Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Data and text mining

Supplementary file of 'DMIL-IsoFun: predicting isoform function using deep multi-instance learning'

Guoxian Yu [1,2,3]*, Guangjie Zhou [1,2], Xiangliang Zhang [3], Carlotta Domeniconi [4] and Maozu Guo [5]*

[1]School of Software, Shandong University, Jinan 250101, China.
[2]College of Computer and Information Sciences, Southwest University, Chongqing 400715, China.
[3]Computer, Electrical, and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, SA.
[4]Department of Computer Science, George Mason University, Fairfax 22030, USA.
[5]School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China.

*To whom correspondence should be addressed. Email: guomaozu@bucea.edu.cn (M. Guo), guoxian85@gmail.com (G. Yu)

Abstract

1 Details of used datasets

Ten RNA-seq datasets in Table S1 were collected from NCBI official site[1] according to the SRA ID. These RNA-seq datasets are pre-processed for experiments as follows. Quality control software: Fastqc (Patel *et al.*, 2012) and Multiqc (Ewels *et al.*, 2016); data filtering software: TrimGalore (Suchan *et al.*, 2016), build index; and comparison software: Hisat2 (Kim *et al.*, 2019); format conversion software: Samtools (Li *et al.*, 2009); differential Expression Analysis Software: Stringtie (Pertea *et al.*, 2015). We extracted the cut site and exon information from the annotation file, which was used to build an index with the B73 v5 genome assembly data by Hisat2. Then, we converted the original RNA-seq data into fastq files through the commands that come with the NCBI database. Fastqc was then used for quality inspection of all fastq files; then, Multiqc was used to integrate fastqc report files. In quality control, the length threshold was fixed to 20bp, and the default Phred score was 20. After that, Hisat2 was used to map the indexed genomic data with the fastq file after quality control and Samtools was used to sort this mapping file. Finally, Stringtie was used to integrate the generated valid data and featureCounts (Liao *et al.*, 2014) was used to construct the isoform expression data matrix.

[1] https://www.ncbi.nlm.nih.gov/sra/

2 Results on gene level and Human data

Table S2 reports the gene-level results of DMIL-IsoFun and of other compared methods on the Maize genome. Due to the lack of isoform-level GO annotations, we have added a max layer after the GCN stage to aggregate the predicted association probabilities between isoforms and GO terms based on the gene-isoform relation and then used gene-level annotations to train the GCN model. We can find that DMIL-IsoFun again achieves a better performance than other compared methods, which proves effectiveness of DMIL-IsoFun in gene-level. Obviously, our method has made a significant improvement in terms of $AUROC$, S_{min} and F_{max}. In Rankloss, DMIL-IsoFun does not exceed the second-best method (DIFFUSU), which is explainable. In our method, an isoform-level co-expression network is used to propagate annotations, when a negative example appears in the gene bag, it will impact the prediction results of the entire gene and cause negative annotations rank ahead of positive ones, and consequently increase the RankLoss.

In addition, we adopt the same Human isoform data from 569 RNA-seq runs of 298 samples from different tissues and conditions of Human ENCODE project, which was also used by Yu *et al.* (2020), to further comparatively study the performance of DMIL-IsoFun and other compared methods. More data information can refer the supplementary file of (Yu *et al.*, 2020). Table S3 reports the results evaluated at the gene-level. Note, alike the evaluation on Maize, the annotations of genes in the validation set are not used for training but used for validation only to avoid self-validation. DIFFUSE is trained with respect to each GO term and very time-consuming to complete on the Human dataset with more than 1000 GO terms. So its results on the Human dataset are not reported. Compared with the GO annotations and available RNA-seq datasets of Maize, the GO annotations of genes and RNA-seq datasets of human are relatively richer.

Table S1. The details of ten used B73 RNA-seq datasets.

Immature tassel	seedling shoot	seedling root	unpollinated ear tip	Root	SAM Apex
SRR445383	SRR445382	SRR445245	SRR445244	SRR395208	SRR424649
				SRR395194	SRR424650
				SRR395192	
				SRR395191	

Although the shallow methods can obtain a relatively better performance on Human than Maize, such as IsoFun and Disofun, DMIL-IsoFun still gets a better performance. The GO terms with larger sizes represent more heterogeneous functions and are therefore more difficult to distinguish among similar GO terms. It is difficult for DisoFun based on matrix factorization to learn discriminative features from the isoforms annotated with these GO terms. IsoFun based on label propagation on sparse isoform network to propagate GO terms, and it further considers both the positive and negative GO annotations of genes. As a result, it can obtain a better performance than DisoFun with respect to selected GO terms, each of which is annotated to at least 150 genes.

Table S5 reports the results of DMIL-IsoFun, DIFFUSE and DeepIsoFun on single-isoform genes in the CC branch, we can find that DMIL-IsoFun also obtains a better performance than other deep learning based methods on SIGs.

Table S5. Prediction results of deep learning based methods on SIGs of Maize in CC branch.

		AUC	AUPRC	$S_{min} \downarrow$	F_{max}
CC	DMIL-IsoFun	**0.704**	**0.698**	**0.727**	**0.564**
	DIFFUSE	0.504	0.564	0.737	0.507
	DeepIsoFun	0.589	0.556	0.789	0.378

3 Data sources analysis

From the results in the main text and those in the supplementary file, we can observe that DMIL-IsoFun can effectively fuse isoform sequence and expression data, and thus achieve a better performance than compared methods, which separately use or fuse RNA-seq datasets, gene-level interactions and sequence data. We further investigate the necessity and effectiveness of mining the composite isoform functional association network $\mathbf{A}$, which is composed with isoform sequences similarity network $\mathbf{W}^s$ and co-expression network $\mathbf{W}^c$. For this investigation, we randomly selected 500 isoforms, and then separately visualized the two networks and the composite one in Fig. S1. We can see that many isoforms co-expressed at the center of the co-expression network, along with many small islands far from the center, each of which is made of several isoforms with co-expression. The sequence similarity network also has dense connections in the center and many small islands made of several isoform. However, the global patterns between these two networks are still different. In the composite network, we can find that these islands are connected with center dense area of the composite network. In this way, the negative impact of noisy edges between isoforms can be reduced to some extent, and the further differentiation of functions of individual isoforms in the composite network can be more credibly made.

In practice, we find if DMIL-Isofun uses the isoform features extracted from MILCNN and the co-expression network derived from RNA-seq datasets to differentiate the functions of individual isoforms, the AUROC, AUPRC, F_{max} and S_{min} of DMIL-IsoFun reduces by 16.5%, 17.5% , 7.9% and 31.2%, respectively. In addition, we also try to construct a composite network by summing up $\mathbf{W}^s$ and $\mathbf{W}^c$, the AUROC, AUPRC, F_{max} and S_{min} of DMIL-IsoFun on this composite network drop by 2.3%, 9.2%, 3.9% and 7.0%, respectively.

4 Parameter sensitivity analysis

There are several key input parameters (k, γ and α) that may impact the performance DMIL-IsoFun. For γ and α, we use the optimal parameters recommended by Lin *et al.* (2020). In the main text, we adopt $k = 10$ to construct the isoform co-expression network and isoform sequence similarity network. Following the experimental setup in the main text, we study the performance variation of DMIL-IsoFun by varying k in $\{0, 5, 10, 15, 20\}$ and report the AUROC, AUPRC, F_{max} values of DMIL-IsoFun in BP subontology in Fig S2. The results in other subontology are similar and not reported. Here, $k = 0$ means that the isoform network only uses the sequence similarity network.

We observe that the performance of DMIL-IsoFun is improved by merging the isoform co-expression network, which can complement the tiny sequence difference between isoforms spliced from the same gene. However, when each isoform considers more than 10 neighbors in the co-expression network derived from Pearson correlation coefficients, the performance begins to decrease. A too sparse co-expression network (small k) can not provide sufficient functional associations among isoforms to enable GCN-based isoform function differentiation, even our DMIL-IsoFun considers S-order connections between isoforms. On the other hand, a too dense co-expression network (large k) results in some noisy/trivial associations between isoforms, and thus compromises the performance of attributed network embedding and the prediction of isoform functions. From these results, we can conclude that the fusion of isoform co-expression network and sequence similarity network indeed contributes to an improved performance for predicting the individual functions of isoforms. Based on these results, we adopt $k = 10$ for experiments.

5 Datasets statistics

We counted the number of isoforms spliced from each gene and reported the counts via histogram graph in Fig. S3 (for Maize dataset) and Fig. S4 (for Human dataset). The maximum number of isoforms spliced from the same genes of Maize is 16 and that of Human is 227. Since genes spliced into $\leq$ 20 isoforms accounts for 92.6% of all Human genes, we adopt $\tau = 20$ for both datasets.

We also studied the quantitative functionality difference of 7,587 MIGs of Maize and found 1,016 MIGs with spliced isoforms having different GO annotations. We firstly computed the functional annotations difference of any pairwise isoforms within these 1,016 MIGs, and then the average value of these differences within a gene, next presented the distribution of average values by boxplot in Fig. S5. The mean value across 1,016 MIGs is 5.18, as shown in Figure S5. These statistics confirm that isoforms spliced from the same gene indeed have different functions and the necessity to differentiate the individual functions of isoforms.

Table S2. Approximate gene-level evaluation results of isoform function prediction on Maize. The gene-level annotations are aggregated from isoforms spliced from respective genes.

	CC				MF				BP			
	AUROC	AUPRC	$S_{min}\downarrow$	F_{max}	AUROC	AUPRC	$S_{min}\downarrow$	F_{max}	AUROC	AUPRC	$S_{min}\downarrow$	F_{max}
miSVM	0.470	0.492	2.218	0.417	0.505	0.074	1.255	0.063	0.528	0.033	**1.006**	0.107
iMILP	0.628	0.494	2.265	0.357	0.530	0.119	3.445	0.089	0.578	0.044	4.621	0.106
IsoFun	0.557	0.467	2.030	0.442	0.561	0.149	3.351	0.250	0.529	0.099	4.361	0.347
Disofun	0.627	0.507	1.969	0.360	0.521	0.151	3.530	0.209	0.574	0.064	4.664	0.249
DeepIsoFun	0.604	0.564	0.875	0.397	0.579	**0.219**	2.380	0.308	0.552	**0.179**	3.265	0.422
DIFFUSE	0.518	0.565	0.804	0.517	0.502	0.197	2.407	0.299	0.496	0.178	3.315	0.399
DMIL-IsoFun	**0.754**	**0.647**	**0.612**	**0.654**	**0.775**	0.200	**1.780**	**0.532**	**0.747**	0.065	2.435	**0.561**

Table S3. Approximate gene-level evaluation results of isoform function prediction on Human. The gene-level predicted annotations are aggregated from isoforms of respective genes. We selected GO terms annotated to at least 150 genes, and obtained 148 MF terms, 175 CC terms and 855 BP terms with respect to 26,866 isoforms alternatively spliced from 12,371 Human genes for experiments.

	CC				MF				BP			
	AUROC	AUPRC	$S_{min}\downarrow$	F_{max}	AUROC	AUPRC	$S_{min}\downarrow$	F_{max}	AUROC	AUPRC	$S_{min}\downarrow$	F_{max}
miSVM	0.661	0.069	4.661	0.122	0.522	0.057	3.568	0.099	0.549	0.043	20.691	0.081
iMILP	0.566	0.117	5.673	0.548	0.534	0.068	3.732	0.519	0.540	0.109	21.969	0.278
IsoFun	0.632	0.322	4.757	0.607	0.528	0.144	3.829	0.471	**0.598**	0.163	19.969	0.339
Disofun	0.570	0.233	4.840	0.537	0.536	0.106	3.642	0.422	0.544	**0.187**	19.920	0.257
DeepIsoFun	0.509	0.244	4.751	0.606	0.506	0.070	3.662	0.570	0.511	0.077	21.346	0.329
DMIL-IsoFun	**0.649**	**0.339**	**4.452**	**0.634**	**0.550**	**0.225**	**3.482**	**0.578**	0.551	0.159	**19.875**	**0.380**

Table S4. The parameters of DMIL-IsoFun.

Name	convolution kernels	Instances pyramid pooling	Depth of GCN	Per-parameter adaptive learning rate	Learning rate	Activation function
Range	8, 16, 24, 32, 128	10	3	RMSprop, Adam	0.01	Relu, LeakyRelu

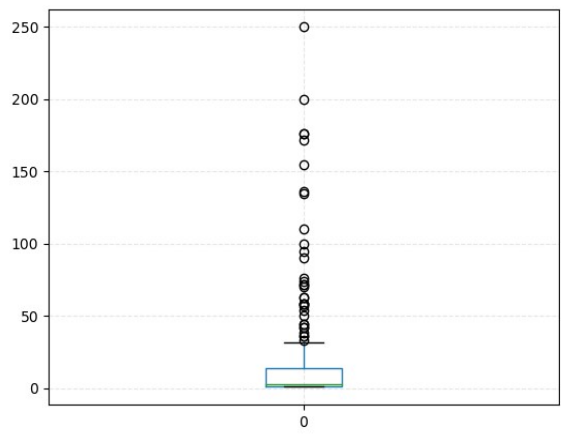


Fig. S5. Functionality differences of spliced isoforms within the same MIGs of Maize.

References

Ewels, P. *et al* (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics,* **32**(19), 3047–3048.

Kim, D. *et al* (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology,* **37**(8), 907–915.

Li, H. *et al* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics,* **25**(16), 2078–2079.

Liao Y. *et al* (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics,* **30**(7), 923–930.

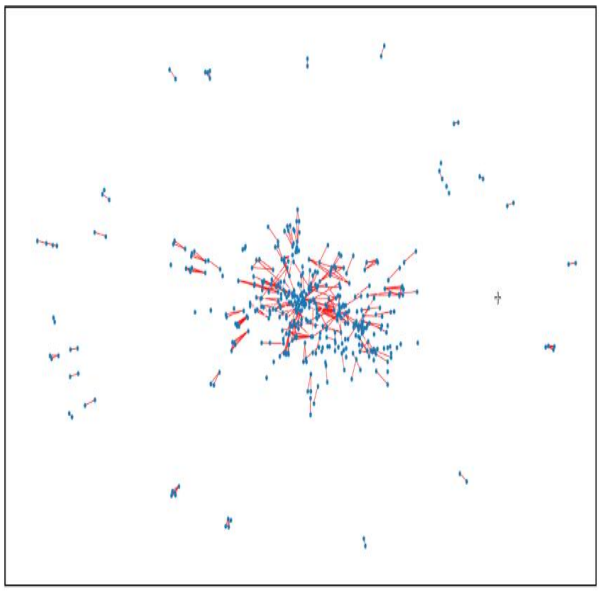
Lin, T.Y. *et al* (2020) Focal loss for dense object detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* **42** (2), 318–327.

Pertea, M. *et al* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology,* **33**(3), 290–295.

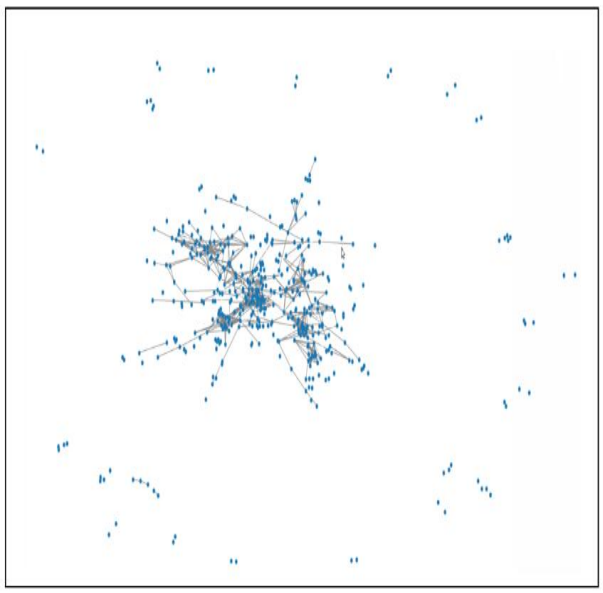
Patel, R.K. *et al* (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE,* **7**(2), e30619.

Suchan, T. *et al* (2016) Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS ONE,* **11**(3), e0151651.

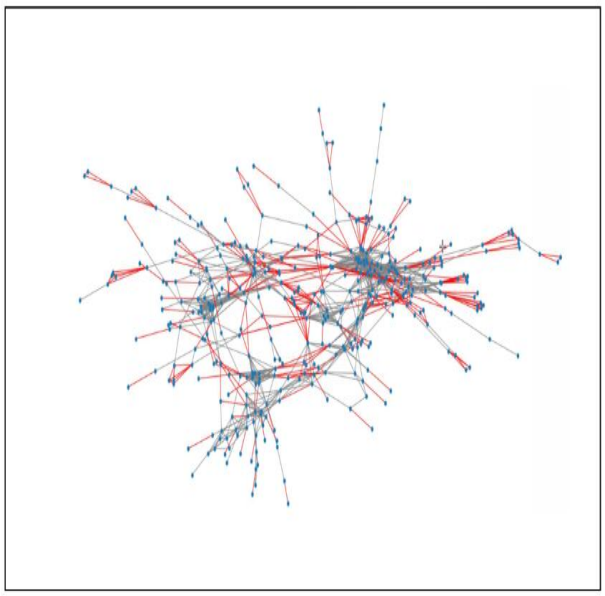
Yu, G. *et al* (2020) Isoform function prediction based on bi-random walks on a heterogeneous network. *Bioinformatics,* **36**(1), 303–310.



(a) Co-expression Network $\mathbf{W}^c$



(b) Sequence Similarity Network $\mathbf{W}^s$



(c) Composite Network $\mathbf{A}$

Fig. S1. Visualization of the isoform co-expression network derived from RNA-seq datasets (a), sequence similarity network induced by BLAST (b), and the composite network by fusing the above two networks (c).

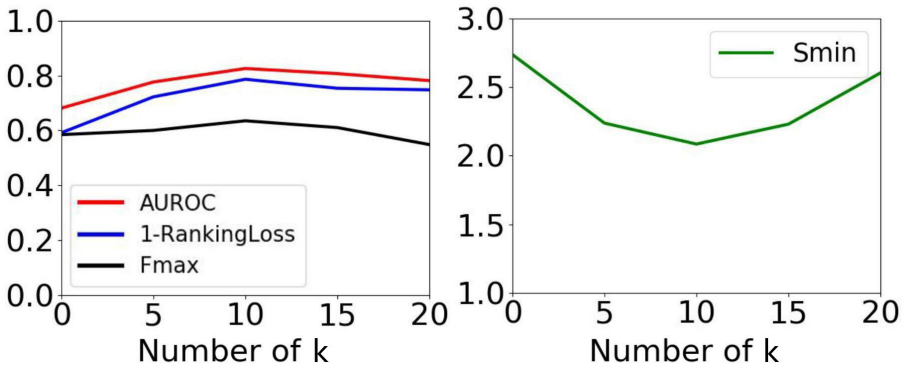


Fig. S2. Results of DMIL-IsoFun under different input values of k.

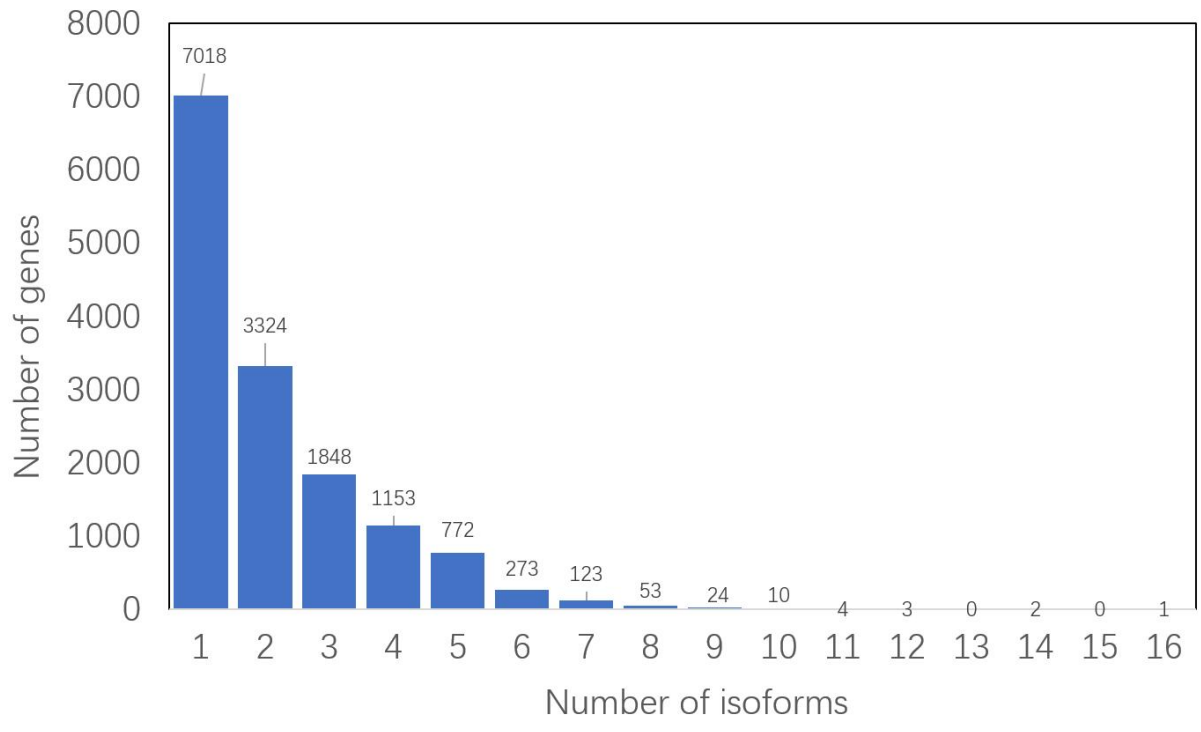


Fig. S3. Distribution of the number of isoforms spliced from the same genes of Maize.

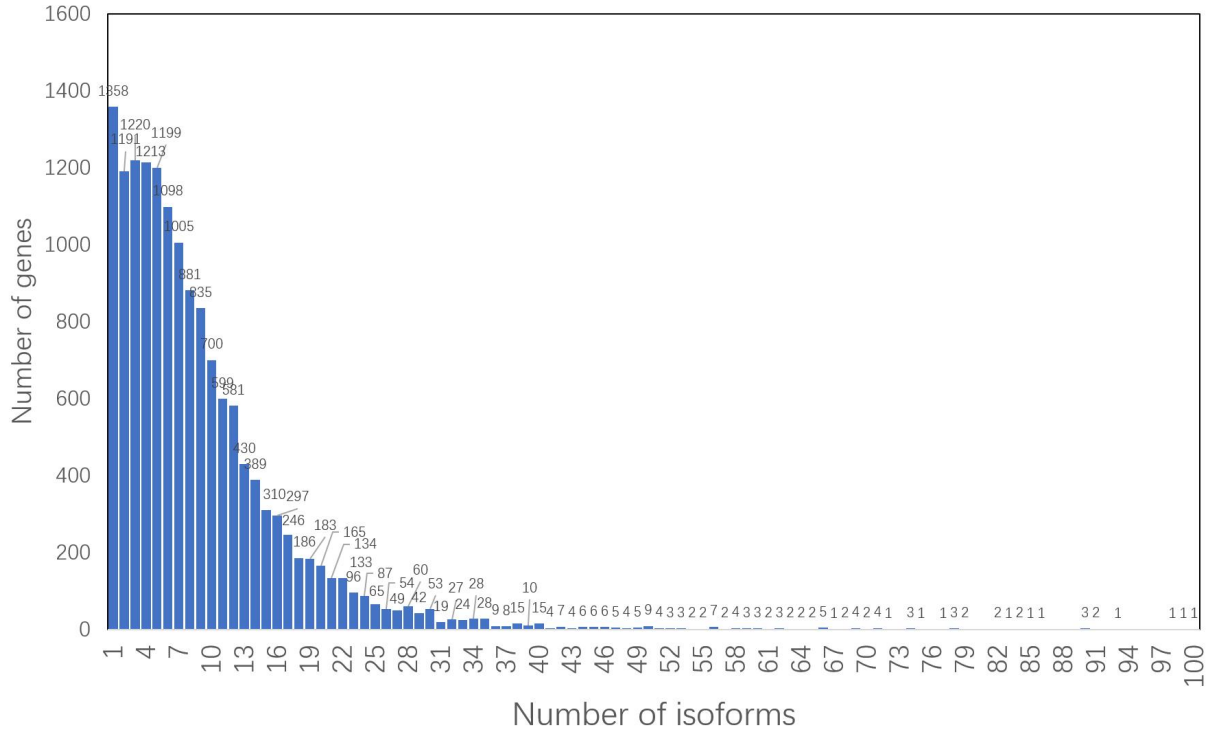


Fig. S4. Distribution of the number of spliced from the same genes of Human.