVM formulations of these two approaches are labeled mi-SVM and MI-SVM by Andrews et al. [50]; we implemented and tested several alternatives of these algorithms.

Without loss of generality, we assume that the *i*th gene with *m* isoforms is denoted as $X_i = \{x_{i1}, x_{i2}, ..., x_{im}\}, x_{ij} \in X = R^d$. The corresponding class label y_i is set to 1 if the gene is annotated to the function under consideration and *0* if not. For each gene, we hypothesize that if a gene is annotated to a function, at least one of its isoforms should be annotated to the function; if a gene is a negative example used in training, none of its isoforms can be annotated to the function, *i.e.*,

$$y_i = \begin{cases} 1, & if\ \exists j \quad s.t.\ \ y_{ij} = 1 \\ 0, & if \quad for\ \forall j \ \ y_{ij} = 0 \end{cases} \tag{1}$$

Then, the mi-SVM formulation of MIL can be written as follows

$$\begin{aligned} & \min_{\{y_{ij}\}} \min_{w,b,\xi} \frac{1}{2}\|w\|^2 + c\sum_{i,j}\xi_{ij} \\ \text{Subject to:} \quad & \left(\langle w, x_{ij}\rangle + b\right) \geq 1-\xi_{ij}, \ \ \forall i \ if \ \ y_{ij} = 1 \\ & \left(\langle w, x_{ij}\rangle + b\right) \leq -1+\xi_{ij}, \forall i \ if \ \ y_{ij} = 0 \\ & \xi_{ij} > 0 \end{aligned} \tag{2}$$

In the standard classification setting, the labels y_{ij} of the isoform x_{ij} would be given; however, in equation (2) labels of isoforms that belong to a positive gene are treated as unobserved hidden integer variables. Therefore, the soft-margin criterion is maximized jointly over hyperplanes and over all possible label assignments. The algorithm is looking for a separating hyperplane such that all isoforms of negative genes are in the negative half-space, whereas there is at least one isoform from every positive gene in the positive half-space. Meantime, the margin is maximized with respect to the selected labels.

Alternatively, in the MI-SVM formulation, the definition of margins is extended to bag-level. Margin of a bag with respect to a separating hyperplane can be defined as the maximum margin of its instances. In the case of positive bags, bag margin is defined by the most positive instance, whereas, for negative bags, the bag margin is defined by the least negative instance. Using this definition of bag margin, the MI-SVM formulation can be written as follows:

$$\min_{s(i)} \min_{w,b,\xi} \frac{1}{2}\|w\|^2 + c\sum_{i,j}\xi_{ij}$$

Subject to:

$$\left(\left\langle w, x_{is(i)}\right\rangle + b\right) \geq 1-\xi_{ij},\ \forall i\ if\ \ y_{ij}=1 \qquad (3)$$

$$\left(\left\langle w, x_{ij}\right\rangle + b\right) \leq -1+\xi_{ij},\ \forall i\ if\ \ y_{ij}=0$$

$$\xi_{ij} > 0$$

where y_i takes the form $y_i = \operatorname{sgn}\max_j\left(\left\langle w, x_{ij}\right\rangle + b\right)$ and *s(i)* is the selector variable that denotes the isoform selected as witness from each positive gene. Note that, for the mi-SVM formulation, every instance in a positive bag has an effect on the margin maximization equation, whereas, in the MI-SVM formulation, only one instance per positive bag is taken into account, because this instance alone will determine the margin of the bag.

Finding the optimum solution to (2) and (3) is a combinatorial optimization problem, which cannot be found efficiently with the state-of-the-art tools. Therefore, we approximated the solution by using the following optimization heuristics which is proposed by Andrews *et al.* [50]. Both formulations of MIL explained above can be considered as mixed-integer problems. In the mi-SVM formulation, instance margin is maximized over hidden labels of instances in positive bags, whereas, in the MI-SVM formulation, bag margin is maximized over selector variable, which selects a single witness from every positive bag. Optimization heuristic uses the fact that, given these integer variables, the problem reduces to a quadratic programming problem which can be solved exactly. The optimization heuristic includes two steps: (i) for a set of given integer variables (i.e. hidden labels in mi-SVM and selector variable in MI-SVM), solve the soft-margin maximization problem and find the optimal separating hyper-plane, (ii) for a given separating hyper-plane, update all integer variables so that they maximize the objective locally. These two steps are run iteratively until integer variables are not updated anymore in step (ii). The following workflow explains this optimization heuristic further in detail for both formulations.

1. Initialization: Initially, we assign all instances (isoforms) in positive bags (genes) as positives, *i.e.*,

$$y_{ij} = \begin{cases} 1, & \text{if} \quad y_i = 1 \\ 0, & \text{if} \quad y_i = 0 \end{cases} \tag{4}$$

where y_i=1 if the gene is annotated to the function under consideration, and y_i=0 if otherwise.

2. Loop:

    (2.1) Model building: construct a maximum margin classification model using positive and negative instances. Using this model, we calculate a prediction score for all instances (including those instances in positive bags which are not "witnesses") in the training set.

    (2.2) Integer variable updating:

    In the mi-SVM formulation, instances in each positive bag are assigned to a label based on prediction score calculated in the previous step. One can choose different thresholds for scores to partition instances into positive and negative classes. Here, we investigated following three thresholds and chose the second threshold, since it gave the best performance (Figure 3-2)

        (i) The first threshold is equal to the mode of the distribution of scores from negative instances in training set.

        (ii) The second threshold is equal to the 75% percentile of scores of all negative instances in training set.

        (iii) The third threshold is equal to the maximum score of negative instances in the training set.

These three thresholds represent different degrees of strictness for assigning labels. The first threshold is the least strict; it assigns most of the instances from positive bags as positive, whereas the third threshold is the most strict, generally leaving only one positive instance in every positive bag.

In the MI-SVM formulation, we chose only the instance with the maximum score in each positive bag as the "witness"; the remaining instances from this bag are not assigned to any class. (*i.e.* dropped out of the margin calculation).

Note that in mi-SVM every instance in positive bags is assigned to either positive or negative, but in MI-SVM only one instance per positive bag is assigned positive and other instances from the same bag are discarded.

(2.3) Stop criterion checking: When the assignment of integer variables does not change anymore (*i.e.* label assignments of instances in positive bags for mi-SVM, witness selector variable for MI-SVM), or the assignment of integer variables reverts to one of the assignments in previous iterations, go to Step (3); otherwise go back to step (2.1).

3. Ending iteration: use the model built in the last iteration to predict all instances. At the gene level, the score of each gene is assigned as the maximum score of all its isoforms.

3.6.4 Estimation of probability score for isoforms using bootstrap bagging

For every function, we need to assign each isoform a score no matter whether the gene that the isoform belongs to has an annotation or not. We therefore used bootstrap bagging to estimate the probability that an isoform is associated with a specific biological process. Essentially all genes are sampled with replacement (0.632 bootstrap) to construct a training set.

The scores of the held-out set are recorded and the process is iterated 30 times. For each isoform, the final score is assigned with the median across all iterations.

3.6.5 Proteomic data processing

Functions of splice variants must eventually be delivered at that protein level. To test whether the predicted differential functions are correct, we compiled the data from LC-MS/MS of normal mammary tissue [81]. The original study reported that normal tissues were harvested from 5 normal mice, processed into tissue lysates and pooled. The pooled sample was digested by trypsin for mass spectrometric analysis. The mzXML files were searched against our modified ECgene database for alternative splice variant analysis using X!Tandem [82].

3.6.6 Web implementation

All prediction results are stored in MySQL databases and delivered through a searchable website: http://guanlab.ccmb.med.umich.edu/isoPred.

3.7 Author Contributions

Experiments are conceived and designed by Ridvan Eksi and Yuanfang Guan. Experiments are performed by Ridvan Eksi. The data is analyzed by Ridvan Eksi, Hong-Dong Li, Rajasree Menon, Gilbert S. Omenn, Matthias Kretzler and Yuanfang Guan. Paper is written by Ridvan Eksi, Hong-Dong Li and Yuanfang Guan. Website is developed by Ridvan Eksi and Yuchen Wen.

3.8 Figures

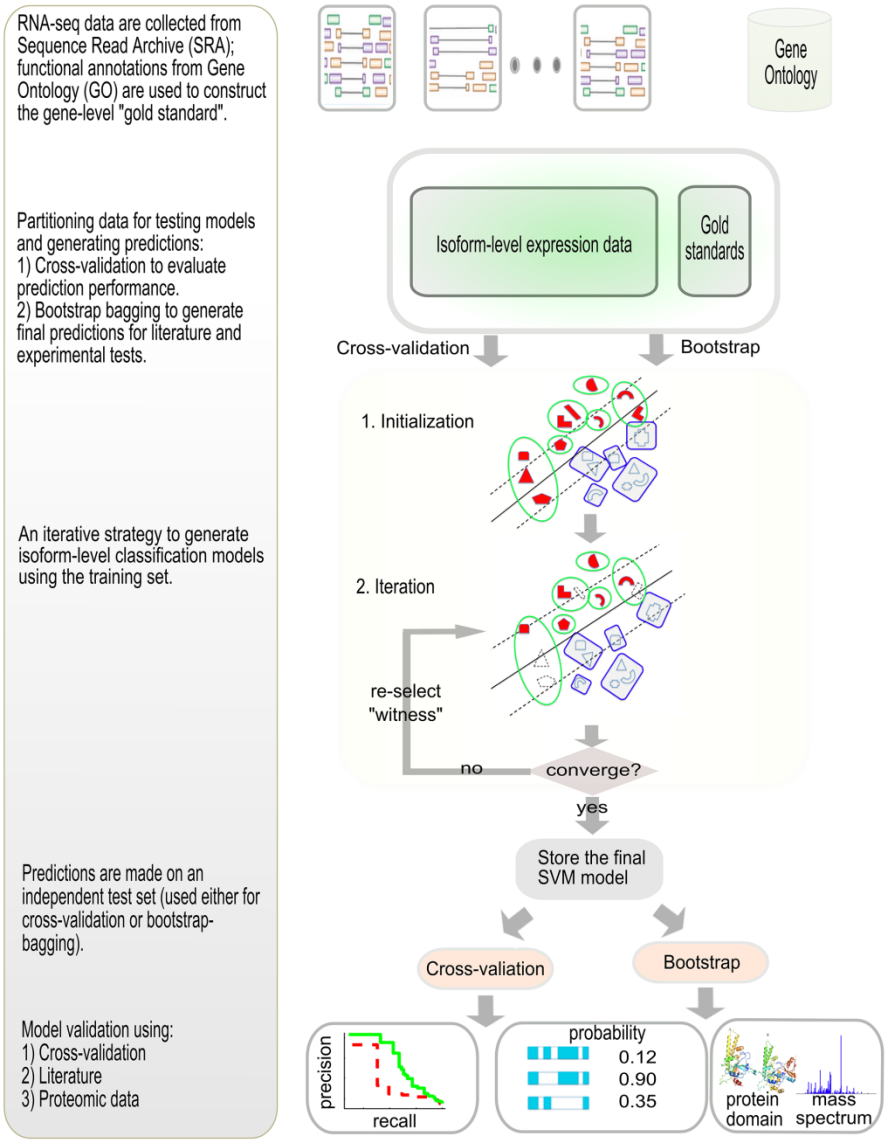


Figure 3-1 : Overview of the computational approach for predicting functions for alternatively spliced isoforms.

We collected RNA-seq data from the sequence read archive (SRA) database and estimated isoform-level expression values using state-of-the-art software [32,34]. We then generated a gene-level gold standard using Gene Ontology (GO) annotations. For each biological function, this gold standard contains positive genes (annotated to the function under investigation) and negative genes (other genes). Our study contains two major parts: cross-validation for performance estimation and bootstrap bagging for generating final predictions as well as performance evaluation. For cross-validation, we partitioned the examples into a training set for model development and a test set for model validation. For generating final predictions for all isoforms, we sampled with replacement to construct a training set, and then used this training set to construct models to assign prediction probabilities to the out-of-bag set. The final predictions

for all isoforms were made by calculating the median prediction values of all out-of-bag sets. For each training set, a model was derived from the RNA-seq data to delineate the positives and the negatives. This model was used to classify the training set and update the labels of the isoforms of the positive genes, under the criterion that at least one isoform of a positive gene must remain positive. This new assignment is then used to construct the model in the next iteration. This process is iterated until the assignment of positive isoforms no longer changes, and then the final model was used to assign a prediction value to the test or the out-of-the-bag set. Bootstrap was done for 30 iterations and the median value for each out-of-the-bag isoform was taken as the final prediction value. The predictive performance of our model was assessed through three approaches: (1) cross-validation of gene-level predictive performances, focusing on comparison between single-isoform genes and multiple-isoform genes, (2) literature validation and (3) experimental validation of top predictions using proteomic data.

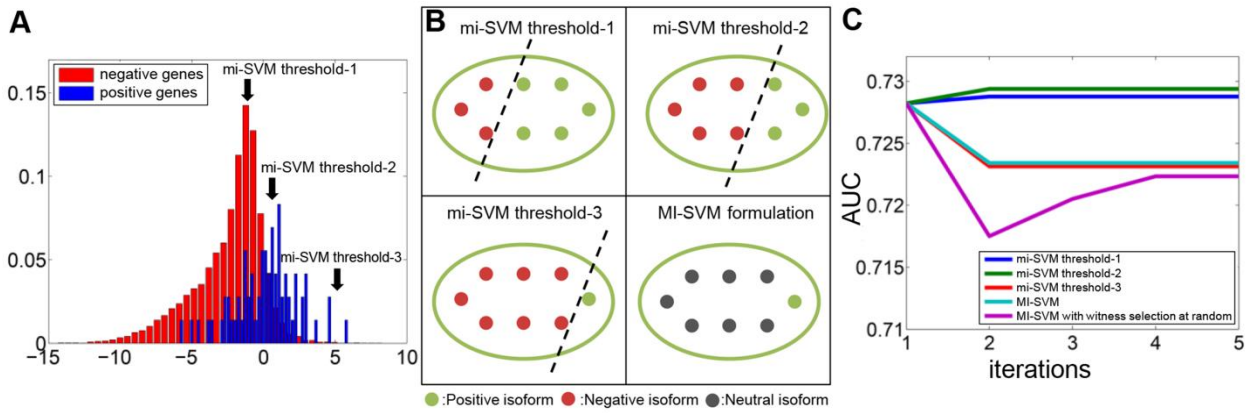


Figure 3-2: Performance comparison of different formulations of the SVM-MIL algorithm in predicting isoform functions.

A. The histogram shows the score distribution of the instances in the positive bags and the negative bags in the training set. Different threshold choices in mi-SVM are based on the distribution of scores of negative genes. The first threshold is equal to the mode of distribution of scores from negative instances in the training set. The second threshold is equal to the 75% percentile of scores of the negative instances in the training set. The third threshold is equal to the maximum score of negative instances in the training set. B. This panel illustrates how different thresholds and formulations can divide the isoforms in a positive bag into positive, negative and neutral classes. Three thresholds in mi-SVM represent different degrees of strictness for assigning labels. The first threshold is the least strict, which assigns most of the isoforms from positive genes as positive, whereas the third threshold is the strictest, which in general leaves only one positive instance in every positive bag. For the MI-SVM formulation, only one isoform per positive gene is assigned as positive, and other isoforms are dropped (*i.e.* neutral class). C. Performance comparison of three different threshold choices for the mi-SVM formulation, the MI-SVM formulation and the MI-SVM formulation with random witness selection. This plot shows that the mi-SVM formulation with threshold-2 performs best in terms of AUC.

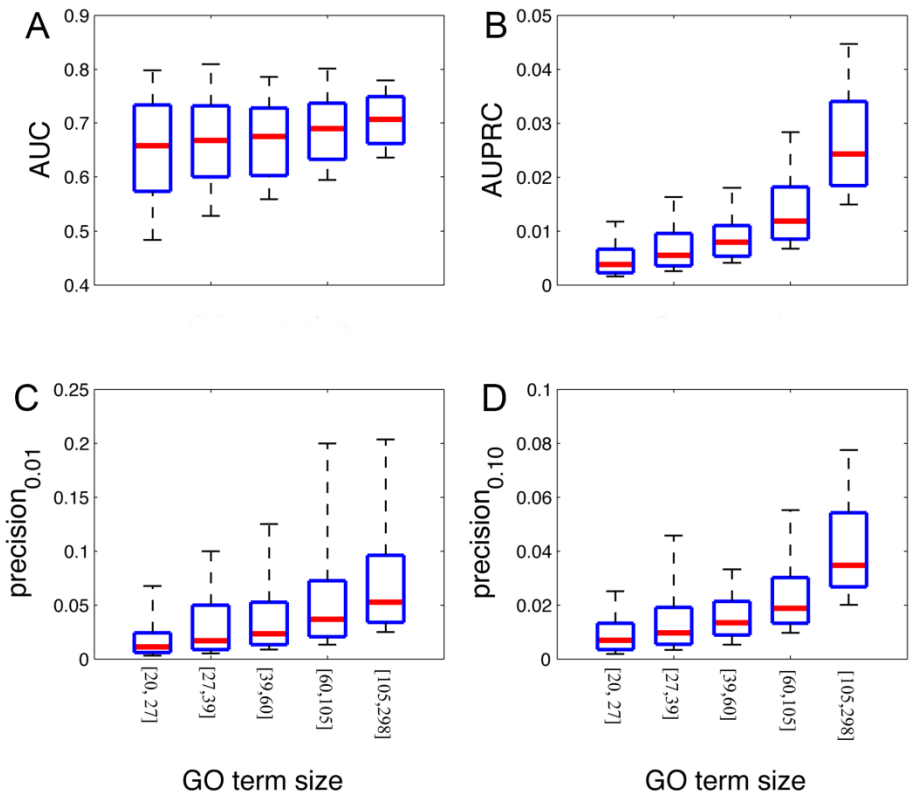


Figure 3-3: Robust performance of our algorithm to predicting functions using RNA-seq data.

We carried out five-fold cross validation to test the performance of our algorithm. For each function, the prediction value for each gene is assigned the maximum prediction value of all of its isoforms, under the assumption that at least one of its isoforms should carry out the function. Because the number of known genes of each GO term systematically affects the prediction performance, we group these terms into 5 groups according to their GO term sizes. (A)–(D) shows the distribution (10, 25, 50, 75, 90%) of the AUCs, the AUPRCs, the precisions at 1% recall and the precisions at 10% recall, respectively.

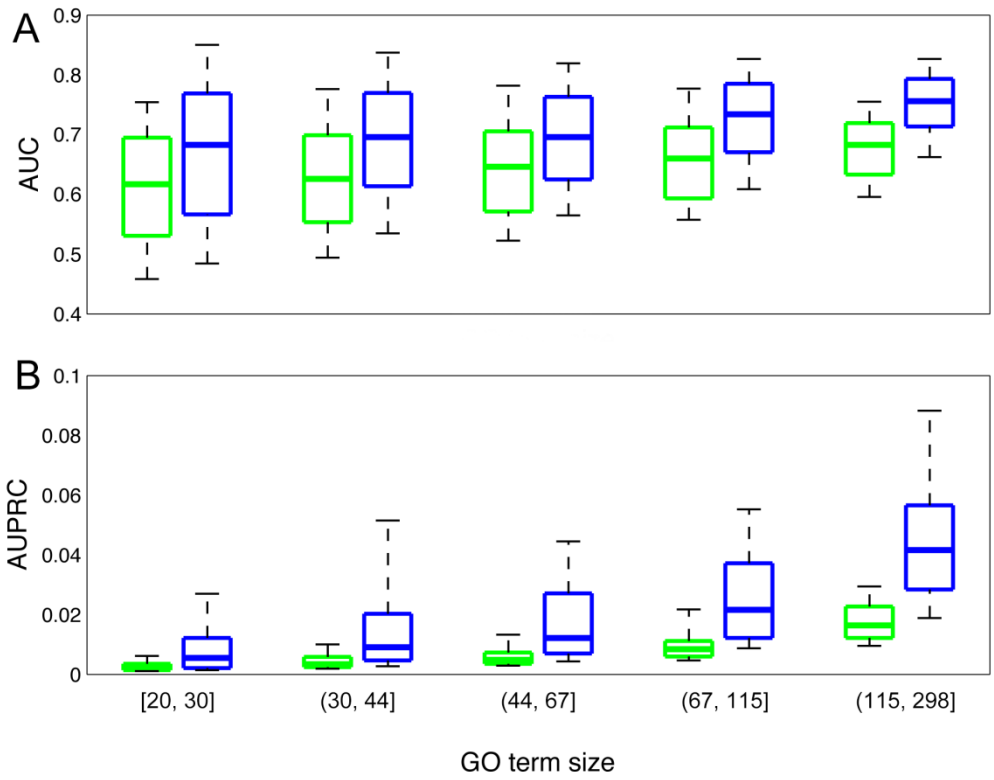


Figure 3-4: Prediction performance comparison of single-isoform genes (green) and multi-isoform gene (blue) based on AUC (upper panel) and AUPRC (lower panel).

We separately evaluated its prediction performance for single-isoform genes and multiple-isoform genes. Two-fold cross-validation was carried out to ensure enough examples in both groups. To ensure comparability, the negatives were randomly selected to ensure that the ratios of positive to negative genes for the multi-isoform group and the single isoform group are the same for each GO term. GO terms were grouped according to the number of genes in the test set. Shown in the box-plot are the AUC (**A**) and AUPRC (**B**) at 10, 25, 50, 75 and 90 percentile, respectively.

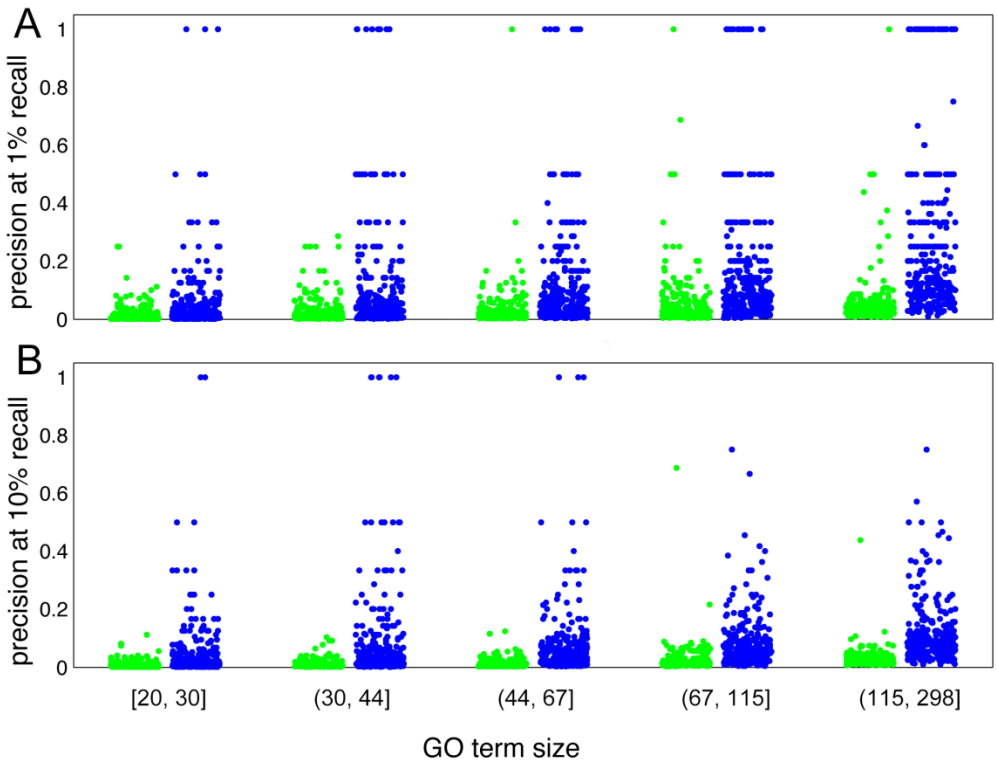


Figure 3-5: Prediction precision between single-isoform genes (green) with multi-isoform gene (blue).

Two-fold cross-validation was carried out to ensure that enough examples are included in both the single-isoform group and the multi-isoform group. The negatives were randomly selected to ensure that the ratios of positive to negative genes for the multi-isoform group and the single isoform group are the same for each GO term, so that the baseline precision for each GO term is equal for the two groups. GO terms were grouped according to the number of genes in the test set. Each dot represents the precision value of an individual GO term. A. Precision at one percent recall. B. Precision at ten percent recall.

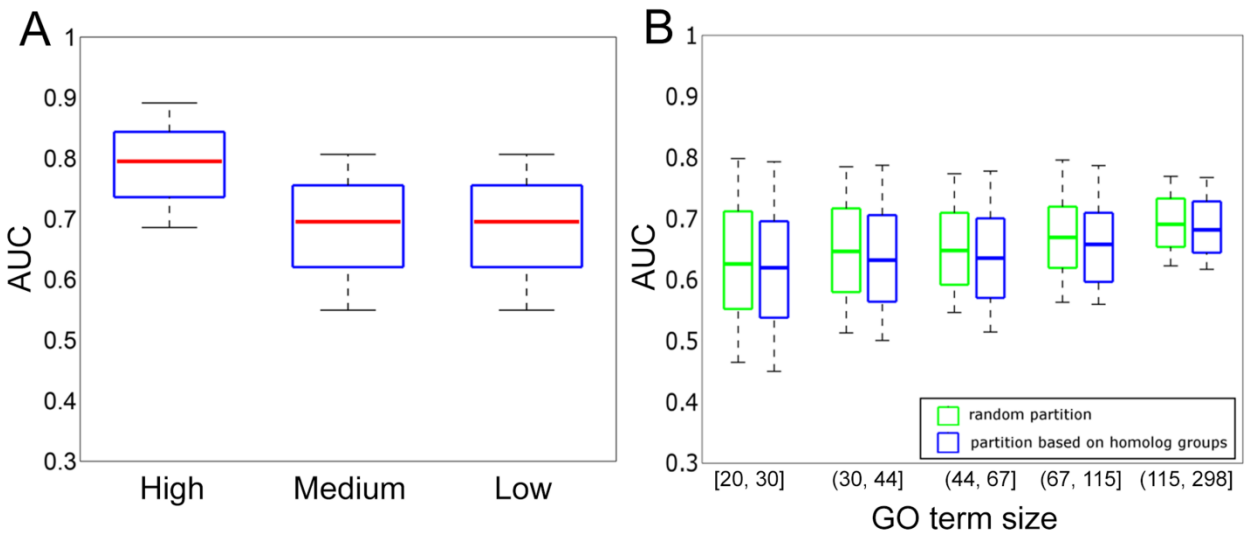


Figure 3-6: Robust performance of our algorithm in predicting isoform functions.

A. Genes are grouped according to their expression levels averaged across all samples in our RNA-seq data collection. The distribution of the performance in AUC across all GO terms is plotted using box-plot. B. The performance in AUC across all GO terms by partitioning the genes according to homologous groups between the training and the test set is compared against the performance of partitioning the genes randomly.

A

B

GO term ID	GO term name	NM_001040654.1	NM_009877.2
GO:0030262	apoptotic nuclear changes	73.9	1.4
GO:0090100	positive regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	1.2	3.1
GO:0031575	mitotic cell cycle G1/S transition checkpoint	3.5	2.4

C

D

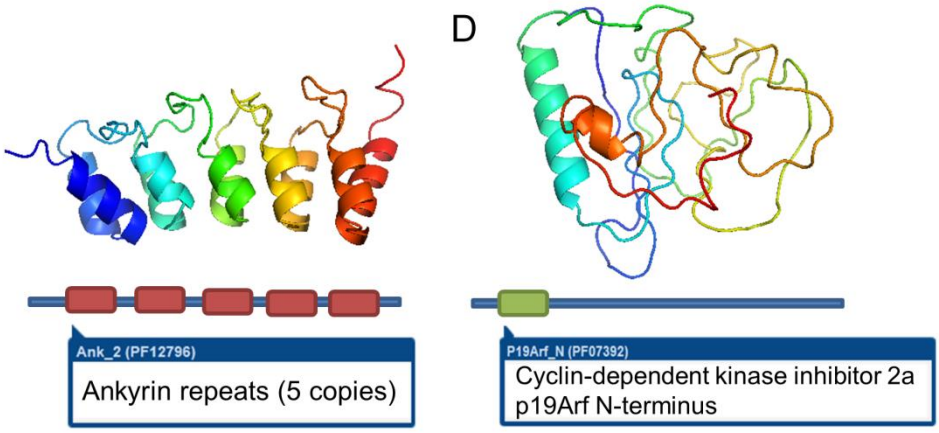


Figure 3-7: Predicted functions for isoforms of CDKN2a and their predicted protein structures.

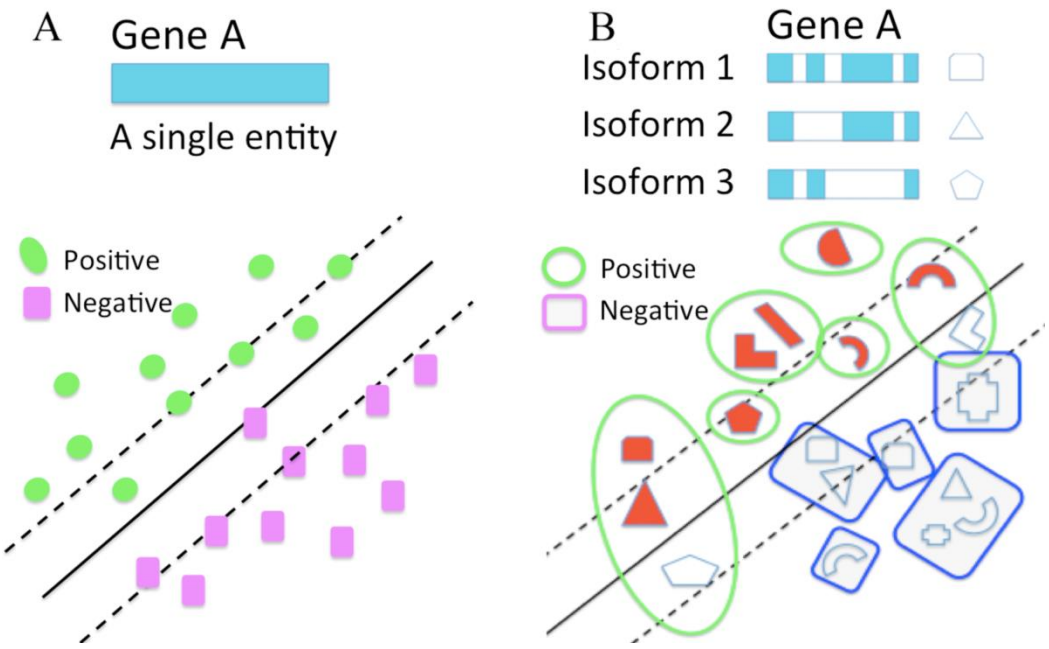
A. Gene model for NM_001040654.1 and NM_009877.2. **B**. Predicted functions for NM_001040654.1 and NM_009877.2. **C**. The computationally modeled structure of NM_001040654.1 is characterized by five ankryin repeats. **D**. The modeled structure of NM_009877.2 has a CDKN2a N-terminus domain.

3.9 Tables

Table 3-1: Examples for predicted functional isoforms that are validated using proteomic data.

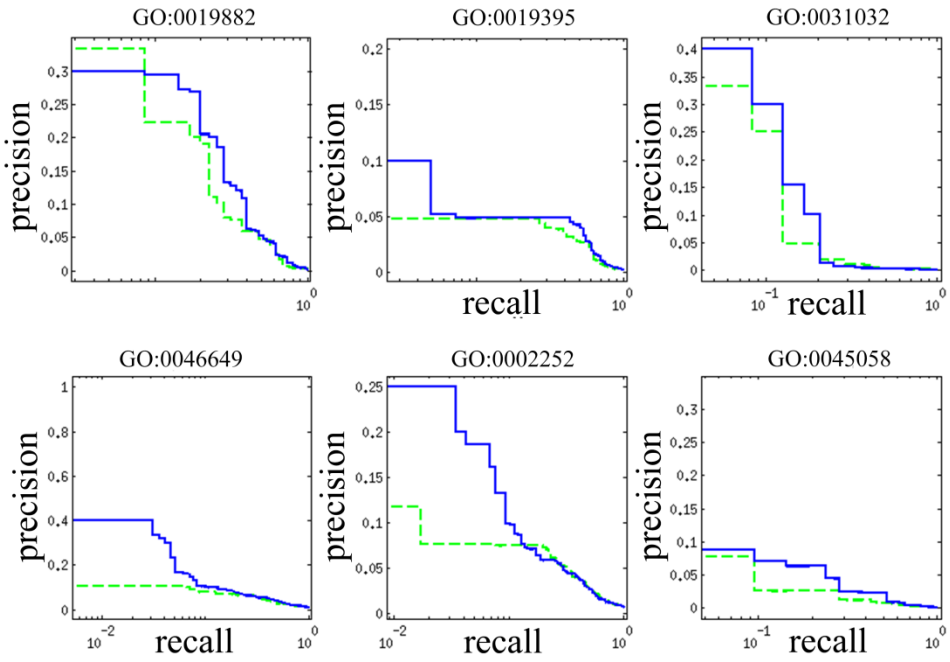
Gene Name	Identified transcript in proteomic data	GO term ID	GO term name	Fold change of prediction score
Tufm	NM_172745.3	GO:0006412	translation	34.10
Tardbp	NM_145556.4	GO:0006396	RNA processing	10.48
Tardbp	NM_145556.4	GO:0008380	RNA splicing	15.08
Tardbp	NM_145556.4	GO:0016071	mRNA metabolic process	6.15
Tardbp	NM_145556.4	GO:0043487	regulation of RNA stability	3.28
Tardbp	NM_145556.4	GO:0043488	regulation of mRNA stability	7.00
Tardbp	NM_145556.4	GO:0043489	RNA stabilization	7.50
Tardbp	NM_145556.4	GO:0048255	mRNA stabilization	6.70
Tardbp	NM_145556.4	GO:0051817	modification of morphology or physiology of other organism involved in symbiotic interaction	3.08
Ola1	NM_025942.2	GO:0006163	purine nucleotide metabolic process	2.23
Ola1	NM_025942.2	GO:0006195	purine nucleotide catabolic process	2.20
Ola1	NM_025942.2	GO:0006200	ATP catabolic process	2.92
Ola1	NM_025942.2	GO:0009141	nucleoside triphosphate metabolic process	2.53
Ola1	NM_025942.2	GO:0009143	nucleoside triphosphate catabolic process	2.32
Ola1	NM_025942.2	GO:0009144	purine nucleoside triphosphate metabolic process	2.60
Ola1	NM_025942.2	GO:0009146	purine nucleoside triphosphate catabolic process	2.38
Ola1	NM_025942.2	GO:0009154	purine ribonucleotide catabolic process	2.59
Ola1	NM_025942.2	GO:0009166	nucleotide catabolic process	2.08
Ola1	NM_025942.2	GO:0009199	ribonucleoside triphosphate metabolic process	2.29
Ola1	NM_025942.2	GO:0009203	ribonucleoside triphosphate catabolic process	2.24
Ola1	NM_025942.2	GO:0009205	purine ribonucleoside triphosphate metabolic process	2.42
Ola1	NM_025942.2	GO:0009207	purine ribonucleoside triphosphate catabolic process	2.66
Ola1	NM_025942.2	GO:0046034	ATP metabolic process	3.55
Ola1	NM_025942.2	GO:0046700	heterocycle catabolic process	2.08
Ola1	NM_025942.2	GO:0072521	purine-containing compound metabolic process	2.35
Ola1	NM_025942.2	GO:0072523	purine-containing compound catabolic process	2.10
Myom1	NM_010867.2	GO:0003012	muscle system process	21.02
Lmna	NM_001002011.2	GO:0007517	muscle organ development	3.04
Lmna	NM_001002011.2	GO:0014706	striated muscle tissue development	4.03
Lmna	NM_001002011.2	GO:0042692	muscle cell differentiation	4.69
Lmna	NM_001002011.2	GO:0051146	striated muscle cell differentiation	7.05
Lmna	NM_001002011.2	GO:0055001	muscle cell development	4.57
Lmna	NM_001002011.2	GO:0060537	muscle tissue development	5.35
Lmna	NM_001002011.2	GO:0061061	muscle structure development	4.28
Gpx3	NM_008161.2	GO:0006518	peptide metabolic process	2.16
Gpx3	NM_008161.2	GO:0006749	glutathione metabolic process	2.25
Gpx3	NM_008161.2	GO:0042743	hydrogen peroxide metabolic process	2.18
Gpx3	NM_008161.2	GO:0044106	cellular amine metabolic process	2.14
Gpx3	NM_008161.2	GO:0072593	reactive oxygen species metabolic process	2.38

3.10 Supplementary Files



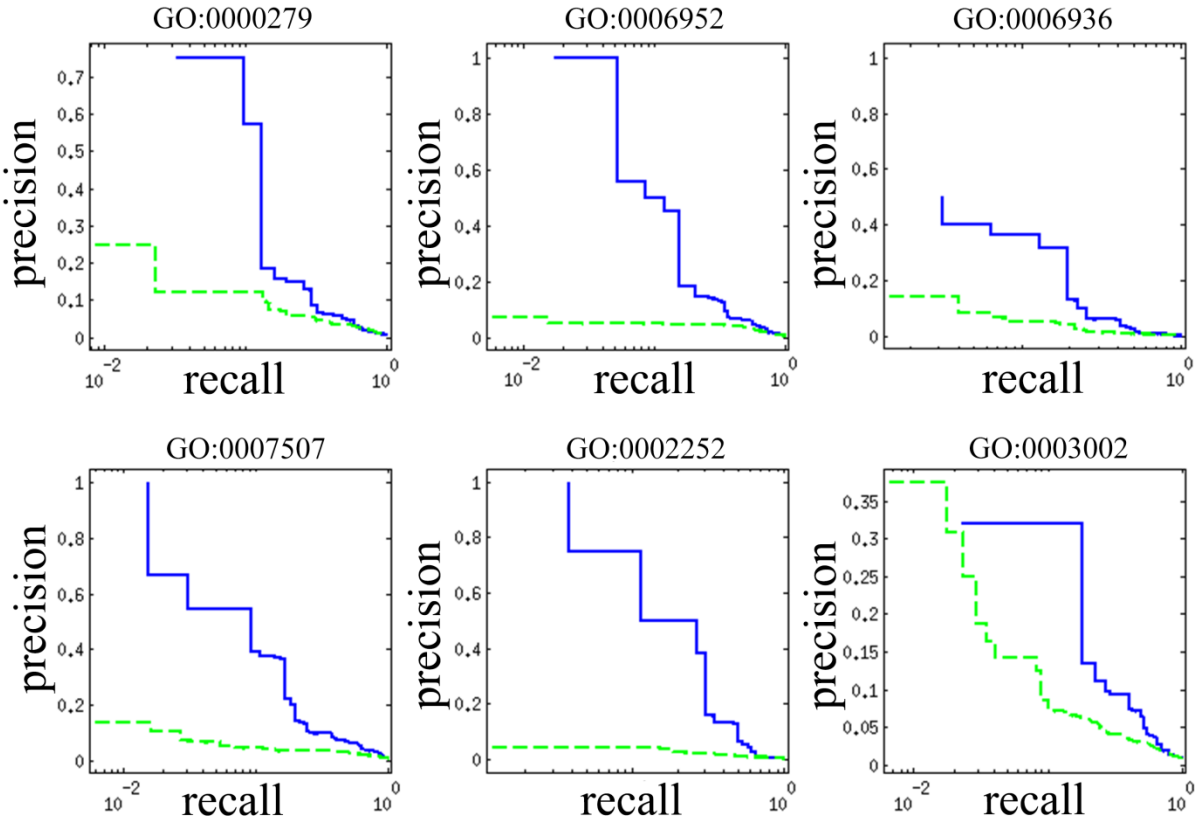
Supplementary Figure 3-1: Differences between the traditional gene function prediction problem and the isoform function prediction problem.

We use maximum margin as a base learner to illustrate the differences between a traditional classification problem for gene function prediction and our scheme for predicting isoform functions. **A**. Traditionally, a gene is treated as one single entity. The positive examples, defined as genes annotated to a specific function, are separated from the negative examples (other genes) by an SVM classifier. **B**. A single gene may contain several isoforms of which only some carry out the function under investigation. Genes here are considered as 'bags', each of which may contain one to several isoforms, defined as 'instances'. A positive gene must have at least one of its isoforms carrying out the function under consideration. None of the isoforms of a negative gene can carry out the function under study. The hyperplane trained to separate the positive isoforms and negative isoforms must satisfy the above criteria.

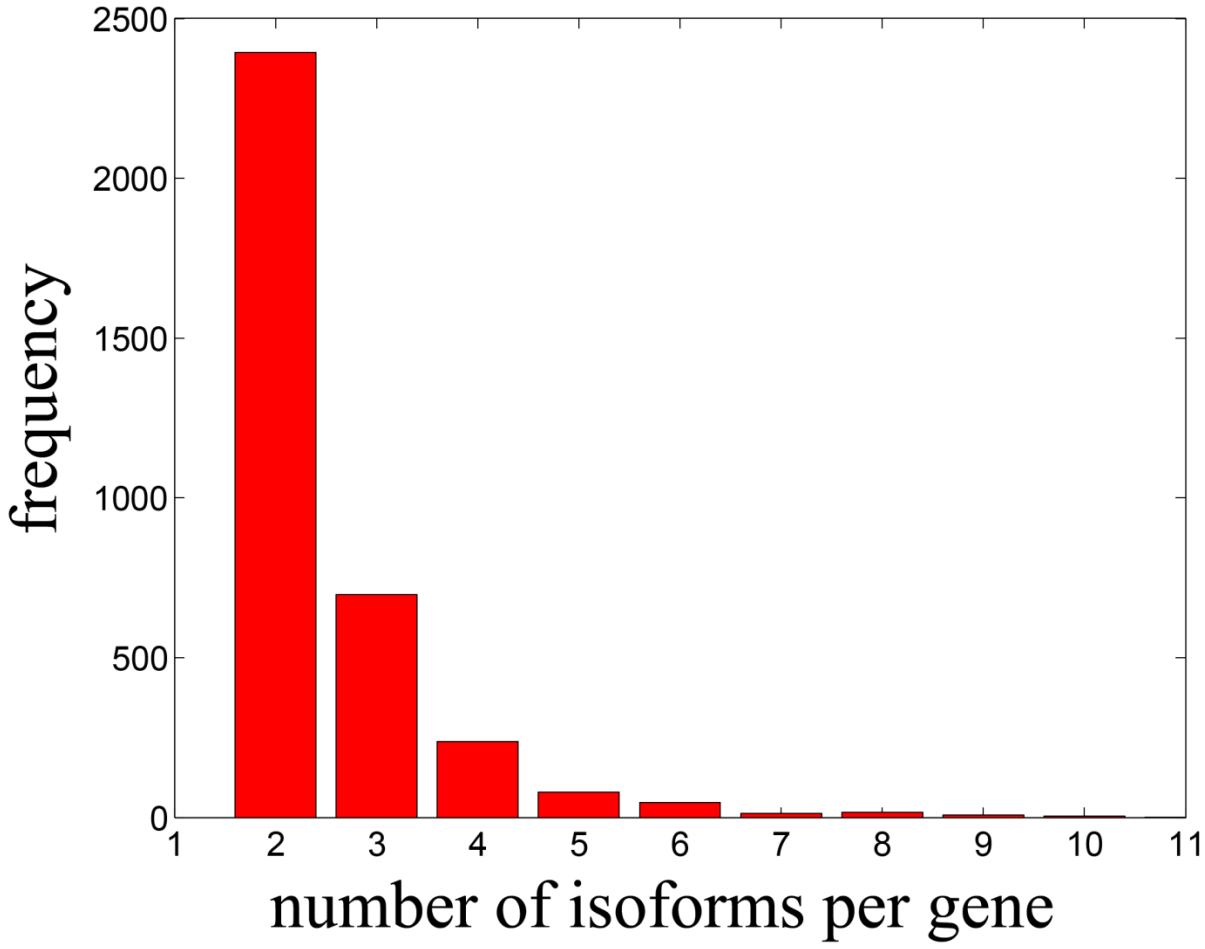


Supplementary Figure 3-2 : Comparison of gene-level prediction performance resulting from gene expression data (dashed green) and isoform expression data (solid blue).

For each GO term-specific gold standard, we developed models using gene-expression data with standard SVM and isoform-expression data with our prediction framework, respectively, and compared their precision recall curves. Shown here are six representative examples, where significant improvements were achieved when using isoform-expression data and our iterative learning strategy.



Supplementary Figure 3-3: Precision recall curve comparison between single-isoform genes (dashed green) and multiple-isoform genes (solid blue) for some GO terms.



Supplementary Figure 3-4: Histogram of number of isoforms per gene according to NCBI annotation file (build 37.2).

This figure shows only multi-isoform genes, single-isoform genes are excluded.

Supplementary Table 3-1: Example isoform groups that are predicted with differential functions.

Gene name	GO term ID	GO term name	Isoform name	Fold change
Tpm1	GO:0031032	actomyosin structure organization	NM_001164248.1	800.37
			NM_001164249.1	33.95
			NM_001164250.1	1.92
			NM_001164251.1	128.06
			NM_001164252.1	1.92
			NM_001164253.1	1.69
			NM_001164254.1	31.87
			NM_001164255.1	10.00
			NM_001164256.1	1.69
			NM_024427.4	2.03
Pde4dip	GO:0030239	Myofibril assembly	NM_001039376.2	1.99
			NM_001110163.1	98.88
			NM_177145.3	1.92
			NM_178080.4	2.19
Rtn4	GO:0006749	glutathione metabolic process	NM_024226.4	61.92
			NM_194051.3	1.57
			NM_194052.3	1.69
			NM_194053.3	1.49
			NM_194054.3	1.07
Rtn4	GO:0022029	telencephalon cell migration	NM_024226.4	1.52
			NM_194051.3	1.60
			NM_194052.3	1.72
			NM_194053.3	1.31
			NM_194054.3	23.83

Dataset S1. Input RNA-seq dataset description.

Dataset S2. Five-fold cross-validation results.

Dataset S3. Separated evaluation for genes annotated with multiple isoforms and a single isoform.

Dataset S4. Comparison between expressed splice variant in the normal breast cell sample and predicted 'functional' isoforms.

3.11 Bibliography

1. Schmitz R, Young RM, Ceribelli M, Jhavar S, Xiao W, Zhang M, et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. Nature. 2012;490(7418):116-20. doi: 10.1038/nature11378.

2. Guan Y, Ackert-Bicknell CL, Kell B, Troyanskaya OG, Hibbs MA. Functional Genomics Complements Quantitative Genetics in Identifying Disease-Gene Associations. PLoS Computational Biology. 2010;6(11):e1000991. doi: 10.1371/journal.pcbi.1000991.

3. Chen KF, Crowther DC. Functional genomics in Drosophila models of human disease. Briefings in Functional Genomics. 2012;11(5):405-15. doi: 10.1093/bfgp/els038.

4. Liang H, Cheung LWT, Li J, Ju Z, Yu S, Stemke-Hale K, et al. Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer. Genome Research. 2012;22(11):2120-9. doi: 10.1101/gr.137596.112.

5. Nelson AC, Pillay N, Henderson S, Presneau N, Tirabosco R, Halai D, et al. An integrated functional genomics approach identifies the regulatory network directed by brachyury (T) in chordoma. The Journal of Pathology. 2012;228(3):274-85. doi: 10.1002/path.4082.

6. Zhang X, Cowper-Sallari R, Bailey SD, Moore JH, Lupien M. Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. Genome Research. 2012;22(8):1437-46. doi: 10.1101/gr.135665.111.

7. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biology. 2011;9(4):e1001046. doi: 10.1371/journal.pbio.1001046.

8. Hu P, Bader G, Wigle DA, Emili A. Computational prediction of cancer-gene function. Nature Reviews Cancer. 2006;7(1):23-34. doi: 10.1038/nrc2036.

9. Guan Y, Myers CL, Hess DC, Barutcuoglu Z, Caudy AA, Troyanskaya OG. Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biology. 2008;9(Suppl 1):S3. doi: 10.1186/gb-2008-9-s1-s3.

10. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics. 2003;19(Suppl 1):i197-i204. doi: 10.1093/bioinformatics/btg1026.

11. Wu H. Prediction of functional modules based on comparative genome analysis and Gene Ontology application. Nucleic Acids Research. 2005;33(9):2822-37. doi: 10.1093/nar/gki573.

12. Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, et al. The functional landscape of mouse gene expression. J Biol. 2004;3(5):21. doi: 10.1186/jbiol16. PubMed PMID: 15588312; PubMed Central PMCID: PMCPMC549719.

13. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics. 2008;40(12):1413-5. doi: 10.1038/ng.259.

14. Matlin AJ, Clark F, Smith CWJ. Understanding alternative splicing: towards a cellular code. Nature Reviews Molecular Cell Biology. 2005;6(5):386-98. doi: 10.1038/nrm1645.

15. Skotheim RI, Nees M. Alternative splicing in cancer: Noise, functional, or systematic? The International Journal of Biochemistry & Cell Biology. 2007;39(7-8):1432-49. doi: 10.1016/j.biocel.2007.02.016.

16. Tazi J, Bakkour N, Stamm S. Alternative splicing and disease. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2009;1792(1):14-26. doi: 10.1016/j.bbadis.2008.09.017.

17. Omenn GS, Yocum AK, Menon R. Alternative Splice Variants, a New Class of Protein Cancer Biomarker Candidates: Findings in Pancreatic Cancer and Breast Cancer with Systems Biology Implications. Disease Markers. 2010;28(4):241-51. doi: 10.1155/2010/705847.

18. Hegyi H, Kalmar L, Horvath T, Tompa P. Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder. Nucleic Acids Research. 2010;39(4):1208-19. doi: 10.1093/nar/gkq843.

19. Wan J, Masuda T, Hackler L, Torres KM, Merbs SL, Zack DJ, et al. Dynamic usage of alternative splicing exons during mouse retina development. Nucleic Acids Research. 2011;39(18):7920-30. doi: 10.1093/nar/gkr545.

20. Severing EI, van Dijk ADJ, van Ham RCHJ. Assessing the contribution of alternative splicing to proteome diversity in Arabidopsis thaliana using proteomics data. BMC Plant Biology. 2011;11(1):82. doi: 10.1186/1471-2229-11-82.

21. de Souza JES, Ramalho RF, Galante PAF, Meyer D, de Souza SJ. Alternative splicing and genetic diversity: silencers are more frequently modified by SNVs associated with alternative exon/intron borders. Nucleic Acids Research. 2011;39(12):4942-8. doi: 10.1093/nar/gkr081.

22. Mittendorf KF, Deatherage CL, Ohi MD, Sanders CR. Tailoring of Membrane Proteins by Alternative Splicing of Pre-mRNA. Biochemistry. 2012;51(28):5541-56. doi: 10.1021/bi3007065.

23. Frühwald J, Camacho Londoño J, Dembla S, Mannebach S, Lis A, Drews A, et al. Alternative Splicing of a Protein Domain Indispensable for Function of Transient Receptor Potential Melastatin 3 (TRPM3) Ion Channels. Journal of Biological Chemistry. 2012;287(44):36663-72. doi: 10.1074/jbc.m112.396663.

24. Oberwinkler J, Lis A, Giehl KM, Flockerzi V, Philipp SE. Alternative Splicing Switches the Divalent Cation Selectivity of TRPM3 Channels. Journal of Biological Chemistry. 2005;280(23):22540-8. doi: 10.1074/jbc.m503092200.

25. Revil T, Toutant J, Shkreta L, Garneau D, Cloutier P, Chabot B. Protein Kinase C-Dependent Control of Bcl-x Alternative Splicing. Molecular and Cellular Biology. 2007;27(24):8431-41. doi: 10.1128/mcb.00565-07.

26. Vegran F, Boidot R, Oudin C, Riedinger JM, Bonnetain F, Lizard-Nacol S. Overexpression of Caspase-3s Splice Variant in Locally Advanced Breast Carcinoma Is Associated with Poor Response to Neoadjuvant Chemotherapy. Clinical Cancer Research. 2006;12(19):5794-800. doi: 10.1158/1078-0432.ccr-06-0725.

27. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008;5(7):621-8. doi: 10.1038/nmeth.1226.

28. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics. 2009;25(8):1026-32. doi: 10.1093/bioinformatics/btp113.

29. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009;10(1):57-63. doi: 10.1038/nrg2484.

30. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Research. 2010;39(Database):D19-D21. doi: 10.1093/nar/gkq1019.

31. Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, Troyanskaya OG. Exploring the functional landscape of gene expression: directed search of large microarray compendia. Bioinformatics. 2007;23(20):2692-9. doi: 10.1093/bioinformatics/btm403.

32. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols. 2012;7(3):562-78. doi: 10.1038/nprot.2012.016.

33. Feng J, Li W, Jiang T. Inference of Isoforms from Short Sequence Reads. Journal of Computational Biology. 2011;18(3):305-21. doi: 10.1089/cmb.2010.0243.

34. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105-11. doi: 10.1093/bioinformatics/btp120.

35. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology. 2010;28(5):511-5. doi: 10.1038/nbt.1621.

36. Kim H, Bi Y, Pal S, Gupta R, Davuluri RV. IsoformEx: isoform level gene expression estimation using weighted non-negative least squares from mRNA-Seq data. BMC Bioinformatics. 2011;12(1):305. doi: 10.1186/1471-2105-12-305.

37. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12(1):323. doi: 10.1186/1471-2105-12-323.

38. Bohnert R, Ratsch G. rQuant.web: a tool for RNA-Seq-based transcript quantitation. Nucleic Acids Research. 2010;38(Web Server):W348-W51. doi: 10.1093/nar/gkq448.

39. Verspoor KM, Cohn JD, Ravikumar KE, Wall ME. Text Mining Improves Prediction of Protein Functional Sites. PLoS ONE. 2012;7(2):e32171. doi: 10.1371/journal.pone.0032171.

40. Fischer JD, Mayer CE, Söding J. Prediction of protein functional residues from sequence by probability density estimation. Bioinformatics. 2008;24(5):613-20. doi: 10.1093/bioinformatics/btm626.

41. Vacic V, Iakoucheva LM, Lonardi S, Radivojac P. Graphlet Kernels for Prediction of Functional Residues in Protein Structures. Journal of Computational Biology. 2010;17(1):55-72. doi: 10.1089/cmb.2009.0029.

42. Thibert B, Bredesen DE, del Rio G. Improved prediction of critical residues for protein function based on network and phylogenetic analyses. BMC Bioinformatics. 2005;6:213. doi: 10.1186/1471-2105-6-213. PubMed PMID: 16124876; PubMed Central PMCID: PMCPMC1208857.

43. Kochańczyk M. Prediction of functionally important residues in globular proteins from unusual central distances of amino acids. BMC Structural Biology. 2011;11(1):34. doi: 10.1186/1472-6807-11-34.

44. Murvai J. Prediction of Protein Functional Domains from Sequences Using Artificial Neural Networks. Genome Research. 2001;11(8):1410-7. doi: 10.1101/gr.168701.

45. Rentzsch R, Orengo CA. Protein function prediction using domain families. BMC Bioinformatics. 2013;14(Suppl 3):S5. doi: 10.1186/1471-2105-14-s3-s5.

46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000;25(1):25-9. doi: 10.1038/75556.

47. Gene Ontology C. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research. 2004;32(90001):258D-61. doi: 10.1093/nar/gkh036.

48. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Research. 2009;38(suppl_1):D355-D60. doi: 10.1093/nar/gkp896.

49. Babenko B. Multiple instance learning: algorithms and applications. View Article PubMed/NCBI Google Scholar. 2008.

50. Andrews S, Hofmann T, Tsochantaridis I. Multiple instance learning with generalized support vector machines. Eighteenth National Conference on Artificial Intelligence (Aaai-02)/Fourteenth Innovative Applications of Artificial Intelligence Conference (Iaai-02), Proceedings. 2002:943-4. PubMed PMID: WOS:000183593700139.

51. Zhang C, Platt JC, Viola PA, editors. Multiple instance boosting for object detection. Advances in neural information processing systems; 2006.

52. Bunescu RC, Mooney RJ. Multiple instance learning for sparse positive bags. Proceedings of the 24th international conference on Machine learning - ICML '07: ACM Press; 2007.

53. Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995. xv, 188 p. p.

54. Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG. Finding function: evaluation methods for functional genomic data. BMC Genomics. 2006;7:187. doi: 10.1186/1471-2164-7-187. PubMed PMID: 16869964; PubMed Central PMCID: PMCPMC1560386.

55. Fu WJ, Carroll RJ, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics. 2005;21(9):1979-86. doi: 10.1093/bioinformatics/bti294.

56. Peña-Castillo L, Tasan M, Myers CL, Lee H, Joshi T, Zhang C, et al. A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biology. 2008;9(Suppl 1):S2. doi: 10.1186/gb-2008-9-s1-s2.

57. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research. 2012;22(9):1760-74. doi: 10.1101/gr.135350.111.

58. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Research. 2008;19(2):327-35. doi: 10.1101/gr.073585.107.

59. Finsterer J. Ataxias with Autosomal, X-Chromosomal or Maternal Inheritance. Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques. 2009;36(04):409-28. doi: 10.1017/s0317167100007733.

60. Menon R, Omenn GS. Identification of Alternatively Spliced Transcripts Using a Proteomic Informatics Approach. Methods in Molecular Biology: Humana Press; 2010. p. 319-26.

61. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols. 2010;5(4):725-38. doi: 10.1038/nprot.2010.5.

62. Liontos M, S. Pateras I, Evangelou K, G. Gorgoulis V. The Tumor Suppressor Gene ARF as a Sensor of Oxidative Stress. Current Molecular Medicine. 2012;12(6):704-15. doi: 10.2174/156652412800792633.

63. Ivanchuk SM, Mondal S, Dirks PB, Rutka JT. The INK4A/ARF locus: role in cell cycle control and apoptosis and implications for glioma growth. J Neurooncol. 2001;51(3):219-29. PubMed PMID: 11407594.

64. Menon R, Roy A, Mukherjee S, Belkin S, Zhang Y, Omenn GS. Functional Implications of Structural Predictions for Alternative Splice Proteins Expressed in Her2/neu–Induced Breast Cancers. Journal of Proteome Research. 2011;10(12):5503-11. doi: 10.1021/pr200772w.

65. Zhang Y. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research. 2005;33(7):2302-9. doi: 10.1093/nar/gki524.

66. Omenn GS, Menon R, Zhang Y. Innovations in proteomic profiling of cancers: Alternative splice variants as a new class of cancer biomarker candidates and bridging of proteomics with structural biology. Journal of Proteomics. 2013;90:28-37. doi: 10.1016/j.jprot.2013.04.007.

67. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. Nature. 2010;465(7294):53-9. doi: 10.1038/nature09000.

68. Bertone P. Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. Science. 2004;306(5705):2242-6. doi: 10.1126/science.1103388.

69. Black DL. Mechanisms of Alternative Pre-Messenger RNA Splicing. Annual Review of Biochemistry. 2003;72(1):291-336. doi: 10.1146/annurev.biochem.72.121801.161720.

70. Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, et al. Revealing Global Regulatory Features of Mammalian Alternative Splicing Using a Quantitative Microarray Platform. Molecular Cell. 2004;16(6):929-41. doi: 10.1016/j.molcel.2004.12.004.

71. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489(7414):101-8. doi: 10.1038/nature11233.

72. Tress ML, Wesselink J-J, Frankish A, López G, Goldman N, Löytynoja A, et al. Determination and validation of principal gene products. Bioinformatics. 2007;24(1):11-7. doi: 10.1093/bioinformatics/btm547.

73. Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink J-J, Lopez G, et al. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Research. 2012;41(D1):D110-D7. doi: 10.1093/nar/gks1058.

74. Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biology. 2013;14(7):R70. doi: 10.1186/gb-2013-14-7-r70.

75. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9(4):357-9. doi: 10.1038/nmeth.1923.

76. Hill DP, Davis AP, Richardson JE, Corradi JP, Ringwald M, Eppig JT, et al. Program description: Strategies for biological annotation of mammalian systems: implementing gene ontologies in mouse genome informatics. Genomics. 2001;74(1):121-8. doi: 10.1006/geno.2001.6513. PubMed PMID: 11374909.

77. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence. 1997;89(1-2):31-71. doi: 10.1016/s0004-3702(96)00034-3.

78. Maron O, Lozano-Pérez T, editors. A framework for multiple-instance learning. Advances in neural information processing systems; 1998.

79. Ray S, Page D, editors. Multiple instance regression. ICML; 2001.

80. Ramon J, De Raedt L, editors. Multi instance neural networks. Proceedings of the ICML-2000 workshop on attribute-value and relational learning; 2000.

81. Whiteaker JR, Zhang H, Zhao L, Wang P, Kelly-Spratt KS, Ivey RG, et al. Integrated Pipeline for Mass Spectrometry-Based Discovery and Confirmation of Biomarkers Demonstrated in a Mouse Model of Breast Cancer. Journal of Proteome Research. 2007;6(10):3962-75. doi: 10.1021/pr070202v.

82. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004;20(9):1466-7. doi: 10.1093/bioinformatics/bth092.

CHAPTER 4

Functional annotation of human protein-coding splice variants[‡]

4.1 Abstract

The vast majority of human multi-exon genes undergoes alternative splicing and produces a variety of splice variant transcripts and proteins, which can perform different functions. These protein-coding splice variants (PCSVs) greatly increase the functional diversity of proteins. Most functional annotation algorithms have been developed at the gene level; the lack of isoform-level gold standards is an important intellectual limitation for currently available machine learning algorithms. The accumulation of a large amount of RNA-seq data in the public domain greatly increases our ability to examine the functional annotation of genes at isoform level. In the present study, we used a multiple instance learning (MIL)-based approach for predicting the function of PCSVs. We used transcript-level expression values and gene-level functional associations from the Gene Ontology database. A support vector machine (SVM)-based 5-fold cross-validation technique was applied. Comparatively, genes with multiple PCSVs performed better than single PCSV genes, and performance also improved when more examples were available to train the models. We demonstrated our predictions using literature evidence of ADAM15, LMNA/C, and DMXL2 genes. All predictions have been implemented in a web

[‡] Chapter 4 is published as Bharat Panwar, Rajasree Menon, **Ridvan Eksi**, Hong-Dong Li, Gilbert S. Omenn, and Yuanfang Guan. "Genome-wide functional annotation of human protein-coding splice variants using multiple instance learning." *Journal of Proteome Research* 15, no. 6 (2016): 1747-1753.

resource called “IsoFunc”, which is freely available for the global scientific community through http://guanlab.ccmb.med.umich.edu/isofunc.

4.2 Introduction

Functional annotation of the gene is an essential task for understanding both biological significance and underlying mechanisms of a particular gene [1-3]. In higher eukaryotes, alternative splicing plays a central role in gene regulation; ~95% of human multiexon genes undergo alternative splicing [4]. It produces different splice variants from the single gene through such mechanisms as exon skipping, mutual exclusion of exons, alternative 5′ donor site, alternative 3′ acceptor site, and intron retention [5]. These numerous splicing events lead to the complex human transcriptome. There are more than 80 000 protein-coding transcripts encoded by fewer than 20 000 genes. It has been estimated that these transcripts synthesize 250 000 to 1 million different proteins [6]; these alternative splicing events significantly increase the diversity of the human proteome [7]. Consequently, different protein sequences from different protein-coding splice variants (PCSVs) can have different biological functions, sometimes even opposite functions. As an example, the two PCSVs Bcl-x(S) and Bcl-x(L) of B-cell lymphoma-x (Bcl-x) gene have pro-apoptotic and antiapoptotic functions, respectively [8]. Additionally, genetic variants alter the splicing patterns, and more of the human disease-causing point mutations affect splicing than coding sequences [9]. Aberrant splicing causes many cellular abnormalities and leads to various human diseases [10-13]. A recent genome-wide variation study suggested that genetic variants affecting RNA splicing contribute to such diseases as colorectal cancer, spinal muscular atrophy, and autism spectrum disorder [14].

The functional diversity of alternative PCSVs is a major challenge for existing functional annotation algorithms, where each gene is commonly considered as a single entity without

recognizing the effects of splicing on protein function [15]. The major problem is the unavailability of a systematic catalog of isoform functions. Most of the previous annotations are based on Gene Ontology (GO), a database of controlled and dynamic vocabulary of many defined biological functions at gene level [16]. Accumulating evidence of isoform functional diversity through different experiments gives us a chance to revisit functional genomics. Large-scale whole transcriptome sequencing (RNA-seq) data provide unprecedented amounts of transcript-level expression data that can be used for developing predictive models; these methods provide high-resolution information about gene function [17]. Coexpression-based functional assignment of a gene is an established and useful strategy across diverse species [18-20]. Furthermore, suitable machine learning algorithms can improve prediction performance significantly [21].

An important challenge is how to coordinate gene-level annotation with transcript-level expression patterns. A multiple-instance learning (MIL) technique [22] has been applied to solve this kind of problem [15, 23]. Eksi et al. [15] developed an MIL-based generic framework for identifying splice variants from a set of well-annotated genes that perform a particular function in the mouse. We adopted a similar MIL-based strategy for functional annotation of human PCSVs and cross-validated prediction results computationally and from published literature. We also developed an open web resource for use by the scientific community.

4.3 Material and Methods

4.3.1 Preprocessing of RNA-seq Data Sets

In this study, initially we used all 573 human RNA-seq data sets (runs) of the ENCODE project [1]. These data contain samples from different tissues and conditions. Therefore, a systematic data processing approach was implemented to ensure the quality of such

heterogeneous data (Figure 4-1). The human genome build GRCh37.75 from Ensembl was used to align the short-reads of each RNA-seq data set using TopHat (v.2.0.11) [24]. A GTF annotation file of the same build was used with an option of no-novel-junctions. Then, we employed Cufflinks to calculate the relative abundance of the transcript as fragment per kilobase of exon per million fragments mapped (FPKM). Because the coverage of mapped reads is different for different RNA-seq samples, we used only 248 of the 573 runs, those containing more than 50% mapping coverage of total reads. These 248 runs originated from 127 different samples, so we calculated the average expression values (FPKM) for each sample separately, comprising a total of 214,292 splice variants of 63,783 genes.

We observed that FPKM values are exceptionally higher for very short transcripts (e.g., tRNAs); therefore, we removed the genes where the average length of all of the splice variants is equal to or less than 100 nucleotides and obtained 59,159 genes for further use. It is important to have sufficient nonzero values in the feature vector during the machine learning step and to enhance signal-to-noise; therefore, we used only the 14,339 genes detected as >1.0 FPKM in >50% of samples. This subset of genes contains 119,041 splice variants that include different biotypes such as protein_coding, nonsense_mediated_decay, non_stop_decay, processed_pseudogene, processed_transcript, and retained_intron. We focused only on protein-coding splice variants (PCSVs); this term was previously used to report evidence of PCSVs of choline acetylase [25] and of CNTNAP2 [26] and distinguish them from noncoding splice variant transcripts. Thus, we analyzed 59,297 splice variants (of 11,946 genes) marked as protein_coding (37,811 known, 7555 novel and 13,931 putative) biotype in the Ensembl database.

There are many zero FPKM values present at the transcript level. The FPKM values were $\log_2$-transformed and while $\log_2$-transforming all FPKM values (transcript level), those zero values give −Infinity value, which is not acceptable as an input for machine learning. Therefore, we have to approximate those zero values with some negative values (e.g., −5). In this way we distinguished zero FPKM values from the FPKM values ≥1.0. We investigated how sensitive the results are to this choice using 94 GO terms (size 20–300) from GO-slim. We found that approximation with −5, −10, and −15 values is giving median AUC values of 0.649, 0.641, and 0.637, respectively. This showed that choice of approximation only slightly changes the performance and all zero FPKM values were approximated with a value of '-5′ as an input for the machine learning.

4.3.2 Gene-Level Gold Standards Based on Gene Ontology

We downloaded Gene Ontology (data-version: releases/2014-05-27) and used biological process terms with their gene annotations. Each biological process GO term has multiple genes annotated to it; therefore, all of those genes and other genes from its descendent GO terms have been assigned as positive for that particular GO term, and the remaining genes have been designated as negative. GO terms-based functional annotation is well known from previous studies [15, 23]. There are 12,584 GO terms for biological processes. The number of positive genes for these GO terms varies. Some GO terms are very broad and contain a very high number of positive genes; while other GO terms contain very few positive genes. Previously, it was shown that GO term size 20–300 is robust for cross-validation [27]. Therefore, we used only GO terms with minimum 20 and maximum 300 positive genes; we found 2129 GO terms in this range. For analyzing results of different sizes, we nearly equally divided all GO terms into five ranges: (A) 20–27, (B) 28–40, (C) 41–64, (D) 65–114, and (E) 115–300.

4.3.3 Multiple Instance Learning and SVM (MIL-SVM)

We adopted a hypothesis similar to that of Eksi et al. [15] that of many splice variants of a positive gene at least one splice variant is responsible for performing the GO term function. Similarly, all splice variants of negative genes are not responsible for that particular function. A gene is termed as a "bag" and each splice variant of that gene is termed as an "instance". The aim of MIL is to identify subsets of splice variants from positive genes and maximize the difference between them and splice variants of negative genes. We used maximum-margin-based classification [28] to maximize the difference between positive and negative PCSVs. An initial subset of positive PCSVs was iteratively refined to get the final selection of PCSVs that optimizes the objective function. Although the MIL approach can be used with any machine learning technique, we chose to use SVM as a base learner because of its success in functional predictions [15, 29, 30].

4.3.4 5-Fold Cross-Validation and Performance Evaluation

Cross-validation is a well-adopted technique for calculating prediction performance. For each GO term, positive and negative genes were partitioned into five subsets. We used a gene-level partition instead of PCSV-level to prevent leaking information during the evaluation process. There was a balancing problem between positive and negative genes; therefore, we used an equal number of positive genes in all five subsets; negative genes were also equally distributed in these five subsets. Then, five sets were created using one positive and one negative subset in each set. During SVM-based machine learning, we used four sets for training and the remaining fifth set for testing. This step was repeated five times, so each set was used once for testing [31-33]. Finally, the average performance of all five test sets was calculated. The

prediction performances were calculated in terms of area under the ROC curve (AUC) and area under the precision-recall curve (AUPRC).

4.3.5 SVM Classification Score at PCSV-Level and Fold Change

The SVM classification scores for each PCSV have been calculated using testing sets in the 5-fold cross-validation. The fold change value was calculated for each PCSV for every GO term. Fold change is a ratio of the rank probability of a PCSV to the base probability [15]. PCSVs were first ranked based on the SVM classification scores, and then rank probability was calculated. Rank probability is a ratio of occurrence of positive PCSVs (PCSVs from positive genes) to the number of PCSVs in the subset of a sorted list that ranked higher than the PCSV of interest. Base probability is a ratio of the total number of positively annotated genes to the total number of genes.

4.4 Results and Discussion

4.4.1 Abundance of Protein-Coding Splice Variants

We analyzed a total of 59,297 PCSVs of 11,946 well-expressed genes. Figure 4-2 shows the distribution of number of PCSVs per gene. Although the average gene contains five PCSVs, there are quite a few genes with more than 10 PCSVs.

4.4.2 Prediction of PCSV-Level Function

We used expression-based input features to learning SVM models and predicted the functions for PCSVs. The prediction performance was calculated for all 2129 GO terms; median performances of 0.641 AUC and 0.011 AUPRC were achieved. GO:0019083 (viral transcription), GO:0006415 (translational termination), GO:0006614 (SRP-dependent cotranslational protein targeting to membrane), GO:0006613 (cotranslational protein targeting to

membrane), and GO:0072599 (establishment of protein localization to endoplasmic reticulum) were the top five performing GO terms and achieved AUC values of 0.989, 0.983, 0.966, 0.961, and 0.961, respectively. We divided GO terms into five equal parts based on their sizes. We calculated separate prediction performances for these five parts and found slightly better performance with increasing GO term size. The median AUC values for these sets A, B, C, D, and E are 0.616, 0.625, 0.641, 0.644, and 0.656 respectively (Figure 4-3 A); it means a higher number of positive examples is useful in the multiple-instance-based SVM learning. Similarly, the median AUPRC values for different GO term sizes A, B, C, D, and E are 0.004, 0.006, 0.010, 0.015, and 0.033 respectively (Figure 4-3 B). The baseline values for these A, B, C, D, and E are 0.0019, 0.0028, 0.0043, 0.0071, and 0.0155 respectively (Figure 4-3 B).

4.4.3 Performance Comparison of Single and Multiple Protein-Coding Splice Variants

In the previous section, mixed performance was evaluated from all the genes, whether those genes contain single PCSV or multiple PCSVs (Figure 4-3). The sole purpose of our prediction tool is to functionally discriminate PCSVs within a gene. Therefore, performance for multiple PCSV genes should be higher than for single PCSV genes [15]. Genes in our data set contain different numbers of PCSVs (Figure 4-2), so we divided genes into four different groups (single, 2–5, 6–15, and 16–66) based on the number of PCSVs. We found that AUC performance consistently improves with a higher number of PCSVs for all GO term sizes (Figure 4-4 A). Performance of single PCSV genes is lower in comparison with multiple PCSVs genes in terms of AUPRC (Figure 4-4 B).

4.4.4 Illustration of Predicted Distinct Functions for PCSVs of ADAM15 and LMNA/C

Although cross-validation-based computational evaluation provided good performance parameters, it is still important to validate predictions with real life examples. In this section, we

highlight two genes, ADAM15 and LMNA/C, for which there are available published experimental studies of splice isoforms, permitting a test of whether our PCSV functional predictions agreed.

4.4.4.1 ADAM15

ADAM Metallopeptidase Domain 15 (ADAM15) is a type I transmembrane glycoprotein known to be involved in cell adhesion. Zhong et al. [34] cloned and characterized alternatively spliced forms of ADAM15 in human breast cancers. They showed that higher levels of two PCSVs (ADAM15A and ADAM15B) were associated with poorer relapse-free survival in node-negative patients. The ADAM15A (ENST00000271836, 814aa) and ADAM15B (ENST00000355956, 839aa) PCSVs differentially affected cell adhesion and invasiveness [34]. Cell adhesion, migration, and invasion were enhanced by expression of ADAM15A, whereas ADAM15B led to reduced adhesion. The fold changes for the GO terms linked to cell adhesion predicted by our function prediction algorithm tend to agree with this study: positive regulation of cell adhesion (GO:0045785) and regulation of cell–substrate adhesion (GO:0010810) were two times higher for ADAM15A compared with that of ADAM15B.

Using our algorithm, the top-ranking GO term for ADAM15B was positive regulation of B cell activation (GO:0050871; 40 fold increase); in contrast, no change was observed for this term for ADAM15A. Similarly, observations with high fold changes for other immune-related terms (immune effector process, positive regulation of immune response, and immune response-activating signal transduction) were found for ADAM15B and not for ADAM15A. Several studies have been published on the role of ADAM15 as a mediator of immune mechanisms underlying inflammation [35]. ADAM15 accounted for the increased level of soluble CD23 in synovial fluid and sera of rheumatoid arthritis patients [36]. Because CD23 is known to stimulate

immune cells [37], ADAM15 could play a role by amplifying inflammation of RA synovitis [35]. These reports and our function predictions imply that isoform ADAM15B may be much more involved in the B-cell-mediated immune mechanisms than ADAM15A.

4.4.4.2 LMNA/C

LMNA/C protein is a component of the nuclear lamina and plays an important role in nuclear assembly, chromatin organization, nuclear membrane, and telomere dynamics.

Experimental functional validation of three main isoforms of LMNA has been reported by Lopez-Mejia et al. [38] The three PCSVs are lamin A (ENST00000368300, 664 aa), progerin (ENST00000368299, 614aa), and lamin C (ENST00000368301, 572aa). Lamin C expression is mutually exclusive with lamin A and progerin splice variants and occurs by alternative polyadenylation. Lopez-Mejia et al. [38] reported antagonistic functions of these three PCSVs in energy expenditure and life span. They found that mice with just lamin C expression live longer and have decreased energy metabolism, increased weight gain, and reduced respiration rate. Increased metabolism was observed in mice that expressed progerin. According to our function predictions, GO terms related to metabolic terms showed high fold changes for lamin A and progerin compared with minimal to no change in lamin C.

Each of these three PCSVs had a unique top-ranking GO term associated with it. Protein targeting to membrane (GO:0006612) was one of the top-ranking GO terms (12-fold increase) for lamin A; no fold change was observed for progerin and lamin C for this term. Substrate adhesion-dependent cell spreading (GO:0034446) was the top-ranking term for progerin (12-fold), with no change observed for that of lamin A and lamin C. Regulation of cation channel activity (GO:2001257) was one of the top terms (~3 fold increase) for lamin C; little change was found for lamin A and progerin. Previously published studies have linked lamin A/C with protein

targeting to membrane [39] and cell spreading [40]; however, experimental evidence of the individual roles of specific human lamin A/C splice variants has not yet been reported, nor has expression been characterized at isoform level across different tissues.

4.4.5 Identification of Alternative Splice Variants with Distinct Functions in HER2+/ER–/PR– Breast Cancers

In a recent publication we highlighted six alternative/noncanonical splice variants that were significantly overexpressed in HER2+/ER–/PR– breast cancers compared with normal mammary, triple-negative breast cancer, and triple-positive breast cancer tissues (HER2+/ER+/PR+) [41]. We were able to infer possible distinct functions for these six splice variants (DMXL2 isoform 3, HIF1A isoform 3, KLC1 isoform c, LNPEP isoform 2, RICTOR isoform 3, and RNF216 isoform 2) compared with their corresponding canonical forms. Biological processes including cell cycle events and glycolysis were linked to four of these six proteins [41]. For example, glycolysis was the top-ranking functional process for DMXL2 isoform 3, with a fold change of 27 compared with just 2 for the canonical protein. No previous reports link DMXL2 with any metabolic processes; the canonical protein is known to participate in signaling pathways [41].

These predictions alone cannot provide a complete functional annotation of each splice variant; however, integrating our predictions with other information such as amino acid sequence, 3D structure, and functional domains can definitely help to infer the potential function of a splice variant [41].

4.4.6 *IsoFunc* Webserver

A user-friendly Web server (http://guanlab.ccmb.med.umich.edu/isofunc) has been developed for the service of the scientific community. Different search options such as gene

symbol, Ensembl gene/transcript ID, and GO term ID and keywords (e.g., apoptosis) have been provided. We have also provided different filtering options for users to select PCSVs based on expression level, Ensembl biotypes, and availability in neXtProt database. PCSVs that are expressed in more than one-third of the total 127 samples with >1.0 FPKM are designated as highly expressed PCSVs. We used only Ensembl transcripts that were marked as protein_coding biotype; this biotype is further subdivided into three different categories: known, novel, and putative; therefore, we provide an option to select the particular subcategory of interest. The proteomics community prefers neXtProt over Ensembl because Ensembl transcript entries are considerably changed with different versions; neXtProt is highly curated and serves as the gold standard for the Human Proteome Project (www.thehpp.org) [42]. To enhance consistency, we provide an option to select only PCSVs that are present in neXtProt release 2015-09-01 [43]. These flexible search options will be useful for a variety of users to explore this web-resource.

All of the GO-based annotation results are displayed with columns of PCSVs sorted horizontally according to their maximum fold change values for any GO-term (rows). This fold change is a ratio of the rank probability of a particular PCSV to the base probability [15]. Higher fold change values have higher chances to perform the corresponding biological process function. Each gene/PCSV on the results page is linked to permit navigation of complementary useful resources (Ensembl, NCBI, UniProt, neXtProt, and GTEx). The GTEx portal is useful to explore relationships between genetic variation and gene/isoform expression in various human tissues [44]. The present *IsoFunc*Web server is also linked with our recently developed in-house tools *Hisonet* [45-47] and *MI-PVT* [48]. The *MI-PVT* is a tool for visualizing the chromosome-centric human proteome, while *Hisonet* displays predicted isoform-level functional networks.

Functionally connected PCSVs in *Hisonet* may be different from *IsoFunc* because *Hisonet* is based on the RefSeq database, while IsoFunc utilizes Ensembl and *Hisonet* used additional information such as pseudoamino acid composition, protein-docking, and protein domains with coexpression values to generate the functional networks. In the future, the *Hisonet* will be updated with Ensembl data to make it more compatible with *IsoFunc*.

4.5 Conclusions

These results demonstrate that our SVM-based algorithm combining RNA-seq expression data with Gene Ontology biological functions is capable of discriminating the functions of specific splice variants arising from genes generating multiple protein-coding splice variants (PCSVs). It is promising for deeper understanding of human gene functions considering the remarkable evolutionary emergence of multiple protein-coding splice variants from multiexonic genes in multicellular organizations. Experimental studies are needed to validate the predictions in this genome-wide resource and to characterize the functional dynamics of splice variants in different tissues and conditions. Integration of multiple omics data and multiple modeling methods will be useful for increasing the efficiency of the prediction tool. The data types that provide isoform-level information include but are not limited to, RNA-seq, exon array, protein domain information, post-transcriptional regulation, nucleotide or amino acid sequence variation/composition, and post-translational modification of protein. Large-scale integration of heterogeneous data may have the potential to improve prediction performance. The transcriptomic data provide mRNA-level expression, but biological functions are defined at the next level of protein expression, and this gap also affects the performance of functional annotation predictions. There is some inherent limitation of computational algorithms to

reconstruct transcript levels through transcriptomic data, and that is an open bioinformatics problem to solve.

4.6 Author Contributions

Experiments are conceived and designed by Bharat Panwar, Ridvan Eksi and Yuanfang Guan. Multiple instance learning framework source code is written by Ridvan Eksi. Experiments are performed by Bharat Panwar. The data is analyzed by Bharat Panwar, Rajasree Menon, Ridvan Eksi, Hong-Dong Li, Gilbert S. Omenn, and Yuanfang Guan. Paper is written by Bharat Panwar. Website is developed by Bharat Panwar.

4.7 Figures

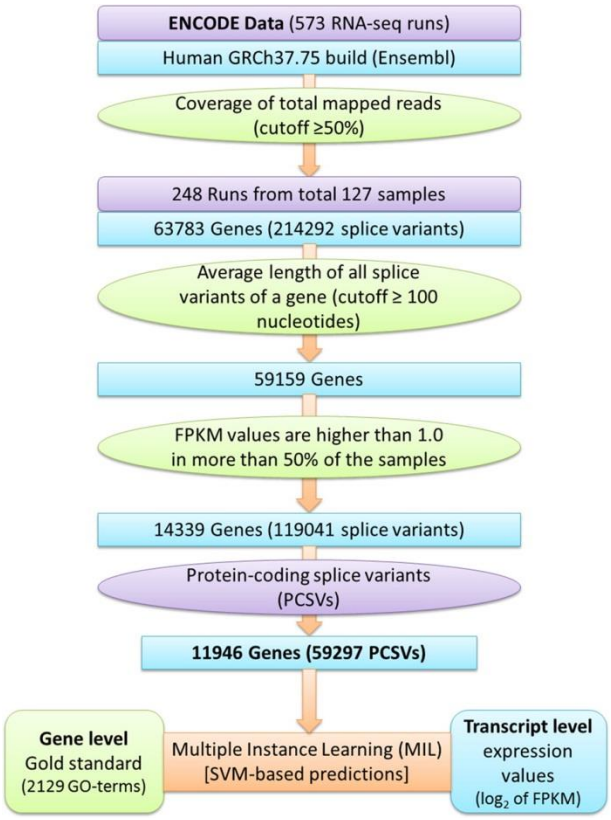


Figure 4-1 Overview of data preprocessing for predicting protein-coding splice variants.

We collected RNA-seq data from the ENCODE project and estimated the expression values using standard tools and thresholds. These values were used as input features for developing SVM models for different GO terms using the multiple instance learning approach.

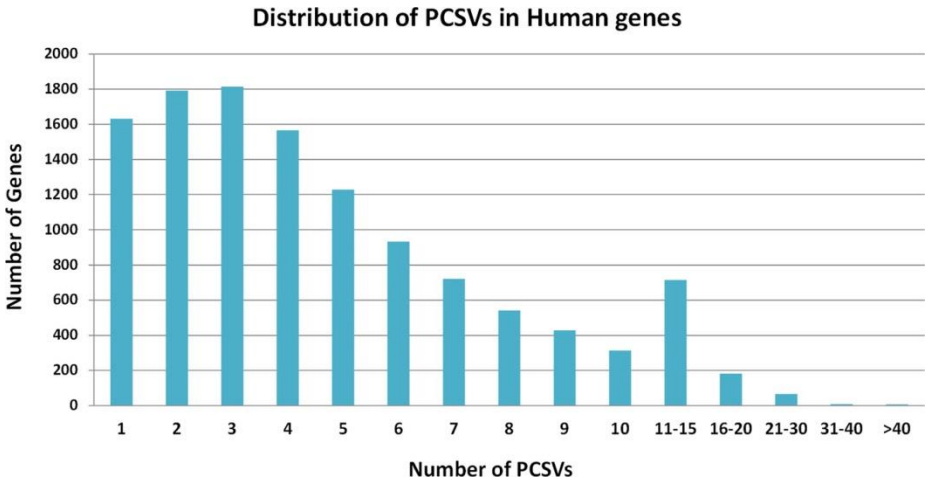


Figure 4-2 Distribution of number of protein-coding splice variants across well-expressed human genes.

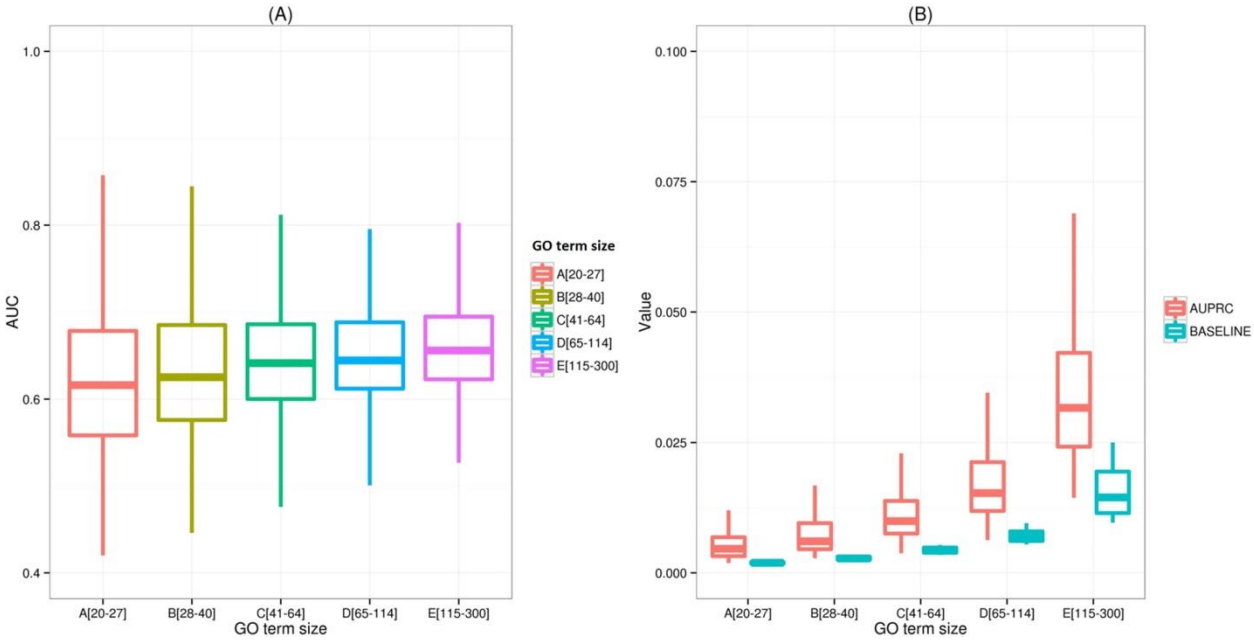


Figure 4-3 Performance of our multiple-instance learning based algorithm for predicting functions of protein-coding splice variants.

We used two different parameters (A) AUC and (B) AUPRC to evaluate prediction performance of the algorithm. The baseline values are also given with AUPRC. The different performances calculated for five different GO term sizes are shown in different colors.

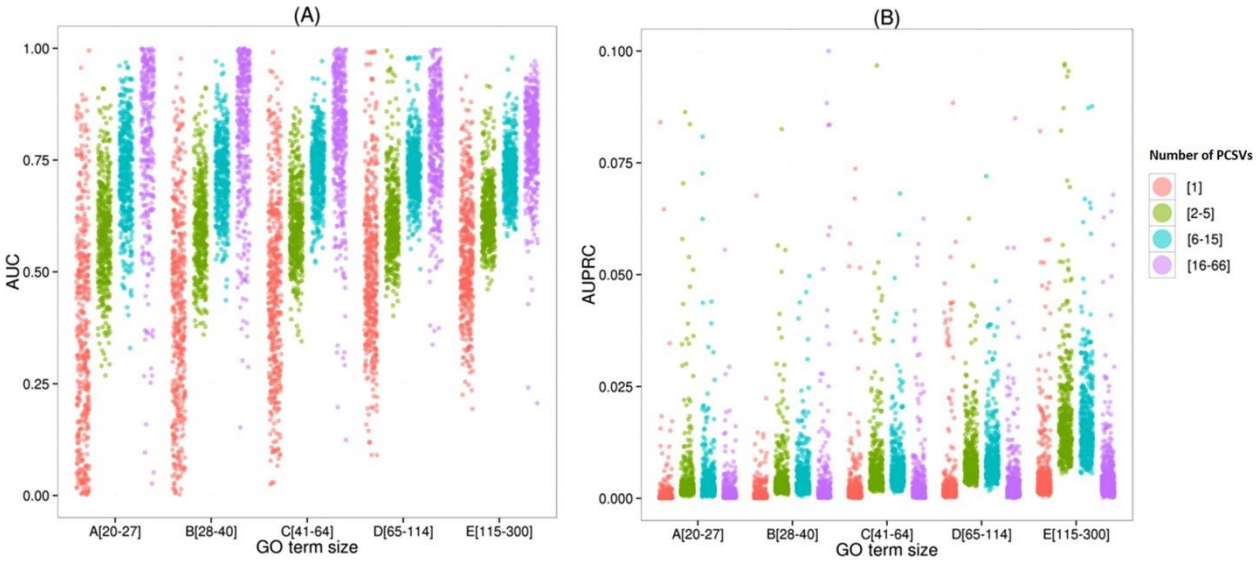


Figure 4-4 Comparative performance of single PCSV genes and multi-PCSVs genes.

Two different parameters (A) AUC and (B) AUPRC have been used to evaluate prediction performances. There are different performances calculated for five different GO term sizes as well as genes with a different number of PCSVs. Genes with single, 2–5, 6–15, and 16–66 PCSVs are shown in red, green, cyan, and magenta color, respectively. Each point shows the performance of a particular GO term in the defined size.

4.8 Bibliography

1. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.

2. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. 2008;4(1):44-57. doi: 10.1038/nprot.2008.211.

3. Murali TM, Wu C-J, Kasif S. The art of gene function prediction. Nature Biotechnology. 2006;24(12):1474-5. doi: 10.1038/nbt1206-1474.

4. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics. 2008;40(12):1413-5. doi: 10.1038/ng.259.

5. Black DL. Mechanisms of Alternative Pre-Messenger RNA Splicing. Annual Review of Biochemistry. 2003;72(1):291-336. doi: 10.1146/annurev.biochem.72.121801.161720.

6. de Klerk E, 't Hoen PAC. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. Trends in Genetics. 2015;31(3):128-39. doi: 10.1016/j.tig.2015.01.001.

7. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature. 2014;509(7502):575-81. doi: 10.1038/nature13302.

8. Revil T, Toutant J, Shkreta L, Garneau D, Cloutier P, Chabot B. Protein Kinase C-Dependent Control of Bcl-x Alternative Splicing. Molecular and Cellular Biology. 2007;27(24):8431-41. doi: 10.1128/mcb.00565-07.

9. López-Bigas N, Audit B, Ouzounis C, Parra G, Guigó R. Are splicing mutations the most frequent cause of hereditary disease? FEBS Letters. 2005;579(9):1900-3. doi: 10.1016/j.febslet.2005.02.047.

10. Skotheim RI, Nees M. Alternative splicing in cancer: Noise, functional, or systematic? The International Journal of Biochemistry & Cell Biology. 2007;39(7-8):1432-49. doi: 10.1016/j.biocel.2007.02.016.

11. He C, Zhou F, Zuo Z, Cheng H, Zhou R. A Global View of Cancer-Specific Transcript Variants by Subtractive Transcriptome-Wide Analysis. PLoS ONE. 2009;4(3):e4732. doi: 10.1371/journal.pone.0004732.

12. Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. Loss of exon identity is a common mechanism of human inherited disease. Genome Research. 2011;21(10):1563-71. doi: 10.1101/gr.118638.110.

13. Sterne-Weiler T, Sanford JR. Exon identity crisis: disease-causing mutations that disrupt the splicing code. Genome Biology. 2014;15(1):201. doi: 10.1186/gb4150.

14. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2014;347(6218):1254806-. doi: 10.1126/science.1254806.

15. Eksi R, Li H-D, Menon R, Wen Y, Omenn GS, Kretzler M, et al. Systematically Differentiating Functions for Alternatively Spliced Isoforms through Integrating RNA-seq Data. PLoS Computational Biology. 2013;9(11):e1003314. doi: 10.1371/journal.pcbi.1003314.

16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000;25(1):25-9. doi: 10.1038/75556.

17. Li H-D, Menon R, Omenn GS, Guan Y. The emerging era of genomic data integration for analyzing splice isoform function. Trends in Genetics. 2014;30(8):340-7. doi: 10.1016/j.tig.2014.05.005.

18. Bréhélin L, Florent I, Gascuel O, Maréchal É. Assessing functional annotation transfers with inter-species conserved coexpression: application to Plasmodium falciparum. BMC Genomics. 2010;11(1):35. doi: 10.1186/1471-2164-11-35.

19. Childs KL, Davidson RM, Buell CR. Gene Coexpression Network Analysis as a Source of Functional Annotation for Rice Genes. PLoS ONE. 2011;6(7):e22196. doi: 10.1371/journal.pone.0022196.

20. Piro RM, Ala U, Molineris I, Grassi E, Bracco C, Perego GP, et al. An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction. European Journal of Human Genetics. 2011;19(11):1173-80. doi: 10.1038/ejhg.2011.96.

21. Pavlidis P, Gillis J. Progress and challenges in the computational prediction of gene function using networks: 2012-2013 update. F1000Research. 2013;2.

22. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence. 1997;89(1-2):31-71. doi: 10.1016/s0004-3702(96)00034-3.

23. Li W, Kang S, Liu C-C, Zhang S, Shi Y, Liu Y, et al. High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. Nucleic Acids Research. 2013;42(6):e39-e. doi: 10.1093/nar/gkt1362.

24. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105-11. doi: 10.1093/bioinformatics/btp120.

25. Matsuo A, Bellier J-P, Hisano T, Aimi Y, Yasuhara O, Tooyama I, et al. Rat choline acetyltransferase of the peripheral type differs from that of the common type in intracellular translocation. Neurochemistry International. 2005;46(5):423-33. doi: 10.1016/j.neuint.2004.11.006.

26. Schneider E, El Hajj N, Richter S, Roche-Santiago J, Nanda I, Schempp W, et al. Widespread differences in cortex DNA methylation of the "language gene"CNTNAP2between humans and chimpanzees. Epigenetics. 2014;9(4):533-45. doi: 10.4161/epi.27689.

27. Fu WJ, Carroll RJ, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics. 2005;21(9):1979-86. doi: 10.1093/bioinformatics/bti294.

28. Andrews S, Tsochantaridis I, Hofmann T, editors. Support vector machines for multiple-instance learning. Advances in neural information processing systems; 2003.

29. Guan Y, Myers CL, Hess DC, Barutcuoglu Z, Caudy AA, Troyanskaya OG. Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biology. 2008;9(Suppl 1):S3. doi: 10.1186/gb-2008-9-s1-s3.

30. Panwar B, Arora A, Raghava GPS. Prediction and classification of ncRNAs using structural information. BMC Genomics. 2014;15(1):127. doi: 10.1186/1471-2164-15-127.

31. Panwar B, Gupta S, Raghava GPS. Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. BMC Bioinformatics. 2013;14(1):44. doi: 10.1186/1471-2105-14-44.

32. Panwar B, Raghava GPS. Prediction of uridine modifications in tRNA sequences. BMC Bioinformatics. 2014;15(1):326. doi: 10.1186/1471-2105-15-326.

33. Panwar B, Raghava GPS. Identification of protein-interacting nucleotides in a RNA sequence using composition profile of tri-nucleotides. Genomics. 2015;105(4):197-203. doi: 10.1016/j.ygeno.2015.01.005.

34. Zhong JL, Poghosyan Z, Pennington CJ, Scott X, Handsley MM, Warn A, et al. Distinct Functions of Natural ADAM-15 Cytoplasmic Domain Variants in Human Mammary Carcinoma. Molecular Cancer Research. 2008;6(3):383-94. doi: 10.1158/1541-7786.mcr-07-2028.

35. Charrier-Hisamuddin L, Laboisse CL, Merlin D. ADAM-15: a metalloprotease that mediates inflammation. The FASEB Journal. 2007;22(3):641-53. doi: 10.1096/fj.07-8876rev.

36. Chomarat P, Briolay J, Banchereau J, Miossec P. Increased production of soluble CD23 in rheumatoid arthritis, and its regulation by interleukin-4. Arthritis & Rheumatism. 1993;36(2):234-42. doi: 10.1002/art.1780360215.

37. Bonnefoy J-Y, Plater-Zyberk C, Lecoanet-Henchoz S, Gauchat J-F, Aubry J-P, Graber P. A new role for CD23 in inflammation. Immunology Today. 1996;17(9):418-20. doi: 10.1016/0167-5699(96)10054-2.

38. Lopez-Mejia IC, de Toledo M, Chavey C, Lapasset L, Cavelier P, Lopez-Herrera C, et al. Antagonistic functions of LMNA isoforms in energy expenditure and lifespan. EMBO reports. 2014;15(5):529-39. doi: 10.1002/embr.201338126.

39. Frangioni JV, Neel BG. Use of a general purpose mammalian expression vector for studying intracellular protein targeting: identification of critical residues in the nuclear lamin A/C nuclear localization signal. Journal of Cell Science. 1993;105(2):481-8.

40. Emerson LJ, Holt MR, Wheeler MA, Wehnert M, Parsons M, Ellis JA. Defects in cell spreading and ERK1/2 activation in fibroblasts with lamin A/C mutations. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2009;1792(8):810-21. doi: 10.1016/j.bbadis.2009.05.007.

41. Menon R, Panwar B, Eksi R, Kleer C, Guan Y, Omenn GS. Computational Inferences of the Functions of Alternative/Noncanonical Splice Isoforms Specific to HER2+/ER–/PR– Breast Cancers, a Chromosome 17 C-HPP Study. Journal of Proteome Research. 2015;14(9):3519-29. doi: 10.1021/acs.jproteome.5b00498.

42. Omenn GS, Lane L, Lundberg EK, Beavis RC, Nesvizhskii AI, Deutsch EW. Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. Journal of Proteome Research. 2015;14(9):3452-60. doi: 10.1021/acs.jproteome.5b00499.

43. Gaudet P, Michel P-A, Zahn-Zabal M, Cusin I, Duek PD, Evalet O, et al. The neXtProt knowledgebase on human proteins: current status. Nucleic Acids Research. 2015;43(D1):D764-D70. doi: 10.1093/nar/gku1178.

44. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. Nature genetics. 2013;45(6):580-5.

45. Li H-D, Menon R, Govindarajoo B, Panwar B, Zhang Y, Omenn GS, et al. Functional Networks of Highest-Connected Splice Isoforms: From The Chromosome 17 Human Proteome Project. Journal of Proteome Research. 2015;14(9):3484-91. doi: 10.1021/acs.jproteome.5b00494.

46. Zhu F, Panwar B, Guan Y. Algorithms for modeling global and context-specific functional relationship networks. Briefings in Bioinformatics. 2015;17(4):686-95. doi: 10.1093/bib/bbv065.

47. Li H-D, Menon R, Eksi R, Guerler A, Zhang Y, Omenn GS, et al. A Network of Splice Isoforms for the Mouse. Scientific Reports. 2016;6(1). doi: 10.1038/srep24507.

48. Panwar B, Menon R, Eksi R, Omenn GS, Guan Y. MI-PVT: A Tool for Visualizing the Chromosome-Centric Human Proteome. Journal of Proteome Research. 2015;14(9):3762-7. doi: 10.1021/acs.jproteome.5b00525.

CHAPTER 5

Conclusions and Future Work

5.1 Conclusions

A vast majority of multi-exon genes in humans undergo alternative splicing, and this can have significant functional consequences. Next-generation sequencing technology has greatly accelerated the study of alternative splicing events. However, there are limitations of using short read sequencing to study alternative splicing events. For example, reconstruction of transcripts from short-read RNA sequencing is not sufficiently accurate [1, 2]. Moreover, relatively few of the protein products of splice isoforms have been characterized functionally. In this dissertation, I outlined a way to use long read sequencing together with short read sequencing to reconstruct transcripts in a given sample. Moreover, I presented two studies, one for mouse and one for human, where we used machine learning tools and rich functional genomics data to predict functions for alternatively spliced isoforms.

In Chapter 2, we described our study where we investigated the overall transcriptome of a kidney using both long reads from PacBio platform and short reads from Illumina platform. We used short reads to validate full-length transcripts found by long PacBio reads, and generated two high quality sets of transcript isoforms that are expressed in glomerular and tubulointerstitial compartments. Then, the confirmed transcript isoforms are compared to the known set of known transcript isoforms in GENCODE and a final list of expressed transcript isoforms along with their annotation status is released to be used by researchers in downstream kidney transcriptome

studies. In this study, integrating data from two different sequencing technologies allowed us to identify and validate nearly 14k known transcripts in tubulointerstitial and glomerular compartments. Moreover, we identified and validated close to 12k and 8k multi-exon potential novel transcripts from each compartment respectively.

In Chapter 3, we stated that it is highly beneficial that the investigation of functions is carried out at the isoform level, because gene functions are delivered through alternatively spliced transcript isoforms that encode proteins of different functions. From this point of view, the standard gene function prediction paradigm has a major drawback in that it considers a gene as one single entity without differentiating its isoforms. The availability of transcript-level expression data from RNA-seq provides a rich resource for addressing this drawback. However, algorithmically, any supervised learning algorithm developed for gene function prediction cannot be directly applied to isoform function prediction because of the lack of isoform-level, 'ground-truth' functional annotations. To address this challenge, we developed an iterative algorithm that predicts functions at the individual isoform level by conceptualizing a gene as a 'bag' of isoforms of potentially different functions. Our key idea was to iteratively extract the common pattern of a subset of isoforms across the positive genes of the function under investigation, aiming at maximizing the coherence within this subset of isoforms and the discriminative power against the other 'negative' genes (genes not related to the specific function under consideration). Through experimental validation, we demonstrated that our approach in combination with publicly available RNA-seq data is capable of differentiating isoform functions, promising better and deeper understanding of gene functions.

In Chapter 4, we described the function prediction study for human protein coding isoforms. We used a multiple instance learning based approach for predicting the function of

protein coding splice variants. Our results demonstrated that our SVM-based algorithm combining RNA-seq expression data with Gene Ontology biological function annotations is capable of discriminating the functions of specific splice variants. This is promising for deeper understanding of human gene functions considering the remarkable emergence of multiple protein-coding splice variants from multi-exonic genes in multicellular organizations. Experimental studies are needed to validate the predictions in this genome-wide resource and to characterize the functional dynamics of splice variants in different tissues and conditions.

5.2 Future directions

In Chapter 2, we explained how we used short reads in conjunction with long reads to infer full structure of transcripts. Using short reads for validation of splice junctions allowed us to eliminate PacBio transcripts with erroneous junctions. On the other hand, none of the sequencing technology is able to reliably and accurately locate transcription start sites for transcripts. PacBio tend to produce transcript with 5'ends truncated. Two possible ways have been proposed to mitigate this problem. Cartolano et al. proposed the implementation of the TeloPrime Full Length cDNA Amplification kit to the Pacific Biosciences Iso-Seq technology in order to enrich for genuine full-length transcripts in the cDNA libraries [3]. In another study, deepCAGE (Cap Analysis of Gene Expression) data is used as a mean to correct for exact 5' end coordinate of transcripts with success [4]. In this study, we imputed the ambiguous start positions by assuming the 5′-most Iso-Seq finding as the true 5′ end which may still result in a shorter than actual transcript. Incorporating either of these methods to ours would improve our ability to reconstruct transcription start sites accurately.

There are many ways to expand upon our existing method for function prediction of alternatively spliced isoforms. First, it is possible to extend the functional genomics dataset used

in our studies. RNA-seq data, which was the only type we used, are the richest resource for genome-wide, isoform-level data so far. But the basic concept is extendable to other large-scale datasets providing isoform-level information such as protein domain data, protein docking scores, exon array, amino acid sequence variation, post-transcriptional regulation, and post-translational regulation datasets. Furthermore, our studies only focused on the base learner SVM. However, our approach is highly extendable to other modeling methods, such as logistical regression, random forests and deep learning methods.

Our function prediction studies are limited to the incomplete isoform catalog maintained by NCBI and GENCODE, but it can be readily updated whenever the genome annotation of isoforms is updated. Additionally, alternatively spliced isoforms often show tissue-specific expression and functions [5-10]. Our generic algorithm does not yet take the tissue-specific functionality into consideration. We expect that more accurate and biologically meaningful isoform function prediction could be achieved if tissue specificity were taken into account. Our method is further limited by the accuracy of current methods that assign reads to individual isoforms and estimate isoform-level expression values. If more advanced algorithms are developed, our algorithm could directly utilize the estimates from those algorithms and generate isoform function predictions.

5.3 Bibliography

1. Tilgner H, Raha D, Habegger L, Mohiuddin M, Gerstein M, Snyder M. Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. G3 (Bethesda). 2013;3(3):387-97. doi: 10.1534/g3.112.004812. PubMed PMID: 23450794; PubMed Central PMCID: PMCPMC3583448.

2. Steijger T, Abril JF, Engstrom PG, Kokocinski F, Consortium R, Hubbard TJ, et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013;10(12):1177-84. doi: 10.1038/nmeth.2714. PubMed PMID: 24185837; PubMed Central PMCID: PMCPMC3851240.

3. Cartolano M, Huettel B, Hartwig B, Reinhardt R, Schneeberger K. cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing. PLoS One. 2016;11(6):e0157779. doi: 10.1371/journal.pone.0157779. PubMed PMID: 27327613; PubMed Central PMCID: PMCPMC4915659.

4. O'Grady T, Wang X, Honer Zu Bentrup K, Baddoo M, Concha M, Flemington EK. Global transcript structure resolution of high gene density genomes through multi-platform data integration. Nucleic Acids Res. 2016;44(18):e145. doi: 10.1093/nar/gkw629. PubMed PMID: 27407110; PubMed Central PMCID: PMCPMC5062983.

5. Fruhwald J, Camacho Londono J, Dembla S, Mannebach S, Lis A, Drews A, et al. Alternative splicing of a protein domain indispensable for function of transient receptor potential melastatin 3 (TRPM3) ion channels. J Biol Chem. 2012;287(44):36663-72. doi: 10.1074/jbc.M112.396663. PubMed PMID: 22961981; PubMed Central PMCID: PMCPMC3481270.

6. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621-8. doi: 10.1038/nmeth.1226. PubMed PMID: 18516045.

7. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. Nature. 2010;465(7294):53-9. doi: 10.1038/nature09000. PubMed PMID: 20445623.

8. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, et al. Global identification of human transcribed sequences with genome tiling arrays. Science. 2004;306(5705):2242-6. doi: 10.1126/science.1103388. PubMed PMID: 15539566.

9. Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem. 2003;72:291-336. doi: 10.1146/annurev.biochem.72.121801.161720. PubMed PMID: 12626338.

10. Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, et al. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell. 2004;16(6):929-41. doi: 10.1016/j.molcel.2004.12.004. PubMed PMID: 15610736.