Published online 22 June 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 2 **1**
https://doi.org/10.1093/nargab/lqab057

FINER: enhancing the prediction of tissue-specific functions of isoforms by refining isoform interaction networks

Hao Chen [1], Dipan Shaw [1], Dongbo Bu [2,3,*] and Tao Jiang [1,4,*]

[1]Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA, [2]Key Lab of Intelligent Information Process, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, [3]University of Chinese Academy of Sciences, Beijing 100049, China and [4]Bioinformatics Division, BNRIST/Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Received February 04, 2021; Revised May 18, 2021; Editorial Decision May 31, 2021; Accepted June 03, 2021

ABSTRACT

Annotating the functions of gene products is a mainstay in biology. A variety of databases have been established to record functional knowledge at the gene level. However, functional annotations at the isoform resolution are in great demand in many biological applications. Although critical information in biological processes such as protein–protein interactions (PPIs) is often used to study gene functions, it does not directly help differentiate the functions of isoforms, as the 'proteins' in the existing PPIs generally refer to 'genes'. On the other hand, the prediction of isoform functions and prediction of isoform–isoform interactions, though inherently intertwined, have so far been treated as independent computational problems in the literature. Here, we present FINER, a unified framework to jointly predict isoform functions and refine PPIs from the gene level to the isoform level, enabling both tasks to benefit from each other. Extensive computational experiments on human tissue-specific data demonstrate that FINER is able to gain at least 5.16% in AUC and 15.1% in AUPRC for functional prediction across multiple tissues by refining noisy PPIs, resulting in significant improvement over the state-of-the-art methods. Some in-depth analyses reveal consistency between FINER's predictions and the tissue specificity as well as subcellular localization of isoforms.

INTRODUCTION

Annotating functions of gene products in complex biological systems is of fundamental importance. A large number of annotation approaches (1,2) have been proposed and a variety of databases have been established to record functional annotation of genes (3,4). However, most of the existing functional annotations are at the gene level, which is coarse-grained and insufficient as a gene might have multiple products. In fact, alternative splicing of mRNAs frequently occurs in eukaryotes, leading a single gene to often produce multiple protein isoforms (5). The isoforms of a gene may carry different or even opposite biological functions (6). For instance, two of the isoforms of BCL2L1 gene, BCL-xL and BCL-xS, exhibit completely opposite functions: BCL-xL inhibits programmed cell death while BCL-xS promotes it (7). The diversity of gene products requires finer functional annotations at the isoform level instead of the gene level.

Since the experimental technologies to determine isoform functions are usually time-consuming and costly, computational approaches to predict isoform functions are highly desired. Many methods have been proposed in recent years (8–16). Most of these approaches apply the multiple instance learning (MIL) technique to explore isoform features, including isoform sequence motifs, conserved domains and expression profiles. More specifically, the MIL technique attempts to learn function-specific isoform features, i.e. features that belong to at least one isoform of each gene possessing the function. The resulting function-specific features are then used to predict the functions of new (or queried) isoforms. However, these methods all suffer from the limitation that some key functional features (such as protein–protein interactions discussed below), which are proved to be effective in predicting gene functions, may not be available at the isoform level. Hence, they have prediction performance less than desirable.

Besides functional features of individual isoforms, the interactions among isoforms also form an important information source of isoform functions. The underlying rationale can be clearly demonstrated by an analogy to protein–protein interactions (PPIs): a protein usually performs spe-

*To whom correspondence should be addressed. Tel: +1 951 8272991; Fax: +1 951 8274643; Email: jiang@cs.ucr.edu
Correspondence may also be addressed to Dongbo Bu. Tel: +86 106 2601019; Fax: +86 106 2601356; Email: dbu@ict.ac.cn

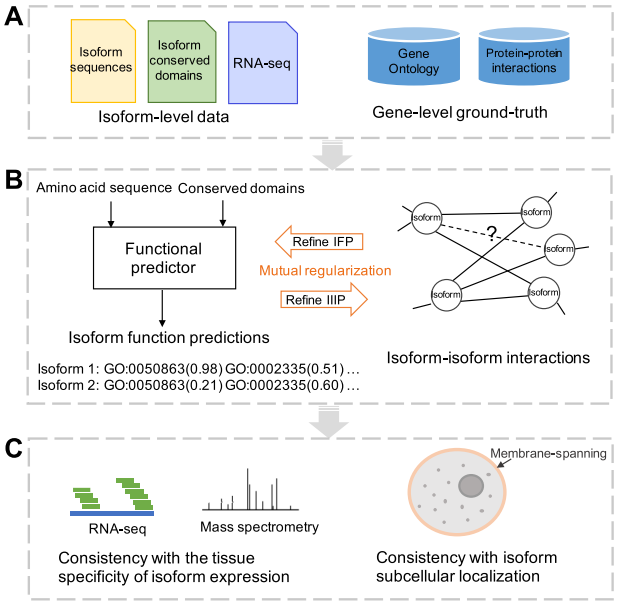

Figure 1. Schematic overview of the FINER workflow.

cific functions through interacting with other proteins (17), thus enabling the prediction of protein functions through analyzing protein interactions. The existing PPI networks are essentially at the gene level as they exhibit only interactions among corresponding genes without providing more detailed information concerning the interaction of isoforms. In fact, the isoforms of a gene may have different interacting partners, possibly due to the difference in their interacting domains resulted from alternative splicing (18). Thus, although PPIs have been successfully used for predicting gene functions (19,20), they cannot be directly applied to infer fine-grained isoform functions. Recently, extensive studies have been performed to refine protein–protein interactions into isoform–isoform interactions (IIIs) (21–25). Clearly, the problems of isoform function prediction (IFP) and isoform–isoform interaction prediction (IIIP) are inherently intertwined, implying that they may not be well addressed if considered separately. Thus, how to solve the two problems jointly and exploit the reciprocal relationship between them remains an interesting challenge.

In this work, we present a novel approach, called FINER (i.e. enhancing the Functional prediction of Isoforms via NEtwork Refinement on their interactions), that jointly solves the IFP and IIIP problems, thus allowing one problem to benefit from the other. Our approach contains three key elements as shown in Figure 1B: (i) The function prediction module predicts the functional labels of isoforms from their amino acid sequences and conserved domains. (ii) The interaction network refinement module identifies real interacting isoform pairs from known interacting gene pairs and denoises the existing IIIs simultaneously. (iii) A mutual regularizer encourages the above two modules to agree with each other, i.e. isoforms with similar predicted functions will be likely connected in refined III networks and vice versa. Through the mutual regularizer, the function prediction module and interaction network refinement module exchange information and, in turn, improve their own prediction of isoform functions and interactions.

To evaluate our approach FINER, we applied it to predict tissue-specific functions and interactions of isoforms in human. Understanding tissue-specific functions of isoforms is an important but challenging task: On one hand, tissue-dependent isoform usages are pervasive across human tissues, since a gene may express various isoforms to perform different functions in different tissues (26,27). On the other hand, less is known about the tissue specificity of PPIs (28). Although PPIs can be associated with tissues through the consideration of tissue-specific expression data (29,30), the derived interactions are perhaps less reliable, thus making the refinement on tissue-specific interactions highly desirable. In addition, diverse tissues serve as multiple sources of datasets for testing our approach. We further analyzed the relationship between our functional prediction and the subcellular localization and the tissue specificity of isoforms (Figure 1C). The experimental results clearly demonstrate the advantages of our approach over the state-of-the-art methods in predicting isoform functions and interactions, as well as its potential in revealing the roles of isoforms in diverse human tissues and diseases.

MATERIALS AND METHODS

Data collection

To predict isoform functions, we need gene-level functional annotation ground-truth, features of individual isoforms (including isoform sequences and conserved domains) and isoform–isoform interactions (derived from gene-level protein–protein interactions and isoform co-expression networks) (Figure 1A). The data used in the study are described in detail as follows.

(i) Isoform sequences: We downloaded 'Coding DNA Sequence' (CDS) of human genome (GRCh38.p13) from the NCBI Reference Sequences database (RefSeq, as of January, 2020) (31). For each CDS, we constructed an isoform by translating it into the amino acid sequence. Two or more isoforms corresponding to the same CDS are treated as a single isoform. To ensure isoform quality, only manually curated RefSeq records were recruited into our study. As a result, we obtained a total of 43 289 isoforms from 19 408 genes, consisting of 33 529 isoforms from 9648 multiple-isoform genes (MIGs) and 9760 single-isoform genes (SIGs).

(ii) Isoform conserved domains: For each isoform, we acquired its conserved domains by searching its amino acid sequence against the NCBI Conserved Domain Database (32).

(iii) Functional annotation ground-truth of genes: We adopted the functional terms defined by Gene Ontology (GO) (3), wherein GO terms are organized in hierarchies represented as directed acyclic graph (DAG) structures, describing functions at different levels of abstraction. For the genes used in the study, we downloaded their functional annotations from the Gene Ontology Annotation (GOA) database (33). To ensure the annotation quality, we kept only manually curated GO

annotations and skipped electronic annotations containing the 'IEA' evidence code.

(iv) Protein–protein interactions: We used the PPI data collected by Zitnik *et al.* (20), in which, various types of physical PPIs from six reputable resources were combined (34–39). All the PPIs have experimental supports. The reader is referred to Zitnik *et al.* (20) for a detailed description of the data. By mapping the collected data to the genes used in our study, we acquired a total of 317 750 interactions among 19 408 genes.

(v) *Isoform expression profiles*: To collect expression profiles of isoforms, we first retrieved RNA-seq experiments for different types of normal human tissues from the NCBI Sequence Read Archive (SRA) database (40), where corresponding accession numbers were obtained from the Human Protein Atlas (HPA) database (41) and the recount-brain project (42) (see Supplementary Table S1 for a list of RNA-seq experiments). Next, we applied the tool Kallisto (43) to obtain quantified isoform expression profiles in each experiment (measured in Transcripts Per Million or TPM).

Construction of tissue-specific datasets

In the study, we applied FINER to predict isoform functions for 12 selected major tissues and three brain sub-tissue of human. These tissues were selected as follows. From the tissues recorded in the BRENDA Tissue Ontology (44), we first selected tissues with valid tissue-specific GO terms. Here, GO terms specifically describing cellular functions of each tissue were retrieved from Greene *et al.* (27), and only GO terms associated with at least five genes were recruited into our experiments (see Supplementary Table S2 for the lists of tissue-specific GO terms). Next, following the criterion used by Li *et al.* (9), we selected tissues associated with at least six RNA-seq experiments to guarantee the quality of co-expression networks to be constructed later. As a result, we obtained a total of 12 major tissues and three brain sub-tissues of human, which are rich enough to represent both diversity and different levels of specificity of human tissues.

Unlike isoform sequences and conserved domains that are tissue-independent, isoform expression profiles and interactions are highly tissue-specific. To construct tissue-specific PPI networks, we first selected genes with high tissue specificity, i.e. the so-called 'tissue enhanced genes' (41). Specifically, for each of the 12 major tissues, we selected genes that have at least four-fold higher mRNA levels over the average levels in the other major tissues. For the three brain sub-tissues, we relaxed the above threshold to 2-fold due to the smaller differences between sub-tissues. Next, a subnetwork was extracted from the global PPI network for each tissue as the tissue-specific PPI network, in which each edge from the global PPI network was included if at least one of the two genes connected by the edge is tissue enhanced. The underlying rationale is that tissue enhanced genes are likely to perform functions specific to the involved tissues, while their interacting partners, if not tissue enhanced, are likely ubiquitously expressed genes that perform tissue-specific functions only when interacting with tissue enhanced genes (27,45).

We further constructed isoform co-expression networks by measuring expression correlations of isoform pairs across all RNA-seq experiments associated with the tissue, wherein only isoforms of genes appearing in the corresponding tissue-specific PPI network were considered. Expression correlation coefficients as edge weights were computed by the absolute value of the leave-one-out Pearson correlation coefficients (46), which is robust against single experimental outliers. To retain reliable co-expression edges but avoid noisy ones, we only kept the top five percent edges with the largest weights in each co-expression network.

The framework of FINER

The architecture of FINER consists of three key modules, namely, the function prediction module that predicts isoform functions (denoted as GO terms) for the input isoforms from their sequences and domains, the III refinement module that iteratively refines the gene-level PPI network to the isoform–isoform interaction network by taking into account isoform co-expression relationship, and a mutual regularization module that enables the exchange of information between the above two modules. A schematic illustration of the architecture is provided in Figure 2. The details of the three modules, together with their training procedures, are described below.

Function prediction module and its learning objective

We constructed the function prediction module on the basis of DIFFUSE (12) with extensions (Figure 2A). The backbone of DIFFUSE is a deep neural network designed for predicting isoform functions based on isoform sequences and conserved domains. Specifically, the neural network contains two components: (i) A convolutional neural network (CNN) component is used to extract sequence features of isoforms, in which the amino acid sequence of each isoform is encoded as a series of overlapping tri-grams $\boldsymbol{s}$. Each tri-gram is encoded as a continuous vector by the dense embedding layer (47). A one-dimensional convolutional layer with multiple convolution filters is then employed to detect functional sites by scanning the encoded sequence and represent the extracted information into a sequence hidden feature vector $\boldsymbol{h}_s$. A pyramid pooling layer was designed in the CNN component to deal with isoform sequences of different lengths. (ii) The other component is a recurrent neural network (RNN) with long short-term memory (LSTM) units (48). In the RNN component, each type of conserved domain is represented as a unique token. Domains of each isoform are ordered as a sequence of tokens $\boldsymbol{d}$, which are encoded by the same dense embedding technique and then input to the LSTM units successively. Content of tokens with their order information for each isoform are thus captured and again represented as a domain hidden feature vector $\boldsymbol{h}_d$.

The two types of hidden features, $\boldsymbol{h}_s$ and $\boldsymbol{h}_d$, are concatenated and then fused as a unified functional feature vector $\boldsymbol{h}$ through a fully connected layer. The feature extraction and fusion process are formally defined as:

$$\boldsymbol{h} = Dense([\boldsymbol{h}_s, \boldsymbol{h}_d]) = Dense([CNN(\boldsymbol{s}), RNN(\boldsymbol{d})]), \quad (1)$$

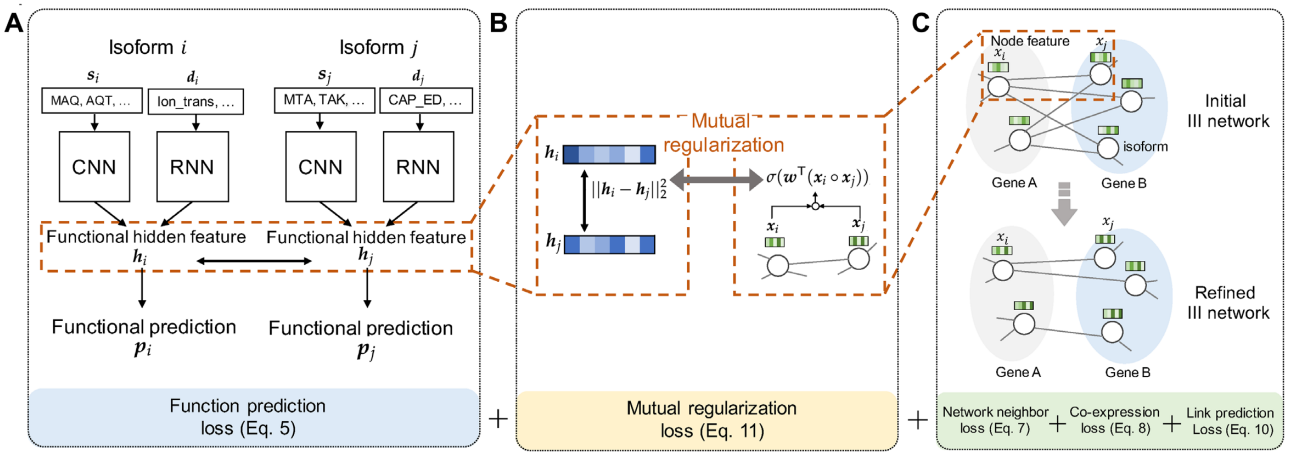

Figure 2. Schematic illustration of the architecture of FINER, which consists of three modules: (**A**) a neural network based function prediction module, (**C**) an III refinement module for iteratively updating III networks and (**B**) a mutual regularization module that is introduced to enable the previous two modules to exchange information with each other. That is, module B encourages isoforms with similar predicted functions to be more likely connected in the refined III networks, and vice versa. See the Materials and methods section for more details.

where $Dense(\cdot)$ denotes the fully connected layer and $[\cdot,\cdot]$ denotes the concatenation of two vectors.

Unlike DIFFUSE, which produces a binary prediction on each individual GO term, FINER produces a multi-label prediction on all the GO terms specific to a given tissue simultaneously, thus making the entire training process more efficient and allowing common knowledge to be shared across all GO terms. Specifically, we used a fully connected layer to map a functional feature vector $\boldsymbol{h}$ to an output vector $\boldsymbol{o}$:

$$\boldsymbol{o} = Dense(\boldsymbol{h}). \tag{2}$$

Here, the output vector $\boldsymbol{o}$ has T dimensions, where T denotes the number of GO terms specific to a tissue. The sigmoid function is applied on each dimension of the output to normalize the prediction on each GO term to a score in the range [0, 1], indicating how likely the input isoform performs the corresponding function.

Because of the hierarchical nature of GO, an isoform is automatically labeled with a GO term if any of its child terms are labeled on the isoform. To ensure consistent prediction on all GO terms, we designed a hierarchical prediction layer as done in Kulmanov *et al.* (19). For each term in the set of T GO terms, we created a binary mask vector, denoted as $\boldsymbol{c}_t$ (where $t = 1, 2, ..., T$), wherein the bits corresponding to the GO term and its children are set as 1. The maximum score from the element-wise product of the output vector and the mask vector is set as the GO term's prediction, which is formally denoted as:

$$a_t = max(\boldsymbol{c}_t \circ \boldsymbol{o}) \quad \text{for } t = 1, 2, ..., T. \tag{3}$$

Finally, the prediction results on all T terms are merged as the functional prediction of the input isoform $\boldsymbol{p} = Hierarchical(\boldsymbol{o}) = (a_1, a_2, ..., a_T)$.

To overcome the difficulty of lack of ground-truth isoform function annotations, we applied the MIL technique, following the previous work on isoform function prediction in (8,12). Specifically, each gene is treated as a bag and the isoforms of a gene are treated as the instances of the bag. For a given function, positive bags refers to genes associated with the function. Clearly, a positive bag should contain at least one positive instance but may also have some negative instances, while a negative bag should contain no positive instances. We initialize all instances of positive bags with positive labels, and the others with negative labels. Given an isoform i and its initial label on GO term t, we can define the following 'binary cross entropy loss':

$$l_{i,t} = -(y_{i,t}\log(p_{i,t}) + (1 - y_{i,t})\log(1 - p_{i,t})), \tag{4}$$

where $y_{i,t}$ is a one-hot indicator for the label of isoform i on GO term t, and $p_{i,t}$ is the corresponding prediction score. To characterize the above bag instance relationship, we weighted each 'binary cross entropy loss' by the corresponding prediction score, so that significant punishment would be applied on large prediction scores with negative labels but not on small prediction scores with positive labels.

Given a set of K isoforms, the learning objective for the function prediction module is to minimize the following 'function prediction loss' defined by the following weighted binary cross entropy (49):

$$L_{fp} = -\sum_{i}^{K}\sum_{t}^{T} \hat{p}_{i,t} l_{i,t}, \tag{5}$$

where $\hat{p}_{i,t}$ is a constant assigned by $p_{i,t}$ to avoid direct minimization of the prediction score. The isoform labels are recalculated after each training iteration under the MIL constraints, which will be described in more detail in the sections below.

III refinement module and its learning objective

For a given tissue, we iteratively refine its isoform–isoform interaction network initialized as the tissue-specific PPI network (Figure 2C). The III network contains the isoforms of genes that appear in the tissue-specific PPI network. Initially, we connect isoforms if and only if their genes have

interactions in the tissue-specific PPI network. We formally define an III network as a undirected graph $G_{III} = (\boldsymbol{V}, \boldsymbol{E})$, in which isoforms are represented as a set of nodes $\boldsymbol{V} = \{v_i\}_{i=1}^{|V|}$, and their interactions are represented as edges $\boldsymbol{E}$ between nodes. Our goal is to produce a refined III network $G'_{III} = (\boldsymbol{V}, \boldsymbol{E}')$ on the same set of nodes but with a new set of edges $\boldsymbol{E}'$, reflecting real interactions among isoforms.

We refine the III network according to isoforms' neighbors in the current III network and the isoform co-expression relationship. For each isoform $v_i \in \boldsymbol{V}$, these two types of information are represented as a node feature vector $\boldsymbol{x}_i$. The details of this representation are described as follows.

(i) Isoform neighborhood: The neighborhood of node v_i is defined as a set of nodes visited by a series of random walks starting from v_i, denoted as $\boldsymbol{N}_i$ (50). The isoforms with similar neighborhoods should share similar node feature vectors as they have similar interacting partners. To characterize this relationship, we specify the following objective function: For a node v_i, the objective seeks to correctly predict $\boldsymbol{N}_i$ from their node feature vectors. As the neighborhood relationship is not certainly bidirectional based on its definition, we use a context vector $\boldsymbol{x}'_i$ to represent each node when it is treated as the prediction target. Thus, predicting the neighborhood is modeled as the conditional likelihood given by a softmax unit parameterized by the products of node vectors. The objective is to minimize the following negative log likelihood through the updating of node feature and context vectors:

$$\begin{aligned} L_{nb} &= -\sum_{i=1}^{|V|}\sum_{j\in \boldsymbol{N}_i} \log p(v_j|v_i) \\ &= -\sum_{i=1}^{|V|}\sum_{j\in \boldsymbol{N}_i} \log \frac{\exp(\boldsymbol{x}'^{\mathsf{T}}_j \boldsymbol{x}_i)}{\sum_{v_k\in \boldsymbol{V}} \exp(\boldsymbol{x}'^{\mathsf{T}}_k \boldsymbol{x}_i)}. \end{aligned} \tag{6}$$

As the computation of the full softmax is expensive, we approximate the objective using negative sampling (51). For each node v_j in the neighborhood of node v_i, we sample a set of non-neighborhood nodes, $\boldsymbol{R}_{ij} \subseteq \boldsymbol{V} - \boldsymbol{N}_i$. Thus, the task becomes to distinguish node v_j from nodes in $\boldsymbol{R}_{ij}$. Then, the above objective can be formulated as the following 'network neighborhood loss':

$$L_{nb} = -\sum_{i=1}^{|V|}\sum_{j\in \boldsymbol{N}_i} (\log \sigma(\boldsymbol{x}'^{\mathsf{T}}_j \boldsymbol{x}_i) - \sum_{k\in \boldsymbol{R}_{ij}} \log \sigma(\boldsymbol{x}'^{\mathsf{T}}_k \boldsymbol{x}_i)). \tag{7}$$

(ii) Co-expression relationship: Co-expressed isoforms are usually those involved in common biological processes and thus may have common interacting partners (52). As introduced in the 'Construction of tissue-specific datasets' section, the tissue-specific co-expression network $G_{EXP} = (\boldsymbol{V}, \boldsymbol{R})$ is constructed on the same set of nodes $\boldsymbol{V}$ as the tissue-specific III network, with a set of weighted edges $\boldsymbol{R}$ where the weight of edge r_{ij} between nodes (v_i, v_j) reflects the expression correlation between two corresponding isoforms. Then, the following 'co-expression loss' introduces a regularization for node feature vectors under the squared Euclidean distance, weighted by the edge weights of the co-expression network, which encourages similar node feature vectors to be shared by co-expressed isoforms:

$$L_{coe} = \sum_{i=1}^{|V|}\sum_{j=1}^{|V|} r_{ij} ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2. \tag{8}$$

To predict interactions from node features, we built a binary classifier. Specifically, for a pair of nodes (v_i, v_j), we first combine their feature vectors using the element-wise multiplication, i.e. $\boldsymbol{x}_i \circ \boldsymbol{x}_j$, which is a commonly used operation in modeling the symmetric relations from vector representations (53–55). Then, the sigmoid function is applied on the weighted summation of the combined representation's dimensions, which outputs a score in the range [0, 1], indicating how likely the interaction happens between the two corresponding isoforms:

$$z_{ij} = \sigma(\boldsymbol{w}^{\mathsf{T}}(\boldsymbol{x}_i \circ \boldsymbol{x}_j)), \tag{9}$$

where $\boldsymbol{w}$ is a vector of trainable parameters which learns to weight the contribution of different dimensions of node feature vectors. To train the weight vector, we treat links in the current III network as labels and apply the same weighted cross entropy introduced in Equation 5 as the 'link prediction loss':

$$L_{lp} = -\sum_{i=1}^{|V|}\sum_{j=1}^{|V|} \hat{z}_{ij}(e_{ij}\log(z_{ij}) + (1-e_{ij})\log(1-z_{ij})), \tag{10}$$

where e_{ij} is a binary indicator for the link between node i and node j in the current III network and $\hat{z}_{ij}$ is a constant assigned by z_{ij}. Feature vectors are also adjusted to adapt to the weight vector based on the objective function, which facilitates the training process.

Mutual regularization and the joint learning objective

The key idea, which forms the cornerstone of this work, is to establish the connection between the two tasks, IFP and IIIP. Inspired by the graph regularizer (56) recently proposed for training neural networks with the help of static graphs, we propose a mutual regularizer for both modules (Figure 2B) that uses edges in the current III network to regularize the learning process of the functional predictor and also encourages dynamic adjustments in the III network consistent with the prediction results made by the functional module:

$$L_{mut} = -\sum_{i=1}^{|V|}\sum_{j=1}^{|V|} z_{ij}(m - ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2^2), \tag{11}$$

where $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ are functional feature vectors of the corresponding isoforms of node i and j, defined in Equation 1, and m is a predefined margin. Intuitively, this 'mutual regularization loss' encourages the functional predictor to learn similar functional feature vectors for two isoforms if

they are connected in the current III network. On the other hand, if two isoforms have similar functional feature vectors, i.e. the squared Euclidean distance over them is smaller than the predefined margin m, a larger prediction score of their interaction is encouraged.

To sum up, the joint objective of FINER is to minimize the following loss function:

$$L = \lambda_1 L_{fp} + \lambda_2 L_{mut} + \lambda_3 L_{nb} + \lambda_4 L_{coe} + \lambda_5 L_{lp}, \quad (12)$$

where λ_1, λ_2, λ_3, λ_4, and λ_5 are the balancing hyper-parameters.

Algorithm 1: Learning algorithm of FINER

Input : Isoform sequences, $\boldsymbol{s}$; Conserved domains, $\boldsymbol{d}$; Initial functional labels, $\boldsymbol{y}$; Initial III network, $G_{III} = (\boldsymbol{V}, \boldsymbol{E})$; Co-expression network, $G_{EXP} = (\boldsymbol{V}, \boldsymbol{R})$.

Output: Functional predictor with parameters Θ; Refined III network $G'_{III} = (\boldsymbol{V}, \boldsymbol{E}')$.

Initialize parameters $\boldsymbol{\Theta}$, $\boldsymbol{x}$, $\boldsymbol{x}'$ and $\boldsymbol{w}$, $G'_{III} = G_{III}$;

while *not converged* **do**

- Sample batches for functional predictor by $\boldsymbol{E}'$;
- **for** *each batch* **do**
  - Update $\boldsymbol{\Theta}$ by Equation (13);
- **for** *each isoform* i **do**
  - Make inference on $\boldsymbol{h}_i$ and $\boldsymbol{p}_i$;
- Update $\boldsymbol{y}$ under the MIL constraints;
- Sample batches for III refinement module by Node2vecWalk(G'_{III});
- **for** *each batch* **do**
  - Update $\boldsymbol{x}$, $\boldsymbol{x}'$, and $\boldsymbol{w}$ by Equation (14);
- **for** *each node pair* (v_i, v_j) **do**
  - Make inference on z_{ij};
- Update $\boldsymbol{E}'$;

Training procedure of FINER

To learn general functional knowledge from sequences and domains, we first pretrain the function prediction module using a large number of proteins retrieved from the SwissProt database (57) as done in DIFFUSE (12). In this study, we collected 98 400 eukaryotic (other than human) protein sequences with GO annotations from the SwissProt database. Conserved domain data were retrieved accordingly using the same method described before. The 'binary cross entropy loss' defined in Equation 4 is used to pretrain the functional predictor.

Next, the function prediction module and the III module are alternately trained with the isoform data, until convergence. The pseudocode for the learning algorithm is given in Algorithm 1, and its basic ideas are sketched below.

(i) Training the function prediction module: In the functional module training phase, parameters of the functional predictor $\boldsymbol{\Theta}$ are updated by minimizing the weighted summation of two components in the loss function with the stochastic gradient descent method:

$$\min_{\boldsymbol{\Theta}} \lambda_1 L_{fp} + \lambda_2 L_{mut}. \quad (13)$$

To speedup learning, isoforms connected in the current III network are sampled into the same training batch. After each training phase of the functional module, the inference is performed for all isoforms on their functional feature vectors $\boldsymbol{h}$ and functional predictions $\boldsymbol{p}$. Under the MIL setting, for each GO term, the labels of all instances in positive bags are updated according to the following criteria: (i) Instances with prediction scores above the predefined threshold are assigned with positive labels, while the others are assigned with negative labels. (ii) For each positive bag, if all its instances are assigned with negative labels, we select the instance with the largest positive prediction score in the bag as positive. The updated labels are used for training in subsequent iterations.

(ii) Training the III network refinement module: In the III module training phase, node vectors and weight parameter $\boldsymbol{w}$ are updated by minimizing the weighted summation of four components in the loss function with the stochastic gradient descent method:

$$\min_{\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{w}} \lambda_2 L_{mut} + \lambda_3 L_{nb} + \lambda_4 L_{coe} + \lambda_5 L_{lp}, \quad (14)$$

After each training phase of the III module, the inference is performed for each node pair (v_i, v_j) on the link prediction z_{ij}, based on which, edges in the current III network are updated to obtain a refined III network. Due to the noisy nature of tissue-specific PPIs, we would like to denoise the existing interactions while discovering *de novo* interactions. Therefore, unlike the label update procedure in the functional module, edge update here does not consider bag-instance constraints. The following criteria are considered when updating edges instead: (i) In the refined III network, edges are set between nodes if the corresponding link prediction scores are above the predefined threshold. (*ii*) To guarantee the inclusion of interaction information for each isoform, the top 10 edges with the largest link prediction scores associated with each node are also included in the refined III network. Edges in the refined network are then used for regularizing the functional module in subsequent iterations.

RESULTS

Prediction of tissue-specific isoform functions

We applied FINER to predict tissue-specific functions of isoforms on the human tissue datasets, including 12 major tissues and three brain sub-tissues. The prediction procedure, together with the calculation of prediction accuracy, are described below:

(i) Dataset partition: For each tissue, we randomly partition its isoforms into training, validation and test sets with the proportions of 70%, 10% and 20%, respectively. Hyper-parameters of the models are manually tuned based on model performance on the validation data (see Supplementary Table S3 for the calibrated hyper-parameter values). The validation data are finally merged with the training data to train a model for performance evaluation on the test data. To avoid potential

information leak (i.e. different components of the partition share isoforms with very similar sequences and thus similar functions), we first require that isoforms of the same gene are partitioned into the same set. In addition, since the function prediction module is pretrained with the SwissProt protein sequences from different eukaryotes and there are closely related paralogous genes in the human genome, we consider clusters of orthologous groups (COGs) defined in the EggNOG database up to the level of eukaryotes (58) (note that such COGs also include many paralogous genes) to prevent closely related homologous genes from being split among different sets. In other words, all genes of the same COG are required to be partitioned together. In addition, all (non-human) SwissProt proteins belonging to COGs that contain (human) genes in the test set are excluded from the pretraining phase.

(ii) Prediction accuracy evaluation: As the ground-truth of isoform functions is generally unavailable, we adopt the widely used alternative evaluation strategy at the gene level (8,9,11,12), with the rationale that if the functions of isoforms are correctly predicted, their gene functions should be correctly predicted automatically. Hence, for each GO term, a prediction score for each gene is generated by taking the maximum prediction score among its isoforms, and the performance is measured by comparing the gene-level prediction with the ground-truth. Both the area under the receiver operating characteristics curve (AUC) and the area under the precision-recall curve (AUPRC) are used to evaluate the performance for each GO term. To make comparisons across different datasets fairly, we unify the AUPRC baseline as 0.1 for all GO terms as done in (11,12).

To evaluate the effect of III refinement on functional prediction, we compare the performance of FINER with that of FINER without III refinement as well as with FINER without co-expression regularization in the III refinement module. Figure 3 summarizes the average AUC and AUPRC values over all the GO terms in each tissue. On average, FINER achieves improvements of 5.80% and 21.5% over FINER without III refinement in terms of AUC and AUPRC, respectively, on the major tissue datasets, as well as improvements of 5.16% and 15.1% in terms of AUC and AUPRC, respectively, on the brain datasets. In addition, FINER achieves improvements of 1.94% and 4.28% over FINER without co-expression regularization in terms of AUC and AUPRC, respectively, on the major tissue datasets, as well as improvements of 1.51% and 7.37% in terms of AUC and AUPRC, respectively, on the brain datasets. The learning curves of the function prediction module in Figure 3C clearly demonstrate that the performance of the module benefited from the refinement of III networks, i.e., the performance of the function prediction module clearly gets better after each III network update, until convergence (see Supplementary Figure S1 for learning curves on all the other tissues).

A concrete example is shown in Figure 4. Isoform NM_000660 is the single isoform of gene TGFB1. According to GO annotations, TGFB1 is labeled as having the heart-specific function of cardiac chamber development (GO:0003205). Without applying III refinement, NM_000660 is predicted to have the function of GO:0003205 with a score of only 0.571, which is just at the boundary between having or not having the function. Meanwhile, most of its interacting partners in the initial III network are predicted as not having the function. In contrast, when applying III refinement, NM_000660 is predicted to be interacting with isoforms that are predicted as having the function, and NM_000660 itself is predicted as having the function with a high score of 0.870.

Comparison with the existing methods

We further make comprehensive comparisons between the functional prediction performance of FINER and that of several state-of-the-art methods with different objectives, including two recent isoform function prediction methods DIFFUSE (12) and DisoFun (14), a tissue-specific protein function prediction method OhmNet (20) and a general biological network refinement method NE (59). Note that three isoform function prediction methods, DisoFun, ISOGO (16) and IsoResolve (15) have been published in the literature after DIFFUSE. Although these methods have not been compared with DIFFUSE directly on the same dataset, their reported overall performance all seem to be worse than that of DIFFUSE. Among the methods, DisoFun adopted a more strict evaluation metric in its performance evaluation. Moreover, it also considers PPI information similar to FINER. We, therefore, choose to include DisoFun in the comparison here.

DisoFun predicts isoform functions using a matrix factorization approach based on isoform expression profiles, where PPIs are used to perform a gene-level regularization. OhmNet first learns protein embeddings from different tissue-specific PPI networks, taking into consideration the dependence between tissues. Independent function classifiers are then trained to predict tissue-specific protein functions at the gene level. NE has been successfully used to denoise tissue-specific PPIs in the literature (59). In the study, we apply NE to denoise initial III networks in our datasets. To compare the effect of their results on enhancing isoform function prediction with that of our refined III networks, we provide FINER with NE's denoised III networks and keep them fixed throughout functional prediction. We denote this model as $\text{FINER}_{\text{fixed}}$+NE. To make the comparisons more clear, we include the performance of FINER without III refinement here, denoted as $\text{FINER}_{\text{fixed}}$+RAW. As shown in Table 1, FINER improved over the best performance of the three isoform/protein function prediction methods (i.e. OhmNet, DisoFun and DIFFUSE) on both the major tissue and brain datasets by 6.94% and 6.57%, respectively, in terms of average AUC, as well as 11.62% and 21.1%, respectively, in terms of average AUPRC. We observe that the standard deviations (SDs) in FINER's performance across different tissues are generally smaller than the other methods. Moreover, a comparison among FINER, $\text{FINER}_{\text{fixed}}$+NE and $\text{FINER}_{\text{fixed}}$+RAW demonstrates that FINER acquires larger performance gains from our iteratively refined III networks than the denoised III networks of NE, even though they are still better than the initial networks.

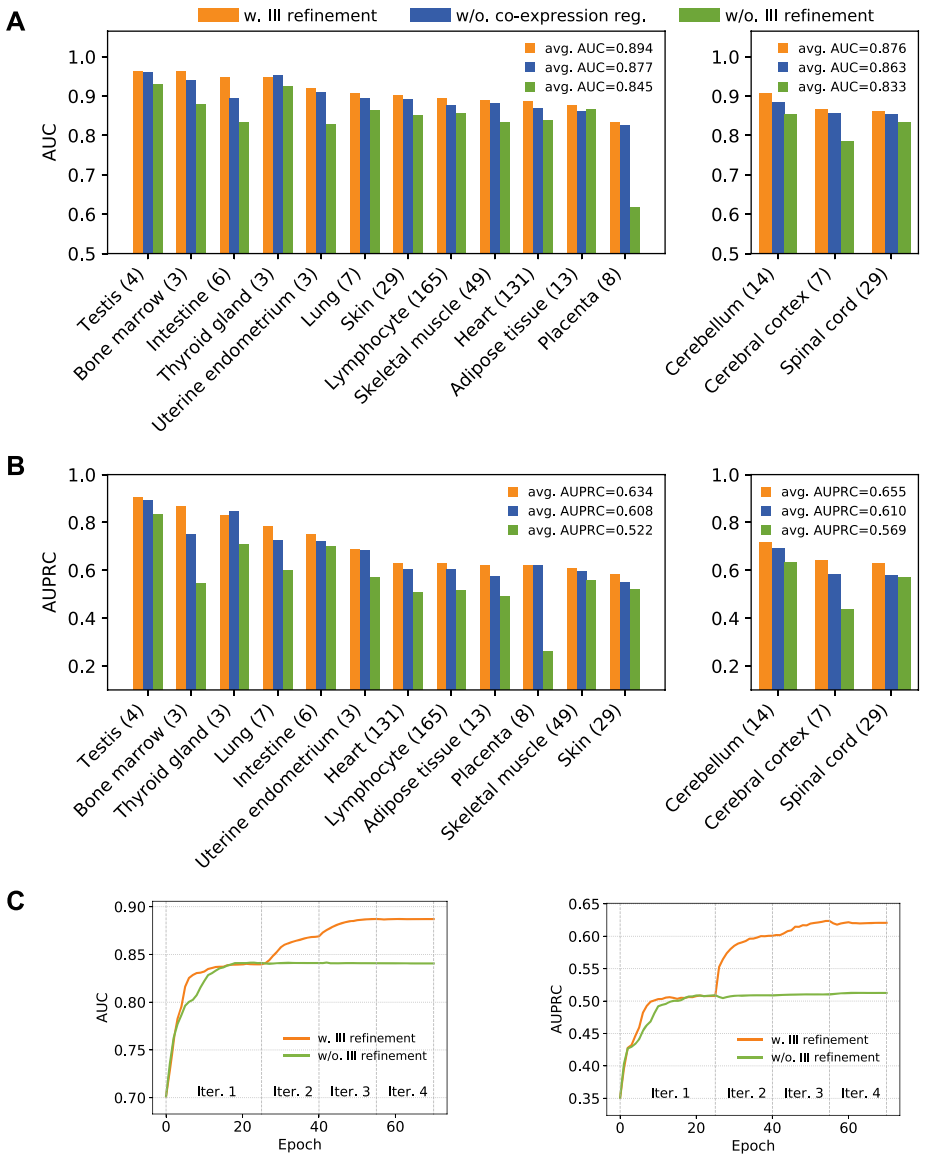

Figure 3. (**A**) Comparison of functional prediction performance measured by the average AUC over GO terms on each dataset, between FINER (orange), FINER without co-expression regularization (blue), and FINER without III refinement (green). The number of GO terms associated with each tissue is noted after the name of the tissue. (**B**) Comparison of functional prediction performance measured by the average AUPRC. (**C**) Learning curves of the function prediction module with (orange) and without (green) III refinement on the Heart tissue dataset in terms of both AUC and AUPRC.

Due to the lack of tissue-specific interaction ground-truth, we measure the consistency between our refined III networks and the results of a state-of-the-art III prediction method. TENSION (24) is compared here as it is the most recent tissue-specific III prediction method. For each tissue, a core subnetwork is extracted from the predicted III network of each method, which is induced by the set of isoforms whose genes are associated with the tissue-specific functions. The Jaccard index is used to measure the similarity between the subnetworks generated by the two methods for each tissue. As shown in Supplementary Table S4, the average of Jaccard indexes across all tissues is 0.332, and they are all significantly larger than the expected ones if two networks are randomly (and independently) generated with the same sets of nodes and number of edges as in the networks predicted by FINER and TENSION (under the column *E*[Jaccard index] in Supplementary Table S4). The moderate similarity between the core parts of III networks on most of the tissues suggests that the III predictions made by the two methods are perhaps informative.

Consistency between the predicted functions of isoforms and their tissue specificity

We validate our isoform-level predictions by investigating their consistency with the tissue specificity of isoforms. It is well-known that the expression of genes is usually tissue-specific. Previous studies have shown that in a certain tissue, the highly expressed genes are usually associated with functions specific to the tissue (60). For example, genes with

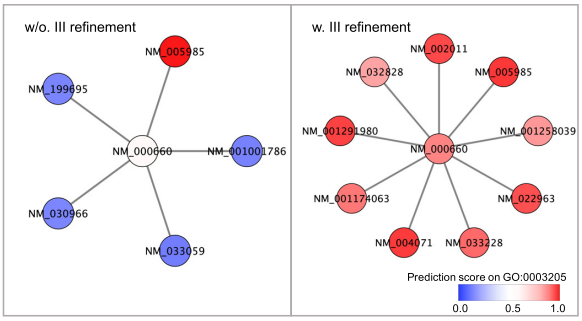

Figure 4. Illustration of the interactions and functional prediction scores on the term GO:0003205 of isoform NM_000660 in both the initial III network and the refined III network. Red nodes represent isoforms predicted as having the function, while blue nodes represent isoforms predicted as not having the function.

elevated expression in skin are associated with functions related to the barrier function, skin pigmentation and hair development, while genes elevated in liver are associated with metabolic processes and glycogen storage (41). As isoforms are actual function carriers, we expect that this relationship also remains true at the isoform level. That is, the set of isoforms elevated in a tissue should be enriched with the corresponding tissue-specific functions. Thus, we quantify the expression specificity of each isoform in a given tissue by the fold change of its mRNA level in the tissue over the average level in other tissues. For each tissue, a set of 'tissue enhanced isoforms' are selected from the test set based on the 'tissue enhanced' criteria same as those in the Materials and methods section. To generate functional annotations of isoforms on each GO term, we binarize the corresponding prediction scores by applying the threshold that optimizes the F1 score with respect to the gene-level ground-truth. Then, Fisher's exact test is performed to test each tissue-specific GO term's enrichment in the set of tissue enhanced isoforms. The multiple testing correction with false discovery rate (FDR) controlling is applied to the P values. Figure 5A shows the fractions of GO terms that are enriched in the tissue enhanced isoform sets of each tissue. Enrichment (i.e. P(corrected) ≤ 0.05) is found in 91.4% (385 out of 421) of the GO terms on the major tissue datasets and 84.0% (42 out of 50) on the brain datasets. These results confirm that the consistency between (predicted) functions and tissue-specific expressions remains at the isoform-level.

We further investigate whether our functional predictions differentiate tissue enhanced isoforms from non-tissue enhanced ones in functional genes. Specifically, for each tissue-specific GO term, we consider only the genes that are associated with the term, and divide isoforms of these genes into two sets, namely, a set of 'tissue enhanced isoforms' and a set of 'non-tissue enhanced isoforms' based on the same criteria as before. Note that either the tissue enhanced isoform set or the non-tissue enhanced isoform set could be empty for a GO term. If this happens, the corresponding GO term is then ignored in the analysis. We compare the fold enrichment of a GO term in both sets. The higher the fold is, the more significant enrichment is found in a set. As shown in Figure 5B, the one-sided Wilcoxon test exhibits significant differences of GO enrichment between such two sets of tissue enhanced and non-tissue enhanced isoforms. The results suggest that FINER was able to identify tissue-enhanced isoforms from genes with tissue-specific functions and assign these functions to such isoforms.

Consistency between the highest connected isoforms and isoform protein-level expression

Previous studies have found that in a given tissue, the isoform of each gene with the most interacting partners usually shows a higher expression level than other isoforms of the same gene and is more likely to play functional roles in the tissue. This observation is consistent across a variety of tissues at both the transcript level and the protein level (21,61). To check the validity of this observation in our refined III networks, we identify the highest connected isoform (HCI) of each MIG in different tissues, where the HCI is defined as the isoform of each MIG that has the highest degree in the III network of a given tissue. An independent dataset for tissue-specific protein-level expression of isoforms was then collected from Wang *et al.* (62). For each tissue, the dataset lists a set of isoforms that are detected at the protein level by mass spectrometry. Due to its low sequence coverage, most genes have only one detected isoform in each tissue, which usually is the highest expressed isoform at the protein level. Ideally, the HCIs of each MIG in different tissues should be the isoforms that have protein expression evidence in the corresponding tissues. As shown in Table 2, the numbers of MIGs whose HCIs in a tissue are detected at the protein level, denoted as N_{FINER}, are significantly higher than the expected numbers of MIGs (N_{chance}) if their HCIs in the tissues are randomly chosen and detected at the protein level. We repeat the same experiment on the III predictions of TENSION. The numbers of MIGs with HCIs in the III networks predicted by TENSION that are detected at the protein level, denoted as N_{TENSION}, are not as significantly different from N_{chance} as the ones of FINER. These results confirm the above observation in our refined III networks.

We also consider more isoforms of each MIG that have high degrees in the predicted IIIs, and found that the numbers of MIGs whose third highest connected isoforms (third HCIs), second highest connected isoforms (second HCIs) or HCIs obtained by FINER are detected at the protein level monotonically increase in all tissues (Supplementary Tables S5 and S6). This suggests that the isoforms detected at the protein levels tend to have higher degrees in the III networks predicted by FINER. However, this monotonicity property does not always hold in the III networks predicted by TENSION.

Consistency between interactions of isoforms and their subcellular localization

Subcellular localization of isoforms determines the environments where they operate. Therefore, subcellular localization plays a significant role in controlling the availability of interacting partners of isoforms and further influencing their functions (63). Thul *et al.* (64) also discovered that interactions among proteins within the same or connected cell organelles are more likely to happen compared to isoforms between disconnected organelles. Inspired by this finding,

Table 1. Comparison of functional prediction performance between FINER and some existing state-of-the-art methods

Method	Major tissue datasets		Brain datasets	
	AUC (SD)	AUPRC (SD)	AUC (SD)	AUPRC (SD)
OhmNet	0.751 (0.099)	0.431 (0.196)	0.743 (0.123)	0.423 (0.223)
DisoFun	0.805 (0.188)	0.460 (0.236)	0.770 (0.214)	0.419 (0.242)
DIFFUSE	0.836 (0.134)	0.568 (0.208)	0.822 (0.190)	0.541 (0.256)
$FINER_{fixed}$+RAW	0.845 (0.102)	0.522 (0.185)	0.833 (0.149)	0.569 (0.254)
$FINER_{fixed}$+NE	0.859 (0.104)	0.540 (0.178)	0.841 (0.131)	0.577 (0.242)
FINER	**0.894** (0.080)	**0.634** (0.158)	**0.876** (0.115)	**0.655** (0.234)

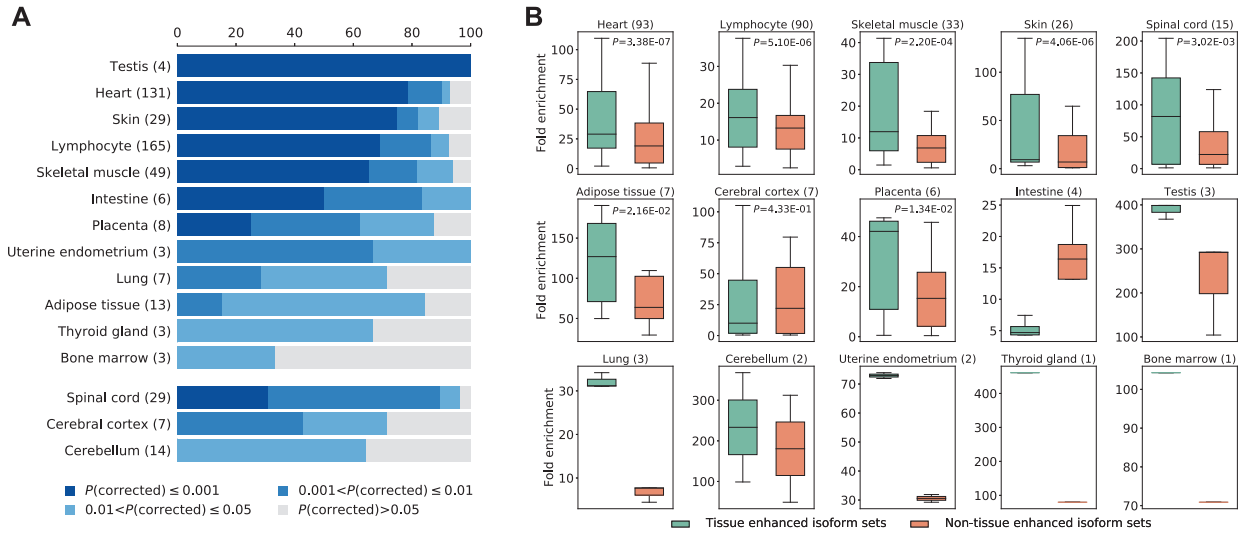

Figure 5. (**A**) The fractions of GO terms that are enriched in the set of tissue enhanced isoforms of each tissue. Different levels of enrichment are colored differently. (**B**) Fold enrichment of GO terms in sets of tissue enhanced isoforms (green) and sets of non-tissue enhanced isoforms (orange), where for each GO term, only isoforms of genes associated with the term are considered. The one-sided Wilcoxon test is performed on the results of each tissue with at least 5 GO terms (numbers of GO terms are noted in the titles) included in this analysis to test the significance of the difference in GO enrichment between tissue enhanced and non-tissue enhanced isoform sets.

Table 2. The numbers of MIGs whose HCIs are detected at the protein level in each tissue. Comparisons are made between HCIs of III networks predicted by FINER and those predicted by TENSION

Tissue	# of MIGs	N_{chance}	N_{FINER} (*P*-value)	$N_{TENSION}$ (*P*-value)
Heart	316	121	217 (2.35E-14)	177 (9.12E-14)
Lung	272	103	150 (1.72E-13)	121 (1.91E-03)
Lymphocyte	399	157	252 (4.33E-14)	196 (5.04E-06)
Placenta	599	228	345 (5.95E-14)	236 (1.29E-01)
Testis	754	287	421 (5.80E-14)	313 (5.40E-03)
Thyroid gland	466	180	287 (3.46E-14)	196 (2.37E-02)
Uterine endometrium	235	91	147 (1.98E-14)	124 (8.47E-07)

we collected some data of isoform subcellular localization from Uhlén *et al.* (41), in which isoforms are annotated with locations predicted from their sequences: soluble (intracellular isoforms), membrane-spanning or secreted. We then examine the enrichment of interactions among isoforms in the same or between different subcellular locations. Figure 6A and C show that, when considering isoforms of SIGs alone, a significant enrichment of interactions is always found between isoforms within the same subcellular location but rarely found between those in different locations, no matter the initial or refined III networks are used. On the other hand, an enhancement of this trend (i.e., enrichment of interactions between isoforms within the same subcellular location) can be seen in refined III networks compared with the initial ones in the Heart, Skeletal muscle, Skin and Thyroid gland tissues. In contrast, Figure 6B and D show that, when considering only isoforms in MIGs, more enrichment of interactions between isoforms in different locations is found in the initial III networks, but the above trend observed in SIGs still remains true in the refined III networks. A plausible conclusion from these observations is that our results concerning the isoforms of SIGs show consistency with the previous findings (64). In other words, even though isoforms at the same subcellular location may not belong to the same or connected organelles, it is conceivable that interactions could be more likely to happen between these isoforms compared to isoforms in different locations, as found consistently in our observations.

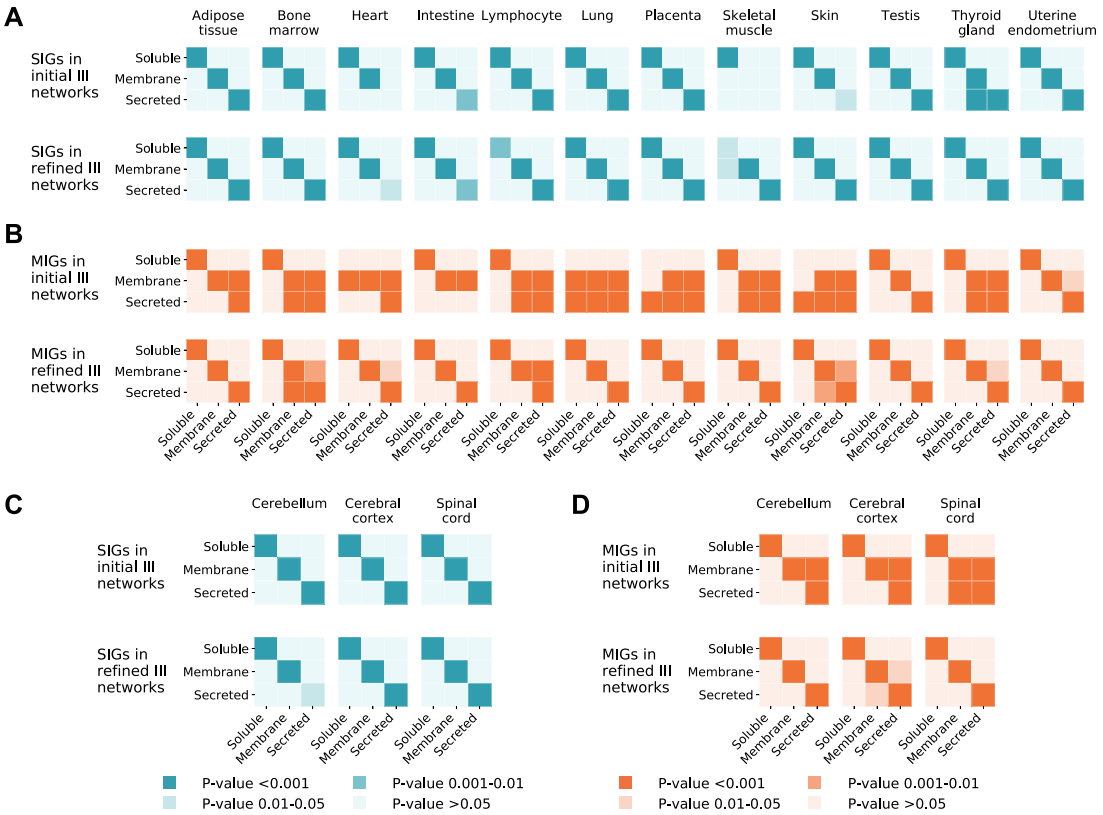

Figure 6. Heat maps describe the probability (measured by the FDR-corrected *P* value for the binomial test) of observing at least as many isoforms in a given location (*y* axis) by chance, given the location of each isoform's interaction partner (*x* axis). (**A**) Comparison of the above probabilities between the isoforms of SIGs in the initial III networks and those in the refined III networks for the 12 major tissues. (**B**) The same comparison for the isoforms of MIGs in the 12 major tissue datasets. (**C**) The same comparison for the isoforms of SIGs in the 3 brain sub-tissue datasets. (**D**) The same comparison for the isoforms of MIGs in the three brain sub-tissue datasets.

However, since different isoforms of MIGs can be localized differently, initializing III networks based on PPIs may introduce many false interactions between isoforms from different locations. Through III network refinement, real interactions are revealed and thus the expected trend is recovered.

Differentiating functions of isoforms with different localization

It is commonly found that a single gene can encode isoforms with different subcellular localization (41), which suggests the potential functional differences between them. We test if FINER can correctly differentiate the functions of isoforms from the same gene, measured in terms of consistency with their localization. We focus on a set of subcellular location enriched GO terms. Specifically, for each subcellular location, we consider the set of genes that encode isoforms located there. Then, GO terms that are enriched in the gene set are selected as the location enriched terms through GO enrichment analysis. For each selected GO term and the corresponding subcellular location, we consider MIGs that are associated with the GO term and encode isoforms with different localization containing at least one isoform in the considered location. Isoforms with prediction scores greater than the background of their genes are annotated with the GO term, where the background of a gene is defined as the average prediction score of all its isoforms. The Jaccard index is used to quantify the agreement that isoforms annotated with a GO term are also located in its corresponding subcellular location.

Figure 7 shows that the predictions of FINER achieve a higher consistency with isoform subcellular localization than those of DIFFUSE and DisoFun in 6 out of the 7 considered GO terms, while DIFFUSE generally outperforms DisoFun. This result suggests that isoform localization information resides in the refined IIIs and isoform sequences may help FINER differentiate the functions of isoforms with different localization.

Case studies with literature support

We finally perform a literature search for experimental evidence to support the predictions of FINER. In particular, some evidence concerning the tissue specificity of isoforms and their functions is collected from the literature for three genes. The first gene FYN encodes isoforms FynB and FynT. Whereas FynB accumulates highly in the brain, FynT is expressed predominantly in lymphocytes. Accordingly, FynT but not FynB serves a tissue-specific function in T-cell activation (65). This evidence is consistent with the relationship between function and tis-

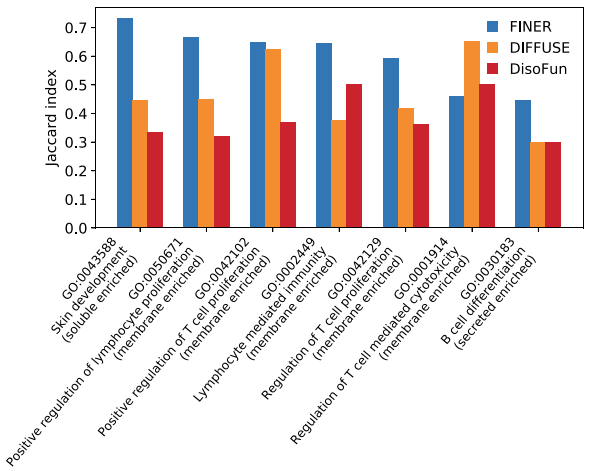

Figure 7. Comparison between FINER (blue), DIFFUSE (yellow) and DisoFun (red) in terms of consistency between their predictions on location enriched GO terms and subcellular localization of isoforms, where the consistency is measured by the Jaccard index.

sue specificity analyzed in Figure 5. FINER correctly predicted the tissue-specific functions of both isoforms. For the lymphocyte-specific GO term 'Regulation of T cell activation (GO:0050863)', FynT has a prediction score 1.3 times the background score of its gene, while FynB only has a score 0.6 times the background. The second gene PPARG involves two isoforms with different tissue specificity. The expression of PPARG2 is restricted mainly to the adipose tissue, whereas PPARG1 is expressed in the adipose tissue and many other tissues. PPARG2 can stimulate the formation of adipocytes (fat cells). However, evidence shows that PPARG1 has no or reduced ability to induce adipogenesis (66,67). Our predictions on the GO term 'Fat cell differentiation (GO:0045444)' accord with the experimental observation. That is, PPARG2 has a prediction score 1.2 times the background, while the score of PPARG1 is 0.8 times the background. The last example concerns three isoforms encoded by gene TITIN. While the isoform N2A is the major isoform of TITIN expressed in skeletal muscles, N2B and N2BA are major TITIN isoforms expressed in the heart, whose expression ratio is related to human heart diseases (68). The III predictions of FINER show that N2A is the highest connected isoform in the refined III network of skeletal muscles, while isoforms N2B and N2BA are the highest connected ones in that of the heart (Supplementary Figure S2), consistent with the relationship analyzed in Table 2.

In addition, we are able to find some experimental evidence that indirectly supports the predictions of FINER concerning the tissue-specific functions of isoforms in four genes. The evidence is collected by the following procedure. For each tissue, among all the MIGs associated with at least one GO term specific to the given tissue, the MIGs whose HCIs have the top five highest degrees (among all HCIs) are selected. An exhaustive literature search is then performed against the selected MIGs. Information about tissue-specific functions and corresponding predictions of FINER concerning the isoforms in these MIGs is listed in Supplementary Table S7 (along with the cases discussed in the previous paragraph). Details of the functional evidence are discussed below. The gene CD40 plays an important signal transduction role in the pathway responsible for B-cell growth and differentiation. Compared with the isoform NM_001250 encoded by CD40, isoform NM_152854 lacks the transmembrane domain, which makes it signal-nontransducible (69). Consistently, for the lymphocyte-specific GO term 'Positive regulation of B cell differentiation (GO:0045579)', FINER predicts NM_001250 to have a score 1.2 times the background score of its gene, while NM_152854 has a score 0.9 times the background. The gene WT1 regulates gonad development through activating the expression of the gene SF1. However, the presence of the KTS motif in WT1 isoforms hinders their interaction with the SF1 promoter (70). Accordingly, among the six isoforms encoded by WT1, FINER gives the three isoforms lacking the KTS motif (NM_001198551, NM_024424, NM_000378) higher scores on the testis-specific GO term 'Male gonad development (GO:0008584)' than the other three isoforms with the KTS motif, as shown in Supplementary Table S7. In the adipose tissue, the gene GATA2 acts as a negative regulator of adipocyte proliferation through interaction with FOG proteins, where the interaction relies on the contact of their zinc fingers (71). Between the two isoforms encoded by GATA2, NM_001145662 lacks a zinc finger compared with NM_032638. Accordingly, FINER predicts NM_032638 to have a score 1.2 times the background on the GO term 'Negative regulation of fat cell proliferation (GO:0070345)', while the score of NM_001145662 is 0.8 times the background. The gene NKX2-5 acts as a transcription factor during the thyroid gland development (72). Among the three isoforms NM_004387, NM_001166175 and NM_001166176 of NKX2-5, the DNA-binding domain is missing in NM_001166175 and NM_001166176 due to alternative splicing. Correspondingly, on the thyroid gland-specific GO term 'Thyroid gland development (GO:0030878)', the isoform with the DNA-binding domain (NM_004387) is predicted with a higher score than the other two as shown in Supplementary Table S7.

DISCUSSION

Isoform function prediction (IFP) and isoform-isoform interaction prediction (IIIP) are two important problems in studying the diversity of gene products. The close ties between functions and interactions of protein isoforms make the IFP and IIIP problems inherently intertwined. In this work, we presented FINER, a unified framework for solving the two problems jointly. FINER establishes the connection between IFP and IIIP by introducing a joint learning objective, which enables both tasks to benefit from each other. We apply FINER to predict tissue-specific isoform functions and interactions on two datasets, which contain 12 major tissues and three brain sub-tissues of human, respectively. FINER outperforms the state-of-the-art methods across different tissue datasets and provides isoform function and interaction predictions that accord with other biological evidence, including isoform tissue specificity and isoform subcellular localization. These results suggest FINER's potential in facilitating the functional ex-

ploration of (individual) isoforms and their roles in diverse human tissues and diseases.

There are several directions for future work. First, the relationship between tissues is not considered in FINER. The reason is that the tissues studied in this work are relatively independent from each other. If tissue-specific functional terms and well-characterized RNA-seq data are available for a wider range of tissues in the future, the dependence between tissues can be considered and transferring functional knowledge between closely related tissues can be explored in FINER. In addition, FINER converges quickly in practice with several rounds of alternately training the function prediction module and the III refinement module, although we do not have a theoretical proof for its convergence yet. We hope to perform more theoretical analysis in the future. Moreover, although this work focuses on the fundamental problem of isoform function prediction, it would be interesting to see whether FINER can be directly applied to predict isoform–disease associations effectively.

DATA AVAILABILITY

Source code of FINER and tissue-specific datasets used in this study are available at https://github.com/haochenucr/FINER.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

National Key Research and Development Program of China [2018YFC0910404, 2018YFC0910405, 2020YFA0907000 (in part)]; National Natural Science Foundation of China [61772197, 62072435, 31671369, 31770775].

Conflict of interest statement. None declared.

REFERENCES

1. Conesa,A., Götz,S., García-Gómez,J.M., Terol,J., Talón,M. and Robles,M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.
2. Huang,D.W., Sherman,B.T., Tan,Q., Kir,J., Liu,D., Bryant,D., Guo,Y., Stephens,R., Baseler,M.W., Lane,H.C. *et al.* (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.*, **35**, W169–W175.
3. 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.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
4. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
5. Nilsen,T.W. and Graveley,B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457–463.
6. Brett,D., Pospisil,H., Valcárcel,J., Reich,J. and Bork,P. (2002) Alternative splicing and genome complexity. *Nat. Genet.*, **30**, 29–30.
7. Urbanski,L.M., Leclair,N. and Anczuków,O. (2018) Alternative-splicing defects in cancer: splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. *WIRES RNA*, **9**, e1476.
8. Eksi,R., Li,H.-D., Menon,R., Wen,Y., Omenn,G.S., Kretzler,M. and Guan,Y. (2013) Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data. *PLoS Comput. Biol.*, **9**, e1003314.
9. Li,W., Kang,S., Liu,C.-C., Zhang,S., Shi,Y., Liu,Y. and Zhou,X.J. (2014) High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. *Nucleic Acids Res.*, **42**, e39.
10. Luo,T., Zhang,W., Qiu,S., Yang,Y., Yi,D., Wang,G., Ye,J. and Wang,J. (2017) Functional annotation of human protein coding isoforms via non-convex multi-instance learning. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 345–354.
11. Shaw,D., Chen,H. and Jiang,T. (2019) DeepIsoFun: a deep domain adaptation approach to predict isoform functions. *Bioinformatics*, **35**, 2535–2544.
12. Chen,H., Shaw,D., Zeng,J., Bu,D. and Jiang,T. (2019) DIFFUSE: predicting isoform functions from sequences and expression profiles via deep learning. *Bioinformatics*, **35**, i284–i294.
13. Yu,G., Wang,K., Domeniconi,C., Guo,M. and Wang,J. (2020) Isoform function prediction based on bi-random walks on a heterogeneous network. *Bioinformatics*, **36**, 303–310.
14. Wang,K., Wang,J., Domeniconi,C., Zhang,X. and Yu,G. (2020) Differentiating isoform functions with collaborative matrix factorization. *Bioinformatics*, **36**, 1864–1871.
15. Li,H.-D., Yang,C., Zhang,Z., Yang,M., Wu,F.-X., Omenn,G.S. and Wang,J. (2020) IsoResolve: predicting splice isoform functions by integrating gene and isoform-level features with domain adaptation. *Bioinformatics*, **37**, 522–530.
16. Ferrer-Bonsoms,J.A., Cassol,I., Fernández-Acín,P., Castilla,C., Carazo,F. and Rubio,A. (2020) ISOGO: Functional annotation of protein-coding splice variants. *Sci. REP-UK*, **10**, 1069.
17. Vazquez,A., Flammini,A., Maritan,A. and Vespignani,A. (2003) Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.*, **21**, 697–700.
18. Taneri,B., Snyder,B., Novoradovsky,A. and Gaasterland,T. (2004) Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. *Genome Biol.*, **5**, R75.
19. Kulmanov,M., Khan,M.A. and Hoehndorf,R. (2018) DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, **34**, 660–668.
20. Zitnik,M. and Leskovec,J. (2017) Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, **33**, i190–i198.
21. Li,H.-D., Menon,R., Govindarajoo,B., Panwar,B., Zhang,Y., Omenn,G.S. and Guan,Y. (2015) Functional networks of highest-connected splice isoforms: from the Chromosome 17 Human Proteome Project. *J. Proteome Res.*, **14**, 3484–3491.
22. Tseng,Y.-T., Li,W., Chen,C.-H., Zhang,S., Chen,J.J., Zhou,X.J. and Liu,C.-C. (2015) IIIDB: a database for isoform-isoform interactions and isoform network modules. In: *BMC genomics*. Springer, Vol. **16**, p. S10.
23. Ghadie,M.A., Lambourne,L., Vidal,M. and Xia,Y. (2017) Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing. *PLoS Comput. Biol.*, **13**, e1005717.
24. Kandoi,G. and Dickerson,J.A. (2019) Tissue-specific mouse mRNA isoform networks. *Sci. REP-UK*, **9**, 13949.
25. Zeng,J., Yu,G., Wang,J., Guo,M. and Zhang,X. (2019) DMIL-III: Isoform-isoform interaction prediction using deep multi-instance learning method. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 171–176.
26. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
27. Greene,C.S., Krishnan,A., Wong,A.K., Ricciotti,E., Zelaya,R.A., Himmelstein,D.S., Zhang,R., Hartmann,B.M., Zaslavsky,E., Sealfon,S.C. *et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.*, **47**, 569–576.
28. Yeger-Lotem,E. and Sharan,R. (2015) Human protein interaction networks across tissues and diseases. *Front. Genet.*, **6**, 257.
29. Basha,O., Barshir,R., Sharon,M., Lerman,E., Kirson,B.F., Hekselman,I. and Yeger-Lotem,E. (2017) The TissueNet v. 2

database: A quantitative view of protein-protein interactions across human tissues. *Nucleic Acids Res.*, **45**, D427–D431.
30. Kotlyar,M., Pastrello,C., Malik,Z. and Jurisica,I. (2019) IID 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.*, **47**, D581–D589.
31. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
32. Marchler-Bauer,A., Derbyshire,M.K., Gonzales,N.R., Lu,S., Chitsaz,F., Geer,L.Y., Geer,R.C., He,J., Gwadz,M., Hurwitz,D.I. *et al.* (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res.*, **43**, D222–D226.
33. Huntley,R.P., Sawford,T., Mutowo-Meullenet,P., Shypitsyna,A., Bonilla,C., Martin,M.J. and O'Donovan,C. (2015) The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res.*, **43**, D1057–D1063.
34. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., Del-Toro,N. *et al.* (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
35. Rolland,T., Taşan,M., Charloteaux,B., Pevzner,S.J., Zhong,Q., Sahni,N., Yi,S., Lemmens,I., Fontanillo,C., Mosca,R. *et al.* (2014) A proteome-scale map of the human interactome network. *Cell*, **159**, 1212–1226.
36. Chatr-Aryamontri,A., Oughtred,R., Boucher,L., Rust,J., Chang,C., Kolas,N.K., O'Donnell,L., Oster,S., Theesfeld,C., Sellam,A. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, **45**, D369–D379.
37. Keshava Prasad,T. T., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. and et,al. (2009) Human protein reference database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
38. Giurgiu,M., Reinhard,J., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C. and Ruepp,A. (2019) CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res.*, **47**, D559–D563.
39. Menche,J., Sharma,A., Kitsak,M., Ghiassian,S.D., Vidal,M., Loscalzo,J. and Barabási,A.-L. (2015) Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**, 1257601.
40. Leinonen,R., Sugawara,H., Shumway,M. and Collaboration I.N.S.D. (2010) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
41. Uhlén,M., Fagerberg,L., Hallström,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,Å., Kampf,C., Sjöstedt,E., Asplund,A. *et al.* (2015) Tissue-based map of the human proteome. *Science*, **347**, 1260419.
42. Razmara,A., Ellis,S.E., Sokolowski,D.J., Davis,S., Wilson,M.D., Leek,J.T., Jaffe,A.E. and Collado-Torres,L. (2019) recount-brain: a curated repository of human brain RNA-seq datasets metadata. bioRxiv doi: https://doi.org/10.1101/618025, 24 April 2019, preprint: not peer reviewed.
43. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
44. Gremse,M., Chang,A., Schomburg,I., Grote,A., Scheer,M., Ebeling,C. and Schomburg,D. (2010) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.*, **39**, D507–D513.
45. Bossi,A. and Lehner,B. (2009) Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.*, **5**, 260.
46. Li,W., Liu,C.-C., Zhang,T., Li,H., Waterman,M.S. and Zhou,X.J. (2011) Integrative analysis of many weighted co-expression networks using tensor computation. *PLoS Comput. Biol.*, **7**, e1001106.
47. Bengio,Y., Ducharme,R., Vincent,P. and Jauvin,C. (2003) A neural probabilistic language model. *J. Mach. Learn. Res.*, **3**, 1137–1155.
48. Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural. Comput.*, **9**, 1735–1780.
49. He,Y., Shi,J., Wang,C., Huang,H., Liu,J., Li,G., Liu,R. and Wang,J. (2019) Semi-supervised skin detection by network with mutual guidance. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2111–2120.
50. Grover,A. and Leskovec,J. (2016) node2vec: Scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 855–864.
51. Mikolov,T., Sutskever,I., Chen,K., Corrado,G.S. and Dean,J. (2013) Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. pp. 3111–3119.
52. Roy,S., Bhattacharyya,D.K. and Kalita,J.K. (2014) Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC Bioinform.*, **15**, S10.
53. Chen,M., Ju,C. J.-T., Zhou,G., Chen,X., Zhang,T., Chang,K.-W., Zaniolo,C. and Wang,W. (2019) Multifaceted protein–protein interaction prediction based on siamese residual rcnn. *Bioinformatics*, **35**, i305–i314.
54. Jiang,J.-Y., Chen,F., Chen,Y.-Y. and Wang,W. (2018) Learning to disentangle interleaved conversational threads with a siamese hierarchical network and similarity ranking. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. **1**, pp. 1812–1822.
55. Hashemifar,S., Neyshabur,B., Khan,A.A. and Xu,J. (2018) Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*, **34**, i802–i810.
56. Bui,T.D., Ravi,S. and Ramavajjala,V. (2017) Neural Graph Learning: Training Neural Networks Using Graphs. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery WSDM'18, NY, pp. 64–71.
57. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
58. Huerta-Cepas,J., Szklarczyk,D., Heller,D., Hernández-Plaza,A., Forslund,S.K., Cook,H., Mende,D.R., Letunic,I., Rattei,T., Jensen,L.J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.
59. Wang,B., Pourshafeie,A., Zitnik,M., Zhu,J., Bustamante,C.D., Batzoglou,S. and Leskovec,J. (2018) Network enhancement as a general method to denoise weighted biological networks. *Nat. Commun.*, **9**, 3108.
60. Fagerberg,L., Hallström,B.M., Oksvold,P., Kampf,C., Djureinovic,D., Odeberg,J., Habuka,M., Tahmasebpoor,S., Danielsson,A., Edlund,K. *et al.* (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell Proteomics.*, **13**, 397–406.
61. Li,H.-D., Menon,R., Omenn,G.S. and Guan,Y. (2014) Revisiting the identification of canonical splice isoforms through integration of functional genomics and proteomics evidence. *Proteomics*, **14**, 2709–2718.
62. Wang,D., Eraslan,B., Wieland,T., Hallström,B., Hopf,T., Zolg,D.P., Zecha,J., Asplund,A., Li,L.-H., Meng,C. *et al.* (2019) A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.*, **15**, e8503.
63. Yu,N.Y., Wagner,J.R., Laird,M.R., Melli,G., Rey,S., Lo,R., Dao,P., Sahinalp,S.C., Ester,M., Foster,L.J. *et al.* (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.
64. Thul,P.J., Åkesson,L., Wiking,M., Mahdessian,D., Geladaki,A., Blal,H.A., Alm,T., Asplund,A., Björk,L., Breckels,L.M. *et al.* (2017) A subcellular map of the human proteome. *Science*, **356**, eaal3321.
65. Davidson,D., Chow,L., Fournel,M. and Veillette,A. (1992) Differential regulation of T cell antigen responsiveness by isoforms of the src-related tyrosine protein kinase p59fyn. *J. Exp. Med.*, **175**, 1483–1492.
66. Mueller,E., Drori,S., Aiyer,A., Yie,J., Sarraf,P., Chen,H., Hauser,S., Rosen,E.D., Ge,K., Roeder,R.G. *et al.* (2002) Genetic analysis of adipogenesis through peroxisome proliferator-activated receptor γ isoforms. *J. Biol. Chem.*, **277**, 41925–41930.
67. Ren,D., Collingwood,T.N., Rebar,E.J., Wolffe,A.P. and Camp,H.S. (2002) PPARγ knockdown by engineered transcription factors:

exogenous PPARγ2 but not PPARγ1 reactivates adipogenesis. *Gene. Dev.*, **16**, 27–32.
68. Makarenko,I., Opitz,C., Leake,M., Neagoe,C., Kulke,M., Gwathmey,J., Del Monte,F., Hajjar,R. and Linke,W. (2004) Passive stiffness changes caused by upregulation of compliant titin isoforms in human dilated cardiomyopathy hearts. *Circ. Res.*, **95**, 708–716.
69. Tone,M., Tone,Y., Fairchild,P.J., Wykes,M. and Waldmann,H. (2001) Regulation of CD40 function by its isoforms generated through alternative splicing. *Proc. Natl. Acad. Sci.*, **98**, 1751–1756.
70. Wilhelm,D. and Englert,C. (2002) The Wilms tumor suppressor WT1 regulates early gonad development by activation of Sf1. *Gene. Dev.*, **16**, 1839–1851.
71. Jack,B.H. and Crossley,M. (2010) GATA proteins work together with friend of GATA (FOG) and C-terminal binding protein (CTBP) co-regulators to control adipogenesis. *J. Biol. Chem.*, **285**, 32405–32414.
72. Dentice,M., Cordeddu,V., Rosica,A., Ferrara,A.M., Santarpia,L., Salvatore,D., Chiovato,L., Perri,A., Moschini,L., Fazzini,C. *et al.* (2006) Missense mutation in the transcription factor NKX2–5: a novel molecular event in the pathogenesis of thyroid dysgenesis. *J. Clin. Endocrinol. Metab.*, **91**, 1428–1433.