Journal of proteome research

Article

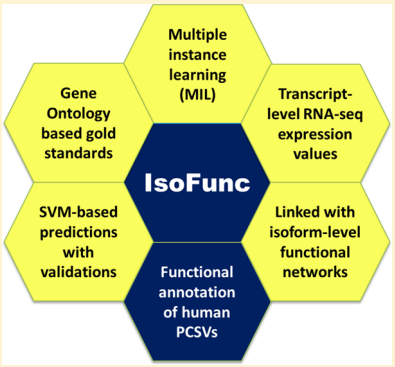
pubs.acs.org/jpr

Genome-Wide Functional Annotation of Human Protein-Coding Splice Variants Using Multiple Instance Learning

Bharat Panwar,*[,†] Rajasree Menon,[†] Ridvan Eksi,[†] Hong-Dong Li,[†] Gilbert S. Omenn,*[,†,‡,§] and Yuanfang Guan*[,†,‡,∥]

[†]Department of Computational Medicine and Bioinformatics, [‡]Department of Internal Medicine, [§]Department of Human Genetics and School of Public Health, and [∥]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109, United States

ABSTRACT: The vast majority of human multiexon genes undergo alternative splicing and produce 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 resource called "IsoFunc", which is freely available for the global scientific community through http://guanlab.ccmb.med.umich.edu/isofunc.



KEYWORDS: *alternative splicing, protein-coding splice variant (PCSV), functional annotation, gene ontology (GO), multiple instance learning (MIL), support vector machine (SVM), RNA-seq, ADAM15, LMNA/C, DMXL2, IsoFunc*

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

Received: September 18, 2015

DOI: 10.1021/acs.jproteome.5b00883
J. Proteome Res. XXXX, XXX, XXX−XXX

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.

MATERIAL AND METHODS

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 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.

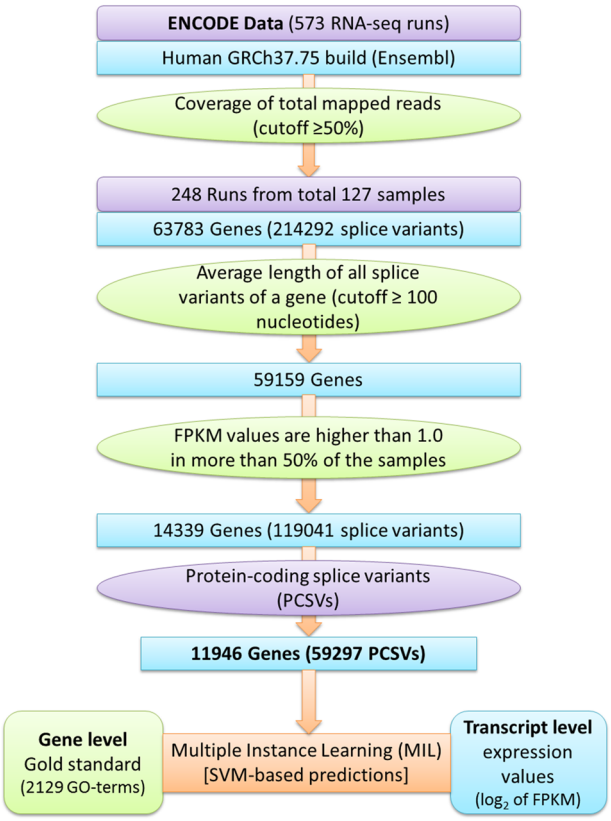


Figure 1. Overview of data preprocessing for predicting protein-coding splice variants. We collected RNA-seq data from the ENCODE project and estimated the expression values using standard tools and thresholds. These values were used as input features for developing SVM models for different GO terms using the multiple instance learning approach.

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 13931 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.

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.

DOI: 10.1021/acs.jproteome.5b00883
J. Proteome Res. XXXX, XXX, XXX−XXX

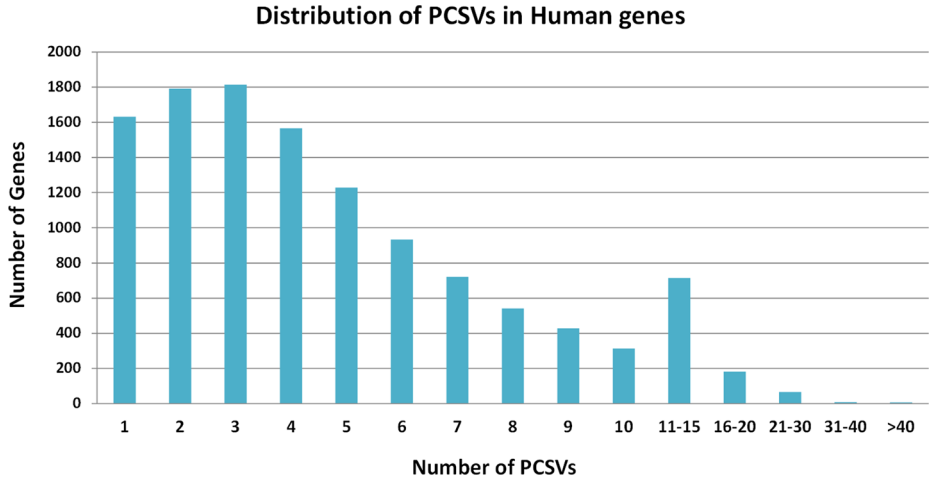


Figure 2. Distribution of number of protein-coding splice variants across well-expressed human genes.

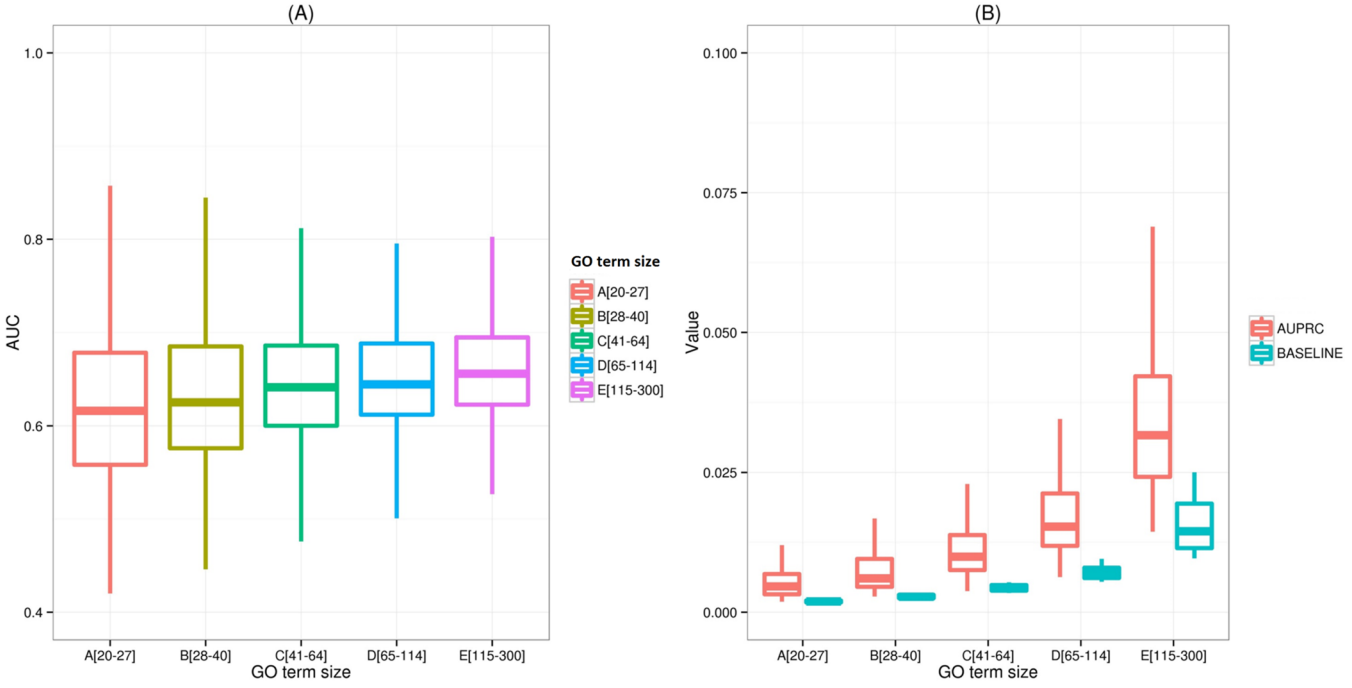


Figure 3. Performance of our multiple-instance learning based algorithm for predicting functions of protein-coding splice variants. We used two different parameters (A) AUC and (B) AUPRC to evaluate prediction performance of the algorithm. The baseline values are also given with AUPRC. The different performances calculated for five different GO term sizes are shown in different colors.

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]

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

DOI: 10.1021/acs.jproteome.5b00883
J. Proteome Res. XXXX, XXX, XXX−XXX

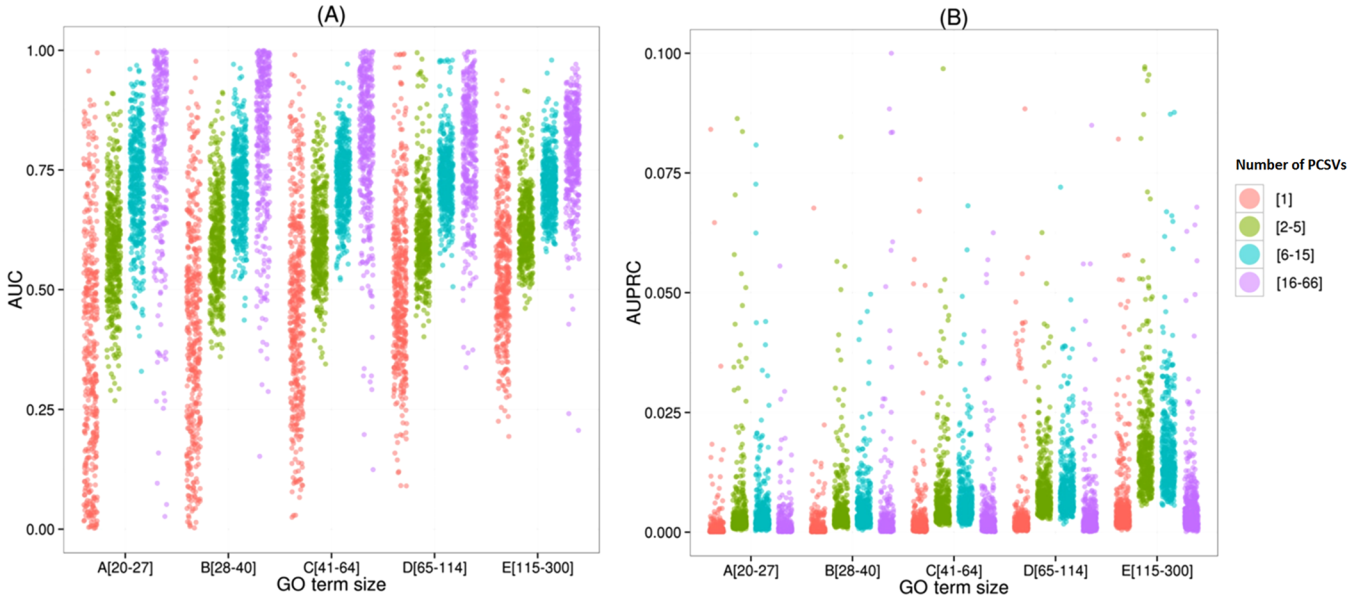


Figure 4. Comparative performance of single PCSV genes and multi-PCSVs genes. Two different parameters (A) AUC and (B) AUPRC have been used to evaluate prediction performances. There are different performances calculated for five different GO term sizes as well as genes with a different number of PCSVs. 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.

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).

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.

RESULTS AND DISCUSSION

Abundance of Protein-Coding Splice Variants

We analyzed a total of 59 297 PCSVs of 11 946 well-expressed genes. Figure 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.

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 3A); 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 3B). 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 3B).

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 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 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 4A). Performance of single PCSV genes is lower in comparison with multiple PCSVs genes in terms of AUPRC (Figure 4B).

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.

DOI: 10.1021/acs.jproteome.5b00883
J. Proteome Res. XXXX, XXX, XXX−XXX

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.

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.

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

In a recent publication we highlighted six alternative/non-canonical 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]

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

DOI: 10.1021/acs.jproteome.5b00883
J. Proteome Res. XXXX, XXX, XXX−XXX

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*.

■ 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.

■ AUTHOR INFORMATION

Corresponding Authors

*B.P.: E-mail: bharatpa@umich.edu. Phone: +1-734-764-0018. Fax: +1-734-615-6553.
*G.S.O.: E-mail: gomenn@umich.edu. Phone: +1-734-763-7583. Fax: +1-734-615-6553.
*Y.G.: E-mail: gyuanfan@umich.edu. Phone: +1-734-764-0018. Fax: +1-734-615-6553.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by NSF 1452656 (Y.G.) and National Institutes of Health grant nos. 1R21NS082212-01 (Y.G.) and U54ES017885 (G.S.O.).

■ REFERENCES

(1) Dunham, I.; Kundaje, A.; Aldred, S. F.; Collins, P. J.; Davis, C. A.; Doyle, F.; Epstein, C. B.; Frietze, S.; Harrow, J.; Kaul, R.; et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489* (7414), 57−74.
(2) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2008**, *4* (1), 44−57.
(3) Murali, T. M.; Wu, C.-J.; Kasif, S. The art of gene function prediction. *Nat. Biotechnol.* **2006**, *24* (12), 1474−1475 author reply 1475−1476..
(4) Pan, Q.; Shai, O.; Lee, L. J.; Frey, B. J.; Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **2008**, *40* (12), 1413−1415.
(5) Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **2003**, *72*, 291−336.
(6) de Klerk, E.; 't Hoen, P. A. C. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet.* **2015**, *31*, 128.
(7) Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; et al. A draft map of the human proteome. *Nature* **2014**, *509* (7502), 575−581.
(8) 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), 8431−8441.
(9) López-Bigas, N.; Audit, B.; Ouzounis, C.; Parra, G.; Guigó, R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.* **2005**, *579* (9), 1900−1903.
(10) Skotheim, R. I.; Nees, M. Alternative splicing in cancer: noise, functional, or systematic? *Int. J. Biochem. Cell Biol.* **2007**, *39* (7−8), 1432−1449.
(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.
(12) Sterne-Weiler, T.; Howard, J.; Mort, M.; Cooper, D. N.; Sanford, J. R. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.* **2011**, *21* (10), 1563−1571.
(13) Sterne-Weiler, T.; Sanford, J. R. Exon identity crisis: disease-causing mutations that disrupt the splicing code. *Genome Biol.* **2014**, *15* (1), 201.
(14) Xiong, H. Y.; Alipanahi, B.; Lee, L. J.; Bretschneider, H.; Merico, D.; Yuen, R. K. C.; Hua, Y.; Gueroussov, S.; Najafabadi, H. S.; Hughes, T. R.; et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science (Washington, DC, U. S.)* **2015**, *347* (6218), 1254806.
(15) Eksi, R.; Li, H.-D.; Menon, R.; Wen, Y.; Omenn, G. S.; Kretzler, M.; Guan, Y. Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data. *PLoS Comput. Biol.* **2013**, *9* (11), e1003314.
(16) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25* (1), 25−29.
(17) Li, H.-D.; Menon, R.; Omenn, G. S.; Guan, Y. The emerging era of genomic data integration for analyzing splice isoform function. *Trends Genet.* **2014**, *30* (8), 340−347.
(18) Bréhélin, L.; Florent, I.; Gascuel, O.; Maréchal, E. Assessing functional annotation transfers with inter-species conserved coexpression: application to Plasmodium falciparum. *BMC Genomics* **2010**, *11*, 35.
(19) Childs, K. L.; Davidson, R. M.; Buell, C. R. Gene coexpression network analysis as a source of functional annotation for rice genes. *PLoS One* **2011**, *6* (7), e22196.
(20) Piro, R. M.; Ala, U.; Molineris, I.; Grassi, E.; Bracco, C.; Perego, G. P.; Provero, P.; Di Cunto, F. An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction. *Eur. J. Hum. Genet.* **2011**, *19* (11), 1173−1180.
(21) Pavlidis, P.; Gillis, J. Progress and challenges in the computational prediction of gene function using networks: 2012−2013 update. *F1000Research* **2013**, *2*, 230.

DOI: 10.1021/acs.jproteome.5b00883
J. Proteome Res. XXXX, XXX, XXX−XXX

(22) Dietterich, T. G.; Lathrop, R. H.; Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* **1997**, *89* (1−2), 31−71.
(23) Li, W.; Kang, S.; Liu, C.-C.; Zhang, S.; Shi, Y.; Liu, Y.; Zhou, X. J. 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.
(24) Trapnell, C.; Pachter, L.; Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **2009**, *25* (9), 1105−1111.
(25) Matsuo, A.; Bellier, J.-P.; Hisano, T.; Aimi, Y.; Yasuhara, O.; Tooyama, I.; Saito, N.; Kimura, H. Rat choline acetyltransferase of the peripheral type differs from that of the common type in intracellular translocation. *Neurochem. Int.* **2005**, *46* (5), 423−433.
(26) Schneider, E.; El Hajj, N.; Richter, S.; Roche-Santiago, J.; Nanda, I.; Schempp, W.; Riederer, P.; Navarro, B.; Bontrop, R. E.; Kondova, I.; et al. Widespread differences in cortex DNA methylation of the "language gene" CNTNAP2 between humans and chimpanzees. *Epigenetics* **2014**, *9* (4), 533−545.
(27) Fu, W. J.; Carroll, R. J.; Wang, S. Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics* **2005**, *21* (9), 1979−1986.
(28) Andrews, S.; Tsochantaridis, I.; Hofmann, T. Support vector machines for multiple-instance learning. *Adv. Neural Inf. Process. Syst.* **2002**, 561−568.
(29) Guan, Y.; Myers, C. L.; Hess, D. C.; Barutcuoglu, Z.; Caudy, A. A.; Troyanskaya, O. G. Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome Biol.* **2008**, *9* (Suppl 1), S3.
(30) Panwar, B.; Arora, A.; Raghava, G. P. S. Prediction and classification of ncRNAs using structural information. *BMC Genomics* **2014**, *15*, 127.
(31) Panwar, B.; Gupta, S.; Raghava, G. P. Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. *BMC Bioinf.* **2013**, *14* (1), 44.
(32) Panwar, B.; Raghava, G. P. Prediction of uridine modifications in tRNA sequences. *BMC Bioinf.* **2014**, *15*, 326.
(33) Panwar, B.; Raghava, G. P. S. Identification of protein-interacting nucleotides in a RNA sequence using composition profile of tri-nucleotides. *Genomics* **2015**, *105* (4), 197−203.
(34) Zhong, J. L.; Poghosyan, Z.; Pennington, C. J.; Scott, X.; Handsley, M. M.; Warn, A.; Gavrilovic, J.; Honert, K.; Krüger, A.; Span, P. N.; et al. Distinct functions of natural ADAM-15 cytoplasmic domain variants in human mammary carcinoma. *Mol. Cancer Res.* **2008**, *6* (3), 383−394.
(35) Charrier-Hisamuddin, L.; Laboisse, C. L.; Merlin, D. ADAM-15: a metalloprotease that mediates inflammation. *FASEB J.* **2007**, *22* (3), 641−653.
(36) Chomarat, P.; Briolay, J.; Banchereau, J.; Miossec, P. Increased production of soluble CD23 in rheumatoid arthritis, and its regulation by interleukin-4. *Arthritis Rheum.* **1993**, *36* (2), 234−242.
(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. *Immunol. Today* **1996**, *17* (9), 418−420.
(38) Lopez-Mejia, I. C.; de Toledo, M.; Chavey, C.; Lapasset, L.; Cavelier, P.; Lopez-Herrera, C.; Chebli, K.; Fort, P.; Beranger, G.; Fajas, L.; et al. Antagonistic functions of LMNA isoforms in energy expenditure and lifespan. *EMBO Rep.* **2014**, *15* (5), 529−539.
(39) Frangioni, J. V.; Neel, B. G. 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. *J. Cell Sci.* **1993**, *105* (Pt 2), 481−488.
(40) Emerson, L. J.; Holt, M. R.; Wheeler, M. A.; Wehnert, M.; Parsons, M.; Ellis, J. A. Defects in cell spreading and ERK1/2 activation in fibroblasts with lamin A/C mutations. *Biochim. Biophys. Acta, Mol. Basis Dis.* **2009**, *1792* (8), 810−821.
(41) Menon, R.; Panwar, B.; Eksi, R.; Kleer, C.; Guan, Y.; Omenn, G. S. Computational Inferences of the Functions of Alternative/Noncanonical Splice Isoforms Specific to HER2+/ER-/PR- Breast Cancers, a Chromosome 17 C-HPP Study. *J. Proteome Res.* **2015**, *14* (9), 3519−3529.
(42) Omenn, G. S.; Lane, L.; Lundberg, E. K.; Beavis, R. C.; Nesvizhskii, A. I.; Deutsch, E. W. Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *J. Proteome Res.* **2015**, *14* (9), 3452−3460.
(43) Gaudet, P.; Michel, P.-A.; Zahn-Zabal, M.; Cusin, I.; Duek, P. D.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Teixeira, D.; et al. The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Res.* **2015**, *43* (D1), D764−D770.
(44) Lonsdale, J.; Thomas, J.; Salvatore, M.; Phillips, R.; Lo, E.; Shad, S.; Hasz, R.; Walters, G.; Garcia, F.; Young, N.; et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **2013**, *45* (6), 580−585.
(45) Li, H.-D.; Menon, R.; Govindarajoo, B.; Panwar, B.; Zhang, Y.; Omenn, G. S.; Guan, Y. Functional Networks of Highest-Connected Splice Isoforms: From The Chromosome 17 Human Proteome Project. *J. Proteome Res.* **2015**, *14* (9), 3484−3491.
(46) Zhu, F.; Panwar, B.; Guan, Y. Algorithms for modeling global and context-specific functional relationship networks. *Briefings Bioinf.* **2015**, bbv065.
(47) Li, H.-D.; Menon, R.; Eksi, R.; Guerler, A.; Zhang, Y.; Omenn, G. S.; Guan, Y. A Network of Splice Isoforms for the Mouse. *Sci. Rep.* **2016**, *6*, 24507.
(48) Panwar, B.; Menon, R.; Eksi, R.; Omenn, G. S.; Guan, Y. MI-PVT: A Tool for Visualizing the Chromosome-Centric Human Proteome. *J. Proteome Res.* **2015**, *14* (9), 3762−3767.

DOI: 10.1021/acs.jproteome.5b00883
J. Proteome Res. XXXX, XXX, XXX−XXX