Published in final edited form as:

Trends Genet. 2014 August ; 30(8): 340–347. doi:10.1016/j.tig.2014.05.005.

The Emerging Era of Genomic Data Integration for Analyzing Splice Isoform Function

Hong-Dong Li¹, Rajasree Menon¹, Gilbert S. Omenn^{1,2}, and Yuanfang Guan^{1,2,3,*}

¹Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, United States

²Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, United States

³Department of Electrical Engineering and Computer Science, Ann Arbor, Michigan, United States

Abstract

The vast majority of multi-exon genes in humans undergo alternative splicing, which greatly increases the functional diversity of protein species. Predicting functions at the isoform level is essential to further our understanding of developmental abnormalities and cancers, which frequently exhibit aberrant splicing and dysregulation of isoform expression. However, determination of isoform function is very difficult, and efforts to predict isoform function have been limited in the functional genomics field. Now deep sequencing of RNA provides an unprecedented amount of expression data at the transcript level. We describe here emerging computational approaches that integrate such large-scale RNA-seq data for predicting functions of alternatively spliced isoforms, and we discuss their applications in developmental and cancer biology. We outline future directions for isoform function prediction, emphasizing the need for heterogeneous genomic data integration and tissue-specific, dynamic isoform-level network modeling, which will allow the field to realize its full potential.

Keywords

Splice isoforms; genomic data integration; function prediction; cancers; development

From one to many: alternative splicing and gene function

Alternative splicing is a process that allows multi-exon genes to produce multiple proteins, referred to as alternatively spliced isoforms, through mechanisms such as exon skipping, alternative 5' donor sites, and intron retention [1, 2] (see Glossary and Figure 1A). This process is tightly regulated by both trans- and cis- factors, as well as by the secondary

© 2014 Elsevier Ltd. All rights reserved.

*To whom correspondence should be addressed: Yuanfang Guan, Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, United States. Tel.: 734-764-0018, gyuanfan@umich.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

structure of the pre-mRNA transcript, which together form the ‘splicing code’ that determines cell typespecific splicing [2]. High-throughput analyses of transcript-level data have been developed to detect alternative splicing. For example, cDNA libraries have revealed breast cancer-associated alternatively spliced transcripts [3]. Expressed sequence tags or RNA-seq data are also informative sources for identifying splicing events [4, 5]. Alternative splicing may have significant functional consequences as it can modify protein sequence/domains [6] and protein-protein interactions [7]. Functional consequences of alternative splicing also include alterations in mRNA decay [8] and the translation process [9]. In humans, an estimated 95% of multi-exon genes undergo alternative splicing [10]. According to the most recent GENCODE human annotation (version 19) [11, 12], 196,520 protein-coding and non-coding transcripts are encoded by 57,820 genes (Figure 1B). This corresponds to approximately three transcripts encoded by a single gene on average. The maximal number of transcripts that a human gene expresses is 82 according to GENCODE (Figure 1B). The protein diversity of mammalian genomes is therefore increased significantly through alternative splicing.

Isoforms resulting from alternative splicing may carry out different or even opposite biological functions [13–21]. For example, alternative splicing of the B-cell lymphoma-x gene (*Bcl-x*) results in two splice variants, anti-apoptotic *Bcl-x_L* and pro-apoptotic *Bcl-x_S* [21]. The balance between the two variants can be modulated by staurosporine, a protein kinase C inhibitor and apoptotic inducer [21]. *Bcl-x_S* is missing two Bcl-2 family motifs, BH1 and BH2, which accounts for the functional difference. Another interesting example is the carboxypeptidase E (*CPE*) gene in the Wnt signal transduction pathway [22]. This pathway is involved in many diseases including cancers and particularly colorectal cancer. It has been shown that the splice variant form of CPE activates the Wnt signal, whereas the full length canonical CPE is an inhibitor of Wnt/B-catenin signaling [22].

Because of the potentially divergent functional consequences, alternative splicing is often tightly regulated in different tissues and developmental processes in order to maintain a healthy physiological state and ensure the correct development of cell lineages or tissues [23, 24], especially in neurons and stem cells [23, 25–32]. In embryonic development, it has been shown that alternative splicing regulation is a common occurrence across tissues and developmental stages based on a genome-wide analysis of isoform expression of the C57BL/6 mouse at three-day points in embryos and the placenta [33]. A specific example is the human forkhead box p1 (*FOXP1*) gene, whose embryonic stem isoform (FOXP1-ES) differs from the FOXP1 isoform in terms of its DNA-binding specificity [27]. During neuronal development, a large number of alternative splicing events have been shown to occur [34]. *Scn9a*, a sodium channel gene, is alternatively spliced in the mouse brain, with one isoform containing exon 17a, which increases *Scn9a* expression, and another isoform lacking this exon, which is targeted for nonsense-mediated degradation [23].

Aberrant splicing events can lead to disease by creating mutant isoforms with disruptive structures or altering isoform expression levels [35–38]. p53 β , an isoform of the human Tumor Protein p53 (*TP53*) gene, can enhance p53 target gene expression and is apoptotic. By contrast, the isoform 133p53 inhibits p53-mediated apoptosis [39]. Splicing at exon 11 of *Brcal* has been associated with breast cancer; a balanced ratio of *Brcal* isoforms was

suggested to maintain normal physiological status [40]. The *CD99* gene pre-mRNA can be alternatively spliced into two distinct isoforms that display opposite functions in tumor malignancy and metastases [41]. By means of forced over-expression, the longer isoform (CD99wt), which has additional domains, was found to inhibit anchorage-independent growth and anoikis resistance, which are instead favored by the shorter variant (CD99sh). For the *CASP3* gene, its *CASP3-L* variant is pro-apoptotic while *CASP3-S* is anti-apoptotic, and overexpression of *CASP3-S* is associated with poor chemotherapy response in breast carcinoma [42]. *NEK2*, which encodes a serine/threonine kinase, was proposed to be involved in the development of breast cancers. In a recent study [43], the expression of its isoform *NEK2C* was significantly upregulated in several breast cancer cell lines as well as in human primary breast cancer tissues.

These observations imply an association between isoform structure/expression and diseases such as cancers, which would be better understood if the differences in isoform functions could be accurately predicted. It is likely, however, that these specific examples represent only a small fraction of the isoforms with distinct functions. The wide availability of genome-wide sequencing data offers the promise of discovering more of these functional important splice variants, but finding them has remained challenging.

Genomic data integration at the gene-level

Traditionally, functional annotation and genomic data analyses have been centered on genes. Functions are frequently studied through knock-out or knock-down experiments. The resulting annotations resolved at the gene level are assembled and recorded in databases such as Gene Ontology (GO) [44] and KEGG pathways [45]. These databases provide a central resource for recording canonical gene functions under generic conditions. Additionally, functional genomic data are traditionally processed at the gene level. For example, when analyzing a microarray dataset, the expression levels are usually assigned to each gene; for a physical interaction dataset, most often it is gene-gene interactions that are recorded [46].

In the past decades, the exponential growth of functional genomic data has stimulated the development of a wealth of computational methods to integrate these data and predict functions at the gene level, including but not limited to Bayesian network-based methods [47], kernel-based classifiers [48], decision trees [49], and the weighted power method. These studies have benefited from widespread expertise and use in the field, solidly establishing the field of gene function prediction.

Another important direction of functional genomic data integration is to build networks that describe the interactions between genes, and then to utilize these networks to prioritize phenotype or disease-associated genes. Such networks include but are not limited to functional relationship networks [50–55], which represent the co-functionality between genes, protein-protein physical interaction networks [56–60], and protein-DNA interaction networks [61, 62]. Analyses of these networks have greatly accelerated the discovery of disease-associated genes [63–65].

The computational approaches described above, however, typically serve as a starting point for experimentally validating the predicted functions or interactions. Unlike traditional gene-level functional studies, for isoform function, prediction is often the only route because experimental identification of isoform functions is difficult. For example, isoform-specific antibodies are generally not available, making experiments such as Western blot and ChIP-chip not routinely resolved at the isoform level. In addition, isoform-specific oligonucleotides are difficult to design for experiments such as qPCR or RNA interference, partly because many isoforms differ by only a small amount, such as a single codon. These technical challenges also mean that the current state of gene function annotation is potentially misleading, as the annotation data are not differentiated for isoforms in the database. Similarly, predictions made at the gene level assign the same function or interaction to different isoforms of the same gene. These limitations have previously been addressed by small-scale investigations focusing on individual isoforms [20, 21, 66], but systematic, genome-scale predictive studies for isoform functions based on functional genomic data has been limited until recently [16, 67, 68]. Here, we discuss emerging methods for predicting isoform functions, highlight recent computational advances, and explore how such predictions can be useful in understanding development and diseases.

Emerging computational approaches for predicting isoform functions

Prior to the recent accumulation of RNA-seq data, protein structure methods [69–75], based on the assumption that structural differences may predispose isoforms to functional diversification, had been the main route for predicting isoform functions. These methods are based on amino acid sequences, protein domains, and 3D conformational structures (Figure 2). One group [18] sought to identify functional differences between isoforms by comparing the 3D conformational structures built using homology modeling [74]. They identified biologically meaningful motifs that may underline the functional differences between isoforms for genes such as *Anxa6* (annexin 6), *Ptbp1* (polypyrimidine tract binding protein 1), *Tax1bp3* (tax1 binding protein 3), *Calu* (Calumenin) and *Cdc42* (cell division cycle 42), whose isoforms were found to be differentially expressed in mouse Her/neu-induced cancers. Because domains are relatively independent functional units, several methods have been developed to utilize domain information to predict functions. For example, two studies correlated domains to protein functions and investigated isoform functions [69, 70]. Another method was developed for predicting functions of uncharacterized proteins by using DNA-binding domains [75]. These types of structure-based predictions have dominated the field of isoform function differentiation and prediction.

Such methods, however, have several potential limitations. Firstly, many structure-based prediction methods rely on threading an unknown protein structure onto established protein structure databases, such as the protein data bank (PDB) [76]. If similar protein structures are not available, then the functions for these isoforms cannot be predicted. Secondly, the majority of the isoforms differ in subtle ways not involving domains. For example, in the mouse, only 34% of the isoform pairs of the same gene differ in domains [16], and the functions of the rest of the isoforms are unlikely to be differentiated solely through structure-based methods. These limitations call for new functional genomics approaches that enable whole-genome prediction of isoform functions.

Recent accumulation of public, large-scale RNA-seq data provides an unprecedented opportunity to integrate these data in the prediction of isoform functions. As of February 2014, more than 10,000 and 8,000 RNA-seq experimental datasets are available for human and mouse, respectively, in the Sequence Read Archive (SRA) of NCBI. These RNA-seq data, like previous microarray data, cover heterogeneous tissues and cell lineages; in contrast to microarray data, they may provide transcriptomic expression at the isoform level, through analysis with software such as TopHat [77], Bowtie [78], BWA [79], Cufflinks [80], and eXpress [81].

Despite the availability of these data sets, however, predicting functions using RNA-seq data faces a major technical challenge: unlike genes, isoforms do not have existing functional annotations. In databases such as GO and KEGG, biological functions are recorded at the gene level. To predict whether a gene is related to a specific function, machine learning algorithms are used to derive models from the genomic data collected for the genes known to be associated with the function under investigation. Such an approach is termed supervised learning, *i.e.*, learning models using existing observations and existing knowledge. However, isoform-level functional annotations are not available, which prevents the use of conventional supervised learning methods to build a classification model.

Recently, a novel strategy was developed to systematically predict functions of alternatively spliced isoforms by integrating RNA-seq data encompassing many conditions and cell types [16], utilizing multiple instance learning (MIL) algorithms (Figure 2). Essentially, MIL is used to build a model for predicting functions for isoforms by taking RNA-seq isoform expression data and a gene-level gold standard including positive and negative genes as its input. A positive gene (termed as a ‘bag’ in MIL) is defined to carry out a particular function under investigation, as long as at least one of its isoforms (‘instances’) carries out this function. By contrast, a negative gene is considered irrelevant to a function, and none of its isoforms should carry out this function [16]. For example, for the biological process mitochondria biogenesis, we have a set of positive genes annotated with this function based on GO and another set of negative genes that do not have this function. For each positive gene, only some of its isoforms are related to mitochondria biogenesis. MIL then tries to identify a subset of isoform(s) (‘witnesses’) from positive genes that most likely carry out the function under study, and then build a model whose accuracy in predicting isoform-level functions is expected to be maximized. An advantage of this approach is that it covers all isoforms that are captured in the RNA-seq datasets, compared to structure-based methods that only cover those that have clear domain information. Furthermore, any classification method can be built into this MIL framework, such as support vector machines [82], Bayesian networks [83], and K-nearest neighbors [84]. This approach was recently used to generate genome-wide predictions of isoform functions and identified functional differences for a large number of multi-isoform mouse genes [16]. The results were validated through literature searches, proteomic experiments, and 3D-structure modeling. However, the evaluation was mainly limited to genes with validated disparate isoform functions. Subsequently, a similar method was used to annotate isoform functions in humans by integrating RNA-seq data into co-expression networks and propagating isoform-level labels based on this multiple instance concept [67]. In this method, the authors defined the

isoform-level function prediction problem in a similar approach as [16], in which genes are defined as ‘bags’ containing isoforms (‘instances’) of potentially different functions. The authors assessed the accuracies of the models by cross-validating single-isoform genes, which may potentially limit estimates of performance because multi-isoform genes were not included in evaluation. The success of these approaches can be illustrated by looking at the five isoforms of *TP53* as an example, for which the known apoptosis regulation function as well as the regulation direction (positive versus negative) were correctly predicted.

Software developed based on the MIL method [16] has been shown to reveal functions related to developmental processes and diseases (Box 1 and Box 2). One example is the cyclin-dependent kinase inhibitor 2A (*Cdkn2a*) gene, a well-known tumor suppressor; it inhibits the oncogenic action of MDM2 by blocking MDM2-induced degradation of p53 and enhancing p53-dependent transactivation and apoptosis [85]. *Cdkn2a* encodes two isoforms with highly different computationally predicted 3D conformations (Figure 3). One was predicted to function in apoptotic nuclear changes [16]. The other was an inhibitor of CDK4 kinase, which is a Ser/Thr kinase family member, and is therefore likely to participate in regulation of protein serine/threonine kinase activity [16]. Another example is the splice isoforms of *Anxa6*, which are differentially regulated in the mouse Her2/neu mammary tumor [18]. The MIL method correctly predicted the differential phosphorylation activities of the isoforms of *Anxa6* [16] (Box 2, Figure 3). These examples, though limited in scope, demonstrate the potential application of MIL in identifying disease-associated isoforms.

Perspectives of isoform-level genomic data integration

The rapid growth of RNA-seq data opens the door for isoform-level genomic data integration and function prediction. This field has several exciting avenues for future development (Figure 2). First, the current integration methods analyze only RNA-seq data. It would be of great interest to integrate heterogeneous genomic data types that provide isoform-level information, including but not limited to RNA-seq, exon array, post-transcriptional regulation, amino acid sequence variation and post-translational modifications of proteins. In such heterogeneous data analyses, the predictive ability (*e.g.* accuracy, sensitivity and specificity) of different types of data needs to be assessed. The most distinguishing feature of the data that underlie different functions of isoforms can then be identified. This would provide information about which kind of data are most predictive and might further help to build a parsimonious and robust prediction model by excluding uninformative features of the data. Additionally, heterogeneous data may provide complementary information, and therefore integration of these data sets would likely improve the accuracy of isoform function predictions. Algorithmically, one limitation of MIL is that it has not theoretically been proven to converge. Thus, effectively identifying all isoforms that carry out the function of positive genes remains a challenging problem facing MIL. Novel well-principled algorithms are in need to integrate multiple types of datasets for isoform function prediction. Since genomic data are usually very large, scalability is an important factor to consider when selecting or designing these algorithms. Overall, such heterogeneous, large-scale data integration may have the potential to significantly improve the performance of the predictions.

Second, genomic data integration at the isoform level can be applied to many of the existing sub-fields of functional genomics. It is worthwhile to explore approaches for integrative network analysis at the isoform level, such as predicting protein physical interaction networks, functional relationship networks, gene-protein interactions and protein-metabolite interactions. Isoform-level data integration methods can also be explored to predict the different properties, cellular locations, and phenotypic relevance of the isoforms. In addition to help prioritizing isoforms of interest, function prediction could also provide a prospective approach to guide the investigation of different classes of transcripts such as noncoding RNAs, which are considered functional elements though their functions are largely unknown [86].

Third, alternative splicing is tightly regulated in different conditions such as different tissues, cell lineages and developmental stages. Recently emerging prediction methods [16] are useful for probing isoform functions, but need next to take temporal and spatial information into account for accurately inferring the functional role or interactions of isoforms. To this end, it would be interesting to analyze multiple datasets generated under different developmental or disease conditions. This could also help reveal novel isoforms that are expressed in specific conditions, and possible functions of these novel isoforms could be explored through prediction methods. Building and comparing condition-specific models would in return advance the development of computational methods as needed. Such methods would promise a refined annotation of gene functions at the isoform level and thus allow for better understanding of developmental biology and diseases.

Concluding remarks

It is widely accepted that splice isoforms of the same gene may show different and even opposite functions, and aberrant alternative splicing is closely linked to developmental disorders and cancers. This poses a challenge to the traditional functional genomic approaches that focus on the gene level, and indicates the necessity to investigate the functions of individual isoforms. Before the accumulation of large amounts of public RNA-seq data, efforts in probing isoform functions at the genome-scale were very limited.

Recent studies have demonstrated that prediction of isoform functions can be achieved through multiple instance learning-based methods by integrating genomic data from heterogeneous conditions [16, 67]. While the prediction was made at the genome-scale, it should be noticed that the number of isoforms of which the functions can be validated with experimental or literature evidence is very limited, mainly due to the lack of functional data at the isoform level. Genes with well-known differentiated functions of isoforms have been used to benchmark these algorithms. Yet the global accuracy of isoform-level function prediction is difficult to assess, because current computational validation are only made on the gene-level [20, 87]. The building and validation of a model is still very challenging due to many factors including the heterogeneity of isoform-level feature data, lack of systematically recorded isoform function, the dynamic nature of isoform expression across cell lineages and the interactions with other types of biological molecules, *etc.*

Despite these limitations, we foresee that data integration at the isoform level will soon add a new dimension to the functional genomics field by providing a high-resolution map of gene functions, which would move us towards a refined understanding of many developmental processes and diseases such as cancer. Surely, many challenges remain for this field to reach its full potential, including network modeling, heterogeneous data integration, and tissue/cell lineage, developmental stage-specific modeling. It is clear, however, that functional genomic data integration at the splice isoform level is now shifting the field away from traditional gene-centered studies.

Acknowledgments

This work is supported by National Institutes of Health [1R21NS082212-01] and NIH [University of Michigan O'Brien Kidney Translational Core Center].

References

1. Black DL. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem.* 2007; 72:291–336. [PubMed: 12626338]
2. Barash Y, et al. Deciphering the splicing code. *Nature.* 2010; 465:53–59. [PubMed: 20445623]
3. Ferreira E, et al. Alternative splicing enriched cDNA libraries identify breast cancer-associated transcripts. *BMC Genomics.* 2010; 11:S4. [PubMed: 21210970]
4. Modrek B, et al. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acid Res.* 2001; 29:2850–2859. [PubMed: 11433032]
5. Griffith M, et al. Alternative expression analysis by RNA sequencing. *Nat Methods.* 2010; 7:843–847. [PubMed: 20835245]
6. Light S, Elofsson A. The impact of splicing on protein domain architecture. *Curr. Opin. Struct. Biol.* 2013; 23:451–458. [PubMed: 23562110]
7. Ellis, Jonathan D., et al. Tissue-specific alternative splicing remodels proteinprotein interaction networks. *Mol cell.* 2012; 46:884–892. [PubMed: 22749401]
8. Lewis BP, et al. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci USA.* 2003; 100:189–192. [PubMed: 12502788]
9. Sanford JR, et al. A novel role for shuttling SR proteins in mRNA translation. *Genes Dev.* 2004; 18:755–768. [PubMed: 15082528]
10. Pan Q, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet.* 2008; 40:1413–1415. [PubMed: 18978789]
11. Harrow J, et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* 2012; 22:1760–1774. [PubMed: 22955987]
12. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science.* 2004; 306:636–640. [PubMed: 15499007]
13. Wang ET, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature.* 2008; 456:470–476. [PubMed: 18978772]
14. Stamm S, et al. Function of alternative splicing. *Gene.* 2005; 344:1–20. [PubMed: 15656968]
15. Tress ML, et al. The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA.* 2007; 104:5495–5500. [PubMed: 17372197]
16. Eksi R, et al. Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data. *PloS Comput Biol.* 2013; 9:e1003314. [PubMed: 24244129]
17. Omenn GS, et al. Alternative splice variants, a new class of protein cancer biomarker candidates: findings in pancreatic cancer and breast cancer with systems biology implications. *Dis Markers.* 2010; 28:241–251. [PubMed: 20534909]

18. Menon R, et al. Functional implications of structural predictions for alternative splice proteins expressed in Her2/neu-induced breast cancers. *J Proteome Res.* 2011; 10:5503–5511. [PubMed: 22003824]
19. Menon R, et al. Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer. *Cancer Res.* 2009; 69:300–309. [PubMed: 19118015]
20. Omenn GS, et al. 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. [PubMed: 23603631]
21. Revil T, et al. Protein Kinase C-Dependent Control of Bcl-x Alternative Splicing. *Mol Cell Biol.* 2007; 27:8431–8441. [PubMed: 17923691]
22. Skalka N, et al. Carboxypeptidase E: a negative regulator of the canonical Wnt signaling pathway. *Oncogene.* 2013; 32:2836–2847. [PubMed: 22824791]
23. Zheng S, Black DL. Alternative pre-mRNA splicing in neurons: growing up and extending its reach. *Trends Genet.* 2013; 29:442–448. [PubMed: 23648015]
24. Yano M, et al. Nova2 Regulates Neuronal Migration through an RNA Switch in Disabled-1 Signaling. *Neuron.* 2010; 66:848–858. [PubMed: 20620871]
25. Ule J, et al. Nova regulates brain-specific splicing to shape the synapse. *Nat Genet.* 2005; 37:844–852. [PubMed: 16041372]
26. Li Q, et al. Neuronal regulation of alternative pre-mRNA splicing. *Nat Rev Neurosci.* 2007; 8:819–831. [PubMed: 17895907]
27. Gabut M, et al. An Alternative Splicing Switch Regulates Embryonic Stem Cell Pluripotency and Reprogramming. *Cell.* 2011; 147:132–146. [PubMed: 21924763]
28. Baumann K. Stem cells: Tailored splicing patterns. *Nat Rev Mol Cell Biol.* 2013; 14:464–465. [PubMed: 23820742]
29. Aaronson Y, Meshorer E. Stem cells: Regulation by alternative splicing. *Nature.* 2013; 498:176–177. [PubMed: 23739330]
30. Han H, et al. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature.* 2013; 498:241–245. [PubMed: 23739326]
31. Salomonis N, et al. Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation. *Proc. Natl Acad. Sci. USA.* 2010; 107:10514–10519. [PubMed: 20498046]
32. Suzuki H, et al. Comprehensive Analysis of Alternative Splicing and Functionality in Neuronal Differentiation of P19 Cells. *PLoS ONE.* 2011; 6:e16880. [PubMed: 21365003]
33. Revil T, et al. Alternative splicing is frequent during early embryonic development in mouse. *BMC Genomics.* 2010; 11:399. [PubMed: 20573213]
34. Boutz PL, et al. A post-transcriptional regulatory switch in polypyrimidine tractbinding proteins reprograms alternative splicing in developing neurons. *Genes Dev.* 2007; 21:1636–1652. [PubMed: 17606642]
35. Srebrow A, Kornblihtt AR. The connection between splicing and cancer. *J Cell Sci.* 2006; 119:2635–2641. [PubMed: 16787944]
36. Wang G-S, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet.* 2007; 8:749–761. [PubMed: 17726481]
37. 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; 70:3440–3449. [PubMed: 20388783]
38. Fackenthal J, Godley LA. Aberrant RNA splicing and its functional consequences in cancer cells. *Dis Model Mech.* 2008; 1:1754–8411.
39. Bourdon J-C, et al. p53 isoforms can regulate p53 transcriptional activity. *Genes Dev.* 2005; 19:2122–2137. [PubMed: 16131611]
40. Tammaro C, et al. BRCA1 exon 11 alternative splicing, multiple functions and the association with cancer. *Biochem. Soc. Trans.* 2012; 40:768–772. [PubMed: 22817731]
41. Scotlandi K, et al. CD99 isoforms dictate opposite functions in tumour malignancy and metastases by activating or repressing c-Src kinase activity. *Oncogene.* 2007; 26:6604–6618. [PubMed: 17471235]

42. Veggran F, et al. Overexpression of Caspase-3s Splice Variant in Locally Advanced Breast Carcinoma Is Associated with Poor Response to Neoadjuvant Chemotherapy. *Clinical Cancer Research*. 2006; 12:5794–5800. [PubMed: 17020986]
43. Liu Z, et al. Nek2C functions as a tumor promoter in human breast tumorigenesis. *Int J Mol Med*. 2012; 30:775–782. [PubMed: 22824957]
44. Ashburner M, et al. Gene Ontology: tool for the unification of biology. *Nature Genet*. 2000; 25:25–29. [PubMed: 10802651]
45. Kanehisa M, et al. The KEGG resource for deciphering the genome. *Nucleic Acids Res*. 2004; 32:D277–D280. [PubMed: 14681412]
46. Subramanian A, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005; 102:15545–15550. [PubMed: 16199517]
47. Troyanskaya OG, et al. A Bayesian framework for combining heterogeneous data source for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl Acad. Sci USA*. 2003; 100:8348–8353. [PubMed: 12826619]
48. Guan Y, et al. Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome Biol*. 2008; 9(Suppl 1):S3. [PubMed: 18613947]
49. Schietgat L, et al. Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics*. 2010; 11:2. [PubMed: 20044933]
50. Guan Y, et al. A genomewide functional network for the laboratory mouse. *PLoS Comput Biol*. 2008; 4:e1000165. [PubMed: 18818725]
51. Franke L, et al. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*. 2006; 78:1011–1025. [PubMed: 16685651]
52. Lee I, et al. A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat Genet*. 2008; 40:181–188. [PubMed: 18223650]
53. Lee I, et al. An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*. *PLoS ONE*. 2007; 2:e988. [PubMed: 17912365]
54. Guan Y, et al. Tissue-specific functional networks for prioritizing phenotypes and disease genes. *PLoS Comput Biol*. 2012; 8:e1002694. [PubMed: 23028291]
55. Guan Y, et al. Functional genomics complements quantitative genetics in identifying disease-gene associations. *PLoS Comput Biol*. 2010; 6:e1000991. [PubMed: 21085640]
56. Stelzl U. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*. 2005; 122:957–968. [PubMed: 16169070]
57. Schwikowski B, et al. A network of protein-protein interactions in yeast. *Nat Biotechnol*. 2000; 18:1257–1261. [PubMed: 11101803]
58. Franceschini A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*. 2013; 41:D808–D815. [PubMed: 23203871]
59. Licata L, et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res*. 2012; 40:D857–D861. [PubMed: 22096227]
60. Giot L, et al. A protein interaction map of *Drosophila melanogaster*. *Science*. 2003; 302:1727–1736. [PubMed: 14605208]
61. Dey B, et al. DNA-protein interactions: methods for detection and analysis. *Mol Cell Biochem*. 2012; 365:279–299. [PubMed: 22399265]
62. Carey MF, et al. Confirming the Functional Importance of a Protein-DNA Interaction. *Cold Spring Harb Protoc*. 2012; 7:733–757. [PubMed: 22753608]
63. Wang X, et al. Network-based methods for human disease gene prediction. *Brief Funct Genomics*. 2011; 10:280–293. [PubMed: 21764832]
64. Wu X, et al. Network-based global inference of human disease genes. *Mol Syst Biol*. 2008; 4:189. [PubMed: 18463613]
65. Moreau Y, Tranchevent L-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet*. 2012; 13:523–536. [PubMed: 22751426]

66. Menon R, Omenn GS. Identification of alternatively spliced transcripts using a proteomic informatics approach. *Methods Mol Biol.* 2011; 696:319–326. [PubMed: 21063957]
67. Li W, et al. High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. *Nucleic Acids Res.* 2013;1–15. [PubMed: 23143271]
68. Corominas R, et al. Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nat Commun.* 2014; 5
69. Rentzsch R, Orengo CA. Protein function prediction using domain families. *BMC Bioinformatics.* 2013; 14(Suppl 3):S5. [PubMed: 23514456]
70. Forslund K, Sonnhammer ELL. Predicting protein function from domain content. *Bioinformatics.* 2008; 24:1681–1687. [PubMed: 18591194]
71. Lee D, et al. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol.* 2007; 8:995–1005. [PubMed: 18037900]
72. Whisstock J, Lesk A. Prediction of protein function from protein sequence and structure. *Q Rev Biophys.* 2003; 36:307–340. [PubMed: 15029827]
73. Laskowski RA, et al. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* 2005; 33:89–W93.
74. Roy A, et al. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc.* 2010; 5:725–738. [PubMed: 20360767]
75. Konc J, et al. Structure-Based Function Prediction of Uncharacterized Protein Using Binding Sites Comparison. *PLoS Comput Biol.* 2013; 9:e1003341. [PubMed: 24244144]
76. Berman HM, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000; 28:235–242. [PubMed: 10592235]
77. Trapnell C, et al. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009; 25:1105–1111. [PubMed: 19289445]
78. Langmead B, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009; 10:R25. [PubMed: 19261174]
79. Li H, Durbin R. Fast and accurate short read alignment with Burrows- Wheeler transform. *Bioinformatics.* 2009; 25:1754–1760. [PubMed: 19451168]
80. Trapnell C, et al. Differential gene and transcript expression analysis of RNAseq experiments with TopHat and Cufflinks. *Nat Prot.* 2012; 7:562–578.
81. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Meth.* 2013; 10:71–73.
82. Andrews S, et al. Multiple instance learning with generalized support vector machines. *AAAI-02 Proceedings.* 2002:943–944.
83. Alam, FI.; Gondra, I. A Bayesian network-based tunable image segmentation algorithm for object recognition; *Signal Processing and Information Technology (ISSPIT)*, 2011 IEEE International Symposium on, 011-016; 2011.
84. Wang, J.; Zucker, J-D. Solving the multiple-instance problem: A lazy learning approach; *Proc. 17th Int'l Conf. on Machine Learning*; 2000. p. 1119-1125.
85. Kamijo T, et al. Functional and physical interactions of the ARF tumor suppressor with p53 and Mdm2. *Proc Natl Acad Sci USA.* 1998; 95:8292–8297. [PubMed: 9653180]
86. Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012; 489:57–74. [PubMed: 22955616]
87. Omenn GS, et al. A new class of protein cancer biomarker candidates: Differentially expressed splice variants of ERBB2 (HER2/neu) and ERBB1 (EGFR) in breast cancer cell lines. *Proteomics.* 2014
88. Du KL, et al. Megakaryoblastic Leukemia Factor-1 Transduces Cytoskeletal Signals and Induces Smooth Muscle Cell Differentiation from Undifferentiated Embryonic Stem Cells. *J Biol Chem.* 2004; 279:17578–17586. [PubMed: 14970199]
89. Xu Z, et al. The MLK Family Mediates c-Jun N-Terminal Kinase Activation in Neuronal Apoptosis. *Mol Cell Biol.* 2001; 21:4713–4724. [PubMed: 11416147]

90. Bjorkholm M, et al. Development of Hodgkin lymphoma in homozygotic triplets with constitutional deletion in MKL1. *Blood*. 2013; 121:4807. [PubMed: 23744493]
91. Scharenberg MA, et al. TGF- β -induced differentiation into myofibroblasts involves specific regulation of two MKL1 isoforms. *J Cell Sci*. 2014; 141:e707.
92. Franco SJ, et al. Reelin Regulates Cadherin Function via Dab1/Rap1 to Control Neuronal Migration and Lamination in the Neocortex. *Neuron*. 2011; 69:482–497. [PubMed: 21315259]
93. Gao Z, et al. Splice-Mediated Motif Switching Regulates Disabled-1 Phosphorylation and SH2 Domain Interactions. *Mol Cell Biol*. 2012; 32:2794–2808. [PubMed: 22586277]
94. Nicholson RI, et al. EGFR and cancer prognosis. *E J Cancer*. 2001; 37:9–15.
95. Baron AT, et al. Soluble Epidermal Growth Factor Receptor (sEGFR/sErbB1) as a Potential Risk, Screening, and Diagnostic Serum Biomarker of Epithelial Ovarian Cancer. *Cancer Epidemiol Biomarkers Prev*. 2003; 12:103–113. [PubMed: 12582019]
96. Wilken JA, et al. The epidermal growth factor receptor conundrum. *Cancer*. 2010; 117:2358–2360. [PubMed: 24048781]

Glossary

Alternative splicing	a highly regulated process that allows a multi-exon gene to generate several RNA sequences, resulting in multiple transcripts and proteins (alternatively spliced isoforms).
Functional annotation	Assignment of attributes to genes and gene products. The most commonly used functional annotation database is Gene Ontology (GO), which uses standardized vocabulary to describe gene characteristics.
RNA-seq	A massively parallel next-generation sequencing technology for profiling the genome-wide RNA levels in a given sample. It provides quantitative estimation of expression levels for alternatively spliced isoforms.
Machine learning	A branch of artificial intelligence that constructs models from the data to predict the properties of unseen examples. In function prediction, it refers to the methods that construct models from the genes annotated with a specific function to predict the function of uncharacterized genes.
Multiple instance learning (MIL)	A machine learning framework that is able to predict functions of isoforms by taking the isoform-level feature data and gene-level functional data as input. The functional data can be obtained from <i>e.g.</i> GO. The isoform-level feature data can be protein domain, amino acid composition, and RNA-seq expression data. In the context of MIL, a gene is termed a bag and its isoforms are termed instances.

Box 1. Isoform function predictions in the context of developmental processes

Computational prediction is a promising tool for prioritizing the isoforms that carry out different roles across tissues and developmental stages. The function prediction algorithm on isoforms of mouse genes has correctly identified the key isoforms involved in developmental processes for several genes [16]. For example, the protein encoded by the megakaryoblastic leukemia factor-1 (*MKL1*) gene interacts with the transcription factor myocardin, a key regulator of smooth muscle cell differentiation [88]. However, the specific role of *MKL1* in cell differentiation is not well studied; it is mostly known to be involved in apoptotic processes [89, 90]. Recently, a specific role was reported [91] for a novel short isoform of *MKL1* in the transformation of adipose cells into myofibroblasts. The shorter protein differs from the long canonical isoform (MKL1-L) in the N-terminal region. A functional N-terminal motif was identified in MKL1-S that allows specific induction of a group of genes including the extracellular matrix modifiers MMP16 and SPOCK3/testican-3 [91]. Except for T-cell and B-cell differentiation, the MIL algorithm predicted [16] the shorter isoform to be highly enriched in cell type differentiation. Moreover, MKL1-L was predicted to be more likely involved in activation of pro-apoptotic gene products compared to MKL1-S, reflecting the functional role of the canonical MKL1-L isoform.

Another interesting example is Disabled-1 (*Dab1*), which is involved in reelin-mediated neuronal migration during brain development [92]. Alternative splicing results in different isoforms of *Dab1* during development [93]. A shorter variant of *Dab1* lacking a YQXI motif is expressed at early developmental stages, whereas a longer variant with this motif is mainly expressed at late developmental stages [93]. Consistent with this, the MIL algorithm [16] predicts a prominent role for the shorter variant in the embryonic stage. The role of the canonical isoform of *Dab1*, which contains two YQXI and two YXVP phosphorylation sites, has been well characterized in neuronal migration [92, 93]. In accordance with this observation, the functional prediction based on MIL showed higher enrichment for neuron migration for the longer *Dab1* variant compared to the shorter one [16]. The above examples reveal that RNAseq data integration at the isoform level can predict the key isoforms linked to specific development processes, making it informative when there are no functional annotations available for the isoforms of the genes regulating the developmental processes.

Box 2. Isoform function predictions in the context of cancers

Splice isoform functions are associated with the development and progression of cancers. Predicted isoform functions can help to identify isoforms whose aberrant expression may lead to cancers. Two interesting isoform function predictions in relation to cancers are illustrated below. Overexpression of the epidermal growth factor receptor (EGFR) has been observed in multiple cancer types and in solid tumour growth [94]. The soluble long isoform EGFR-L (1210 amino acids) is the source of the major circulating epidermal growth receptor in human blood [95]. As predicted by integrating the RNA-seq data [16], mouse EGFR-L is likely to be involved in the regulation of coagulation, cholesterol metabolic process and response to steroid hormone stimulus, while EGFR-S is not. Conversely, the regulation of anti-apoptosis is more enriched for EGFR-S compared with EGFR-L. The concentration of serum EGFR-S is inversely associated with I/II and stage III/IV epithelial ovarian cancer, compared with levels in healthy women [95] and may complicate the efficacy of EGFR-targeted therapies [96]. The *Anxa6* gene codes two isoforms: NM_001110211.1 and NM_013472.4. The former encodes a shorter protein by six amino acids difference due to the lack of an alternate in-frame exon compared to the latter (Figure 3) [18]. The 3-dimensional (3D) confirmation modeling of these two isoforms shows a difference in the likelihood of phosphorylation in the Thr-Pro-Ser sequence, consistent with the prediction by MIL. In addition, functional differences of the two isoforms on catecholamine secretion were reported [18]. The above examples suggest the potential of isoform-level genomic data integration and function prediction to help us understand cancer biology.

Highlights

- Splice isoforms of the same gene may show different and even opposite functions.
- Accurate prediction of isoform functions can be achieved through integrating RNA-seq data.
- Isoform-level function analysis is valuable for understanding development and cancers.
- Future work includes network modeling and heterogeneous data integration.

Figure 1.
(A) Five basic mechanisms of alternative splicing. (B) Current statistical numbers of human genes and transcripts based on GENCODE annotation (version 19). lnc genes: long non-coding genes; snc genes: small non-coding genes; lnc transcripts: long non-coding transcripts; NMD transcripts: non-sense mediated decay transcripts. The count of genes with different number of transcripts is plotted; on average a gene has three transcripts annotated in GENCODE.

Figure 2.

Methods for isoform function prediction and future perspectives. Current protein structure-based methods take amino acid sequence, secondary or tertiary structure as input for predicting isoform functions. The modeling methods include but are not limited to domain-based methods and homology modeling. Multiple instance learning (MIL) is a recently developed method for predict isoform functions by integrating isoform-level feature data (*e.g.* RNA-seq) and gene-level functional annotation data. For a given function under study, green circles represent positive genes associated with this function; blue rectangles represent negative genes not having this function; each small element in genes represents an isoform. For a positive gene, at least one of the isoforms must carry out the function (red). For a negative gene, none of its isoforms should carry out the function under study (white). In terms of general terminology of MIL, genes and its isoforms are called bags and instances, respectively. Future perspectives of the isoform-level functional analyses include integrating heterogeneous genomic data with additional predictive values, networks for better understanding of isoform-isoform interaction or isoform-phenotype associations, and tissue and developmental stage-specific predictions.

Figure 3.

Structural differences of functionally different isoforms of the same gene. **A.** The upper part shows the gene models of the two isoforms NM_001040654.1 and NM_009877.2 of the *Cdkn2a* gene in mouse. The computationally modeled 3D structure of NM_001040654.1 in mouse is characterized by ankyrin repeats. In contrast, the 3D structure of NM_009877.2 has a CDKN2a N-terminus domain. Adapted from [16]. **B.** Superimposed, predicted 3D structures of the two isoforms of the *Anxa6* gene. Due to the absence of “VAAEIL” residues (existing in NM_013472.4) in the NM_001110211.1 variant (red), the loop region is

smaller. The structural alignment shows that the “TPS” residues are not aligned between the *Anxa6* variants (inset box), which could affect the post-translational modification of Thr and Ser residues (as shown in green and red spheres in the inset figure). Adapted from [18].