AIDE: annotation-assisted isoform discovery and abundance estimation from RNA-seq data

Wei Vivian Li[1,†], Shan Li[2,†], Xin Tong[3], Ling Deng[4], Hubin Shi[2,*], and Jingyi Jessica Li[1,5,*]

[1] *Department of Statistics, University of California, Los Angeles, CA, U.S.A. 90095-1554*
[2] *Laboratory of Tumor Targeted and Immune Therapy, Clinical Research Center for Breast, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University and Collaborative Innovation Center, China 610041*
[3] *Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA 90089-0809*
[4] *Laboratory of Molecular Diagnosis of Cancer, Clinical Research Center for Breast, West China Hospital, Sichuan University, China 610041*
[5] *Department of Human Genetics, University of California, Los Angeles, CA, U.S.A. 90095-7088*
[†] *These authors contributed equally to the work.*
[*] *To whom correspondence should be addressed.*

Abstract

Genome-wide accurate identification and quantification of full-length mRNA isoforms is crucial for investigating transcriptional and post-transcriptional regulatory mechanisms of biological phenomena. Despite continuing efforts in developing effective computational tools to identify or assemble full-length mRNA isoforms from next-generation RNA-seq data, it remains a challenge to accurately identify mRNA isoforms from short sequence reads due to the substantial information loss in RNA-seq experiments. Here we introduce a novel statistical method, AIDE (Annotation-assisted Isoform Discovery and abundance Estimation), the first approach that directly controls false isoform discoveries by implementing the statistical model selection principle. Solving the isoform discovery problem in a stepwise and conservative manner, AIDE prioritizes the annotated isoforms and precisely identifies novel isoforms whose addition significantly improves the explanation of observed RNA-seq reads. We evaluate the performance of AIDE based on multiple simulated and real RNA-seq datasets followed by a PCR-Sanger sequencing validation. Our results show that AIDE effectively leverages the annotation information to compensate the information loss due to short read lengths. AIDE achieves the highest precision in isoform discovery and the lowest error rates in isoform abundance estimation, compared with three state-of-the-art methods Cufflinks, SLIDE, and StringTie. As a robust bioinformatics tool for transcriptome analysis, AIDE will enable researchers to discover novel transcripts with high confidence.

Introduction

A transcriptome refers to the entire set of RNA molecules in a biological sample (e.g., a cell or a tissue sample). Alternative splicing, a post-transcriptional process during which particular exons

of a gene may be included into or excluded from a mature messenger RNA (mRNA) isoform (i.e., transcript) transcribed from that gene, is a key contributor to the diversity of eukaryotic transcriptomes [1]. Alternative splicing is a prevalent phenomenon in multicellular organisms, and it affects approximately 90%-95% of genes in mammals [2]. Understanding the diversity of eukaryotic transcriptomes is essential to interpreting gene functions and activities under different biological conditions [3]. In transcriptome analysis, a key task is to accurately identify the set of existing isoforms and estimate their abundance levels under a specific biological condition, because the information on isoform composition is critical to understanding the isoform-level dynamics of RNA contents in different cells, tissues, and developmental stages. Abnormal splicing events have been known to cause many genetic disorders [4], such as retinitis pigmentosa [5] and spinal muscular atrophy [6]. Accurate isoform identification and quantification will shed light on gene regulatory mechanisms of genetic diseases, thus assisting biomedical researchers in designing targeted therapies for diseases.

The identification of truly expressed isoforms is an indispensable step preceding accurate isoform quantification. However, compared with the quantification task, isoform discovery is an inherently more challenging problem both theoretically and computationally. The reasons behind this challenge are threefold. First, next-generation RNA-seq reads are too short compared with full-length mRNA isoforms. RNA-seq reads are typically no longer than 300 base pairs (bp) in Illumina sequencing [7], while more than 95% of human isoforms are longer than 300 bp, with a mean length of $1,712$ bp (the GENCODE annotation, Release 24) [8]. Hence, RNA-seq reads are short fragments of full-length mRNA isoforms. Due to the fact that most isoforms of the same gene share some overlapping regions, many RNA-seq reads do not unequivocally map to a unique isoform. As a result, isoform origins of those reads are ambiguous and need to be inferred from a huge pool of candidate isoforms. Another consequence of short reads is that "junction reads" spanning more than one exon-exon junctions are underrepresented in next-generation RNA-seq data, due to the difficulty of mapping junction reads (every read needs to be split into at least two segments and has all the segments mapped to different exons in the reference genome). The underrepresentation of those junction reads further increases the difficulty of discovering full-length RNA isoforms accurately. Second, the number of candidate isoforms increases exponentially with the number of exons. Hence, computational efficiency becomes an inevitable factor that every computational method must account for, and an effective isoform screening step is often needed to achieve accurate isoform discovery [9]. Third, it is a known biological phenomenon that often only a small number of isoforms are truly expressed under one biological condition. Given the huge number of candidate isoforms, how isoform discovery methods balance the parsimony and the accuracy of their discovered isoforms becomes a critical and meanwhile difficult issue [10, 11].

Over the past eight years, computational researchers have developed multiple state-of-the-art isoform discovery methods to tackle one or more of the challenges mentioned above. The two earliest annotation-free methods are Cufflinks [12] and Scripture [13], which can assemble mRNA

isoforms solely from RNA-seq data without using annotations of known isoforms. Both methods use graph-based approaches, but they differ in how they construct graphs and then parse a graph into isoforms. Scripture first constructs a connectivity graph with nodes as genomic positions and edges determined by junction reads. It then scans the graph with fixed-sized windows, scores each path for significance, connects the significant paths into candidate isoforms, and finally refines the isoforms using paired-end reads. Cufflinks constructs an overlap graph of mapped reads, and it puts a directed edge based on the genome orientation between two compatible reads that could arise from the same isoform. It then finds a minimal set of paths that cover all the fragments in the overlap graph. A more recent method StringTie also uses the graph idea. It first creates a splice graph with read clusters as nodes to identify isoforms, and it then constructs a flow network to estimate the expression levels of isoforms using a maximum flow algorithm [14].

Another suite of methods utilize different statistical and computational tools and regularization methods to tackle the problems of isoform discovery and abundance estimation. IsoLasso [15], SLIDE [16], and CIDANE [10] all build linear models, where read counts are summarized as the response variable, and isoform abundances are treated as parameters to be estimated. The candidate isoforms with non-zero estimated abundance are then considered as discovered. IsoLasso starts from enumerating valid isoforms and then uses the Lasso algorithm [17] to achieve parsimony in its discovered isoforms. SLIDE incorporates the information of gene and exon boundaries in annotations to enumerate candidate isoforms, and it uses a modified Lasso procedure to select isoforms. CIDANE also uses the Lasso regression in its first phase, and then it employs a delayed column generation technique in its second phase to check if new isoforms should be added to improve the solution. Another method iReckon takes a different approach and tackles the isoform discovery problem via maximum likelihood estimation [11]. It first constructs all the candidate isoforms supported by RNA-seq data, and then it utilizes a regularized expectation-maximization (EM) algorithm [18] to reduce the number of expressed isoforms and avoid over-fitting.

Aside from the intrinsic difficulty of isoform identification due to the short read lengths and the huge number of candidate isoforms, the excess biases in RNA-seq experiments further afflict the isoform discovery problem. Ideally, RNA-seq reads are expected to be uniformly distributed within each isoform. However, the observed distribution of RNA-seq reads significantly violates the uniformity assumption due to multiple sources of biases. The most commonly acknowledged bias source is the different levels of guanine-cytosine contents (GC contents) in different regions of an isoform. The GC content bias was first investigated by Dohm et al. [19], and a significantly positive correlation was observed between the read coverage and the GC contents. Another work later showed that the effects of GC contents tend to be sample-specific [20]. Another major bias source is the positional bias, which causes the uneven read coverage at different relative positions within an isoform. As a result of the positional bias, reads are more likely to be generated from certain regions of an isoform, depending on experimental protocols, e.g., whether cDNA fragmentation or RNA fragmentation is used [21, 22]. Failing to correct these biases will likely lead to high false discovery rates in isoform discovery and unreliable statistical results in downstream analyses e.g.,

differential isoform expression analysis [23].

Current computational methods account for the non-uniformity of reads using three main approaches: to adjust read counts summarized in defined genomic regions to offset the non-uniformity bias [21, 24], to assign a weight to each single read to adjust for bias [25], and to incorporate the bias as a model parameter in likelihood-based methods [26, 27, 28, 29] for isoform discovery or abundance estimation.

Despite continuous efforts the bioinformatics community has spent on developing effective computational methods to identify full-length isoforms from next-generation RNA-seq data, the existing methods still suffer from low accuracy for genes with complex splicing structures [30, 31]. A comprehensive assessment has shown that methods achieving good accuracy in identifying isoforms in *D. melanogaster* ($34,776$ annotated isoforms) and *C. elegans* ($61,109$ annotated isoforms) fail to maintain good performance in *H. sapiens* ($200,310$ annotated isoforms) [30, 32]. Although it is generally believed that deeper sequencing will lead to better isoform discovery results, the improvement is not significant in *H. sapiens*, compared with *D. melanogaster* and *C. elegans*, due to the complex splicing structures of human genes [33, 30]. Moreover, despite increasing precision and recall rates of identified isoforms evaluated at the nucleotide level and the exon level, it remains challenging to improve the isoform-level performance. In other words, even when all sub-isoform elements (i.e., short components of transcribed regions such as exons) are correctly identified, accurate assembly of these elements into full length isoforms remains a big challenge.

Motivated by the observed low accuracy in identifying full-length isoforms solely from RNA-seq data, researchers have considered leveraging information from reference annotations (e.g., ENSEMBL [32], GENCODE [8], and UCSC Genome Browser [34]) to aid isoform discovery. Existing efforts include two approaches. In the first approach, methods extract the coordinates of gene and exon boundaries, i.e., known splicing sites, from annotations and then assemble novel isoforms based on the exons or subexons (the regions between adjacent known splicing sites) of every gene [10, 14, 16]. In the second approach, methods directly incorporate all the isoforms in annotations by simulating faux-reads from the annotated isoforms [35]. However, the above two approaches have strong limitations in their use of annotations. The first approach does not fully use annotation information because it neglects the splicing patterns of annotated isoforms, and these patterns could assist learning the relationship between short reads and full-length isoforms. The second approach is unable to filter out non-expressed annotated isoforms because researchers lack prior knowledge on which annotated isoforms are expressed in an RNA-seq sample; hence, its addition of unnecessary faux-reads will bias the isoform discovery results and lose control of the false discovery rate.

Here we propose a more statistically principled approach, AIDE (**A**nnotation-assisted **I**soform **D**iscovery and abundance **E**stimation), to leverage annotation information in a more advanced manner to increase the precision and robustness of isoform discovery. Our approach is rooted in statistical model selection, which takes a conservative perspective to search for the smallest model that fits the data well, after adjusting for the model complexity. In our context, a model corresponds

to a set of candidate isoforms (including annotated ones and novel ones), and a more complex model contains more isoforms. Our rationale is that a robust and conservative computational method should only consider novel isoforms as credible if adding them would significantly better explain the observed RNA-seq reads than using only the annotated isoforms. AIDE differs from many existing approaches in that it does not aim to find all seemingly novel isoforms. It enables controlling false discoveries in isoform identification by employing a statistical testing procedure, which ensures that the discovered isoforms make statistically significant contributions to explaining the observed RNA-seq reads. Specifically, AIDE learns gene and exon boundaries from annotations and also selectively borrows information from the annotated isoform structures using a stepwise likelihood-based selection approach. Instead of fully relying on the annotation, AIDE is capable of identifying non-expressed annotated isoforms and remove them from the identified isoform set. Moreover, AIDE simultaneously estimates the abundance of the identified isoforms in the process of isoform reconstruction.

Results

The AIDE method utilizes the likelihood ratio test to identify isoforms via a stepwise selection procedure, which gives priority to the annotated isoforms and selectively borrows information from their structures. The stepwise selection in AIDE consists of both forward and backward steps. Each forward step finds an isoform whose addition contributes the most to the explanation of RNA-seq reads given the currently identified isoforms. Each backward step rectifies the identified isoform set by removing the isoform with the most trivial contribution given other identified isoforms. AIDE achieves simultaenous isoform discovery and abundance estimation based on a carefully constructed probabilistic model of RNA-seq read generation, and the stepwise selection process consists of model parameter estimation and likelihood ratio tests (see Methods; Figure 1). We first use a comprehensive transcriptome-wide study to evaluate and compare the performance of AIDE and three other widely used methods (Cufflinks, SLIDE, and StringTie) provided with various read coverages and annotations of different quality. Second, we validate the isoform discovery results of these four methods using a benchmark Nanostring dataset. Third, we further assess the precision and recall rates of these four methods on human and mouse RNA-seq datasets. We finally use a simulation study to demonstrate the necessity and superiority of stepwise selection in AIDE. In all the above four studies, AIDE demonstrates its advantages in achieving the highest precision in isoform discovery and the best accuracy in isoform quantification among the four methods.

AIDE outperforms state-of-the-art methods on simulated data

We compare AIDE with three other state-of-the-art isoform discovery methods, Cufflinks [12], StringTie [14], and SLIDE [16], in a simulation setting that well mimics real RNA-seq data analysis. The four methods tackle the isoform discovery task from different perspectives. Cufflinks assem-

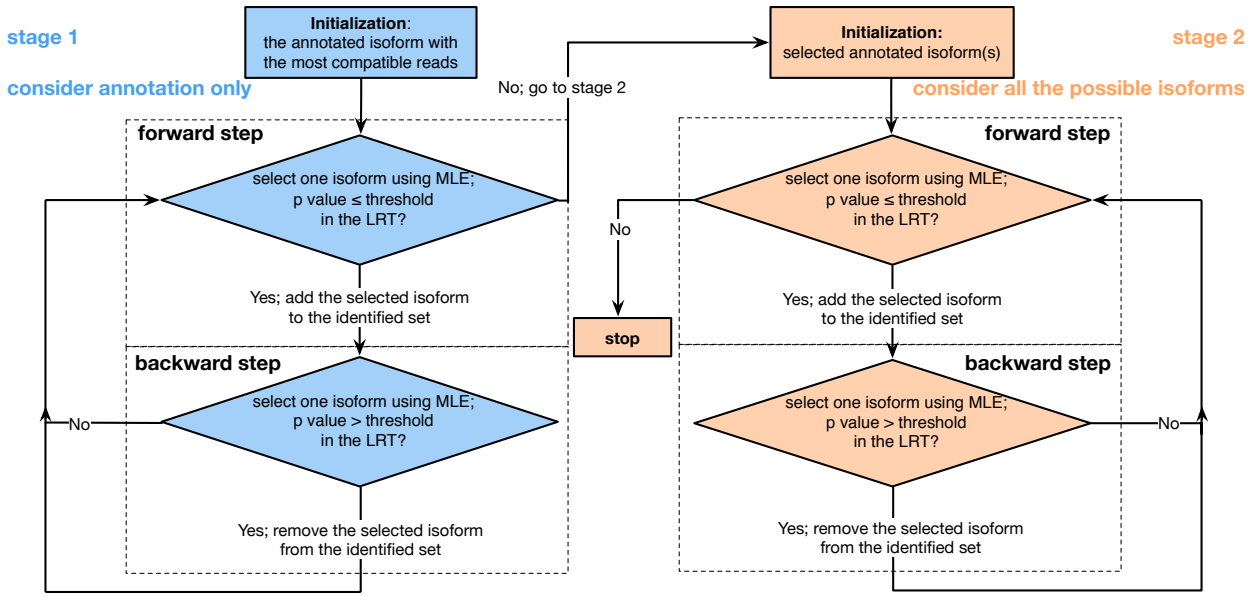

Figure 1: Workflow of the stepwise selection in the AIDE method. Stage 1 starts with a single annotated isoform compatible with the most reads, and all the other annotated isoforms are considered as candidate isoforms. Stage 2 starts with the annotated isoforms selected in stage 1, and all the possible isoforms, including the unselected annotated isoforms, are considered as candidate isoforms. In the forward step in both stages, AIDE identifies the isoform that mostly increases the likelihood, and it uses the likelihood ratio test (LRT) to decide whether this increase is statistically significant. If significant, AIDE adds this isoform to its identified isoform set; otherwise, AIDE keeps its identified set and terminates the current stage. In the backward step in both stages, AIDE finds the isoform in its identified set such that the removal of this isoform decreases the likelihood the least, and it uses the likelihood ratio test to decide whether this decrease is statistically significant. If significant, AIDE removes this isoform from its identified set; otherwise, AIDE keeps the identified set. After the backward step, AIDE returns to the forward step. AIDE stops when the forward step in Stage 2 no longer adds a candidate isoform to the identified set.

bles isoforms by constructing an overlap graph and searching for isoforms as sparse paths in the graph, SLIDE utilizes a regularized linear model, and StringTie uses a network-based algorithm. Unlike all these three methods, our proposed method AIDE, built upon likelihood ratio tests and stepwise selection, converts the isoform discovery problem into a statistical variable selection problem.

To conduct a fair assessment of the four methods, we simulate RNA-seq datasets using the R package `polyester` [22], which uses both built-in models and real RNA-seq datasets to generate synthetic RNA-seq data that exhibit similar properties to those of real data. We simulate eight human RNA-seq datasets with eight different read coverages (10x, 20x, $\ldots$, 80x) and predetermined isoform fractions (see Methods). We compare the accuracy of the four methods supplied with annotations of different quality. In real scenarios, annotations contain both expressed (true) and non-expressed (false) isoforms in a specific RNA-seq sample. Annotations might miss some expressed isoforms in that RNA-seq sample because alternative splicing is known to be condition-specific and widely diverse across different conditions. Therefore, it is critical to evaluate the extent to which different methods rely on the accuracy of annotations, specifically the annotation purity and the annotation completeness, which we define as the proportion of

expressed isoforms among the annotated ones and the proportion of annotated isoforms among the expressed ones, respectively.

We construct nine sets of synthetic annotations (as opposed to the real annotation from GENCODE [8] or Ensembl [36]), with varying purity and completeness (Table 1). For example, annotation 4 has a 60% purity and a 40% completeness, meaning that 60% of the annotated isoforms are truly expressed and the annotated isoforms constitute 40% of the truly expressed isoforms.

Table 1: Synthetic annotations. Purity and completeness of the nine sets of synthetic annotations are calculated based on the truly expressed isoforms in the simulated data.

synthetic annotation set	1	2	3	4	5	6	7	8	9
purity	40%	40%	40%	60%	60%	60%	80%	80%	80%
completeness	40%	60%	80%	40%	60%	80%	40%	60%	80%

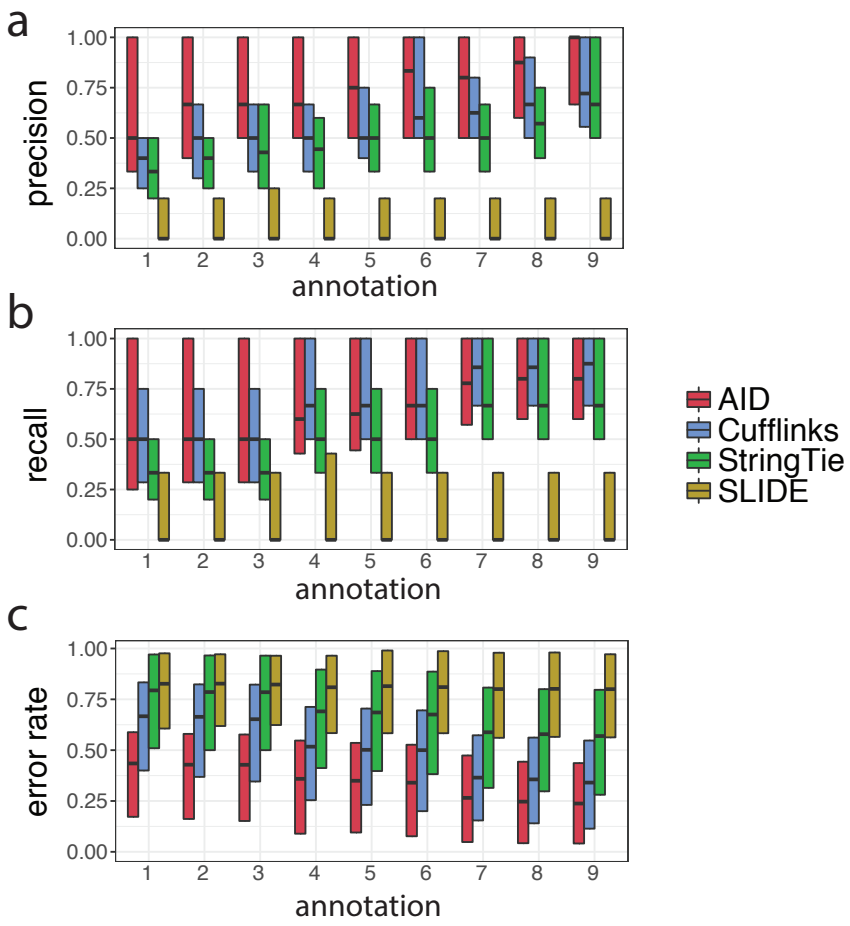

Figure 2: Distributions of per-gene isoform discovery and abundance estimation results of AIDE, Cufflinks, StringTie, and SLIDE on simulated RNA-seq data with 10x coverage. Each box gives the 1st quantile, median, and 3rd quantile of the per-gene results given each of the nine sets of synthetic annotations. **a:** precision rates in isoform discovery; **b:** recall rates in isoform discovery; **c:** error rates (defined as the total variation) in abundance estimation.

We first compare AIDE and the other three methods Cufflinks, SLIDE, and StringTie in terms of their isoform discovery accuracy per gene. Regarding both the precision rates (i.e., the proportions of expressed isoforms in the discovered isoforms) and the recall rates (i.e., the proportions of discovered isoforms in the expressed isoforms), AIDE outperforms the other three methods with all the nine sets of synthetic annotations (Figures 2, S2, and S3). Especially with the less accurate synthetic annotation sets 1-4, AIDE demonstrates its clear advantages in precision and recall rates thanks to its stepwise selection strategy, which prevents AIDE from being misled by the wrongly annotated isoforms. When the read coverage is 10x and the annotations have 40% purity (sets 1-3), the median precision rates of AIDE are as high as the 3rd-quantile precision rates of Cufflinks. In addition, AIDE achieves high precision rates ($> 75\%$) much more frequently than the other three methods (Figure 2a). In terms of the recall rates, AIDE and Cufflinks exhibit better capability in correctly identifying truly expressed isoforms than StringTie and SLIDE. AIDE achieves high recall rates ($> 75\%$) in more genes with the annotation sets 1-3 (purity $= 40\%$), and its recall rates are close to those of Cufflinks when the annotation purity is increased to 60% and 80% (annotation sets 4-9) (Figure 2b). We also find that AIDE is robust to sequencing depths. On the contrary, the recall rates of StringTie and SLIDE deteriote quickly as sequencing depths decrease (Figures S2 and S3). Regarding to what extent does the accuracy of annotations affect isoform discovery, we find that the annotation purity is more important than the annotation completeness for isoform discovery (Figures 2, S2, and S3). This observation suggests that if practitioners have to choose between two annotation sets, one with high purity but low completeness and the other with low purity but high completeness, they should use the former annotation set as input into AIDE.

As a concrete example, we consider the human gene *DPM1* and its annotated isoforms in synthetic annotation sets 1 and 9 (Table 1). For *DPM1*, the annotation set 1 has a 67% purity and a 67% completeness, and the annotation set 9 has a 60% purity and a 100% completeness. In Figures S4 and S5, we plot the distribution of RNA-seq reads in the reference genome, along with the truly expressed isoforms, the annotated isoforms, and the isoforms identified by AIDE, Cufflinks, and StringTie, respectively. Thanks to its capacity to selectively incorporate information from the annotated isoforms, AID successfully identifies the shortest truly expressed isoform, which is missing in the annotation set 1 (Figure S4). With the annotation set 9 that contains two non-expressed isoforms, AID correctly identifies the three truly expressed isoforms (Figure S5). In contrast, Cufflinks, StringTie, and SLIDE miss some of the truly expressed isoforms with both annotation sets. Specifically, they miss the shortest expressed isoform not in the annotation set 1, and they identify too many non-expressed isoforms with the less pure annotation set 9.

We also summarize the genome-wide average precision rates, recall rates, and F scores of AIDE and the other three methods at three different levels: base, exon and isoform levels, with each of the nine synthetic annotation sets (Figure 3). Please see Supplementary Materials for the definitions of the precision rate, the recall rate, and the F score at the base, exon, and isoform levels. All the four methods have high accuracy at the base and exon levels regardless of the accuracy of the synthetic annotations. However, even when exons are correctly identified, it

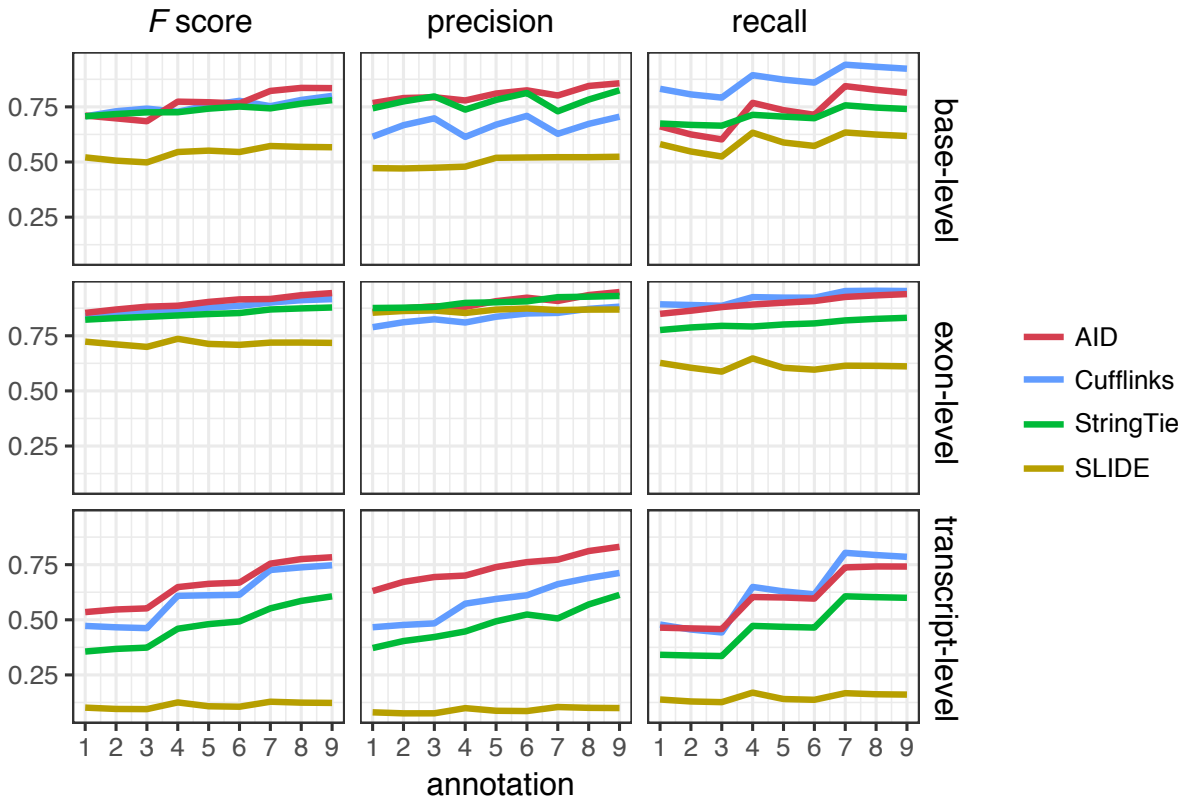

Figure 3: Comparison between AIDE and the other three isoform discovery methods in simulation. What is displayed is the genome-wide average performance of AIDE, Cufflinks, StringTie, and SLIDE given each of the nine synthetic annotation sets. The base-level, exon-level, and isoform-level precision rates, recall rates, and F scores averaged across all the $2,262$ genes are calculated based on RNA-seq data with a 10x coverage.

remains challenging to accurately assemble exons into full-length isoforms. At the isoform level, AIDE achieves the best precision rates, recall rates comparable to Cufflinks, and the best F scores with all the synthetic annotation sets. This result demonstrates the advantage of AIDE in achieving high precision and low false discovery rates in isoform discovery.

As for every gene, the proportions of the expressed isoforms are specified in this simulation study, we also compare AIDE with the other three methods in terms of their accuracy in isoform abundance estimation. We use $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)'$ to denote the proportions of J possible isoforms enumerated from a given gene's known exons, and we use $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_J)'$ to denote the estimated proportions by a method. We note that $\sum_{j=1}^{J} \alpha_j = \sum_{j=1}^{J} \hat{\alpha}_j = 1$. We define the estimation error rate as the total variation: $e(\hat{\boldsymbol{\alpha}}) = \frac{1}{2} \sum_{j=1}^{J} |\alpha_j - \hat{\alpha}_j|$. The error rate is a real value in $[0, 1]$, with a value 0 representing a 100% accuracy. With all the nine synthetic annotation sets, AIDE achieves the overall smallest error rates (Figure 2c). The advantages of AIDE over the other three methods are especially obvious with the first three annotations, whose purity is only 40%.

AIDE improves isoform abundance estimation on real data

It would be ideal if we could evaluate the discovered isoforms and the estimated isoform abundance of different methods on real data. However, the structures and abundance of expressed isoforms are unobservable in real RNA-seq data. Therefore, we seek to evaluate the performance of different methods by comparing their estimated isoform expression levels in the FPKM (Fragments Per Kilobase Million) unit with the NanoString counts, which often serve as benchmark data for isoform abundance [37, 38], because the NanoString nCounter technology is considered as one of the most reproducible and robust assays for quantifying gene and isoform expression levels [39]. We expect an accurate isoform discovery method to discover a set of isoforms close to the expressed isoforms in an RNA-seq sample. If the identified isoforms are accurate, the subsequently estimated isoform abundance is more likely to be accurate and agree better with the NanoString counts.

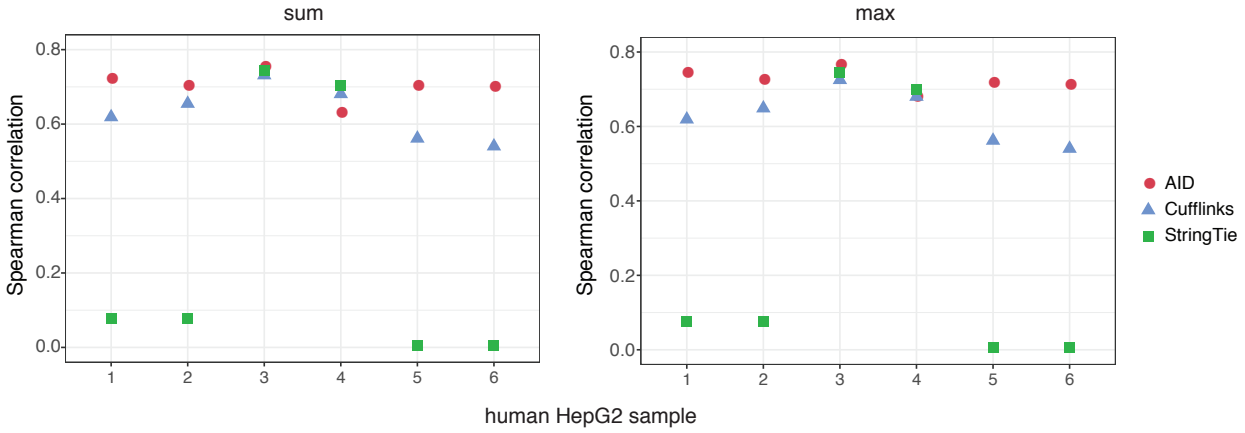

Figure 4: Spearman correlation coefficients between the estimated isoform expression and the benchmark NanoString counts. **a**: For every probe, the sum of the expression levels of its corresponding isoforms is used in the calculation. **b**: For every probe, the maximum of the expression levels of its corresponding isoforms is used in the calculation.

We therefore apply AIDE, Cufflinks, and StringTie to six samples of the human HepG2 (liver hepatocellular carcinoma) immortalized cell line with both RNA-seq and NanoString data [30] (Supplementary Table S1). We supply all the three methods with the GENCODE annotation (version 24) [8] for their isoform discovery. SLIDE requires its input RNA-seq reads to have a unique read length after mapping and is thus not applicable to the six HepG2 mapped RNA-seq datasets. We note that the NanoString nCounter technology is not designed for genome-wide quantification of RNA molecules, and the HepG2 NanoString datasets have measurements of 140 probes corresponding to 470 isoforms of 107 genes. Since one probe may correspond to multiple isoforms, we first find the isoforms compatible with every probe, and we then compare the sum or the maximum of the estimated abundance of these isoforms with the count of that probe. For each HepG2 sample, we calculate the Spearman correlation coefficient between the estimated isoform abundance ("sum" or "max") and the NanoString probe counts to evaluate the accuracy of each

method (Figure 4). AIDE has the highest correlations in five out of the six samples, suggesting that AIDE achieves more accurate isoform discovery as well as better isoform abundance estimation in this application.

AIDE improves isoform discovery on real data

We have demonstrated that AIDE achieves more precise isoform discovery and more accurate isoform abundance estimation than Cufflinks, StringTie, and SLIDE in simulations and on real data with Nanostring counts as a benchmark. In addition to these results, we perform another transcriptome-wide comparison of AIDE and the other three methods to further validate the robustness and reproducibility of AIDE in isoform discovery. Since transcriptome-wide benchmark data are unavailable for real RNA-seq experiments, we use the isoforms in the GENCODE annotation as a surrogate basis for evaluation [8]. For every gene, we randomly select half of the annotated isoforms and input them as partial annotations into every isoform discovery method. For RNA-seq data, we collect from public data repository three human embryonic stem cell (ESC) datasets and three mouse bone marrow-derived macrophage datasets (Supplementary Table S1). For the human ESC data we use the GENCODE annotation version 24; for the mouse macrophage data we use the GENCODE annotation version 10. For each gene, we apply AIDE, Cufflinks, StringTie, and SLIDE to these six datasets for isoform discovery with partial annotations, and we evaluate the precision and recall rates of each method by comparing their identified isoforms with the complete set of annotated isoforms. Although the annotated isoforms are not equivalent to the truly expressed isoforms in the six samples from which RNA-seq data were generated, the identified isoforms, if accurate, are supposed to largely overlap with the annotated isoforms given the quality of human and mouse annotations. Especially if we assume the human and mouse annotations are unions of known isoforms expressed under various well-studied biological conditions including human ESCs and mouse bone marrow-derived macrophages, it is reasonable to use those annotations to estimate the precision rates of the discovered isoforms, i.e., what proportions of the discovered isoforms are expressed. Estimation of the recall rates is more difficult, because the annotations are likely to include some isoforms that are non-expressed in human ESCs or mouse bone marrow-derived macrophages.

We summarize the isoform discovery accuracy of the four methods on each dataset at both the exon level and the isoform level in Figure 5. At the exon level, AIDE has the highest precision and recall rates on all the six datasets, achieving F scores greater than 90% (Figure 5a-b). The second best method at the exon level is StringTie. Since connecting exons into the full-length isoforms is much more challenging than simply identifying the individual exons, all the methods have lower accuracy at the isoform level than at the exon level. While having slightly lower but similar recall rates than Cufflinks, AIDE achieves the highest precision rates ($\sim 80\%$ on human datasets and $\sim 60\%$ on mouse datasets) at the isoform level (Figure 5c-d). Although precision and recall rates are both important measures for isoform discovery, high precision results (equivalently, low false discovery results) of computational methods are often preferable for experimental validation. In

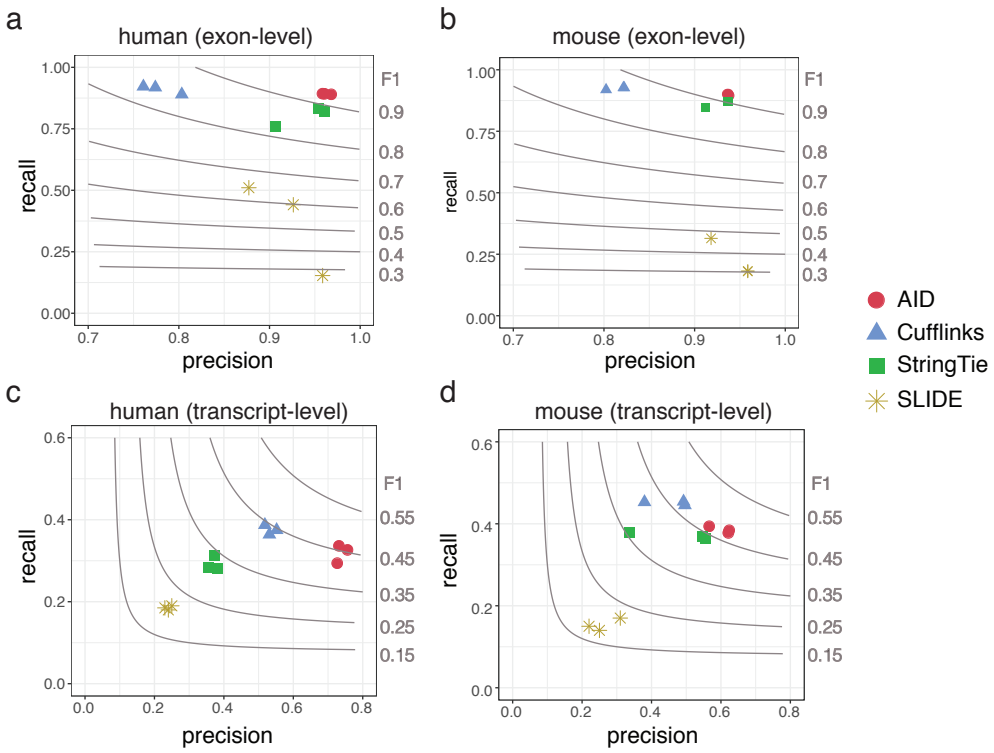

Figure 5: Comparison of AIDE and the other three methods in real data. **a**: exon-level accuracy in the human ESC samples; **b**: exon-level accuracy in the mouse macrophage samples; **c**: transcript-level accuracy in the human ESC samples; **d**: transcript-level accuracy in the mouse macrophage samples. The gray contours denote the F scores, as marked on the right of each panel.

the three mouse macrophage samples, AIDE identifies novel isoforms for genes *MAPKAPK2*, *CXCL16*, and *HIVEP1*, which are known to play important roles in macrophage activation [40, 41, 42]. Since AIDE has higher percision rates than the other three methods, these novel isoforms found by AIDE are worth investigating in macrophage studies. In our previous simulation results, we have shown that the accuracy of annotations is a critical factor determining the performance of AIDE. Even though the partial annotations used in this study only have a 50% completeness, AIDE achieves the best precision rates among the four methods, and we expect AIDE to achieve high accuracy when supplied with annotations of better quality in real applications.

AIDE is able to achieve more precise isoform discovery than existing methods because it utilizes a statistical model selection principle to determine whether to identify a candidate isoform as expressed. Therefore, only those isoforms that are statistically supported by the evidence (i.e., the observed reads) are retained. We use four example genes *ZBTB11, TOR1A, MALSU1,* and *SRF6* to illustrate the superiority of AIDE over the other three methods (Supplementary Figures S6 and S7). The genome browser plots clearly show that AIDE identifies the annotated isoforms with the best precision, while the other methods either miss some annotated isoforms or lead to too many false discoveries. We also use the same genes *ZBTB11, TOR1A, MALSU1,* and *SRF6* to show that AIDE is robust to the choice of the p-value threshold used in the likelihood ratio tests (see

Methods). The default choice of the threshold is $\frac{0.01}{\text{total number of genes}}$, which is 4.93×10^{-7} for human samples and 4.54×10^{-7} for mouse samples. We use AIDE to identify isoforms with different thresholds chosen from $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}, 10^{-10}\}$ and track how the results change while the threshold decreases from 10^{-2} to 10^{-10} (Supplementary Figures S8 and S9). As expected, AIDE tends to discover slightly more isoforms when the threshold is larger, and it becomes more conservative with a smaller threshold. However, the default threshold leads to accurate results for those four genes, and the discovered isoform set remains stable around the default threshold.

PCR-Sanger sequencing validates the effective reduction of false discoveries by AIDE

Since AIDE and Cufflinks have demonstrated higher accuracy than other methods in the above assessment of genome-wide isoform discovery from synthetic and real RNA-seq datasets, we further evaluated the performance of these two methods on a small cohort RNA-seq datasets using polymerase chain reaction (PCR) followed by Sanger sequencing. We applied both AIDE and Cufflinks to five breast cancer RNA-seq datasets for isoform identification with GENCODE annotation version 24. After summarizing and comparing the genome-wide isoform discovery results, we randomly selected ten genes that have annotated transcripts uniquely predicted by only AIDE or Cufflinks with FPKM > 2 for experimental validation (Table S2). We summarized the genes into two categories: six genes with annotated isoforms identified only by Cufflinks but not by AIDE (category 1), and four genes with annotated isoforms identified only by AIDE but not by Cufflinks (category 2). For four out of the six genes in category 1, *MTHFD2*, *NPC2*, *RBM7*, and *CD164*, our experimental validation found that the isoforms uniquely predicted by Cufflinks were false positives (Figure 6a-d). Specifically, both AIDE and Cufflinks correctly identified the full-length isoforms *MTHFD2-201*, *NPC2-207*, *RBM7-203*, *CD164-003* for the four genes, respectively. However, the isoforms *MTHFD2-203*, *NPC2-205*, *RBM7-208*, *CD164-210* predicted only by Cufflinks were all false discoveries. The validation results of category 1 indicates that AIDE effectively reduces false positive prediction of full-length mRNA isoforms compared with Cufflinks. For two out of the four genes in category 2, we validated the isoforms uniquely predicted by AIDE as true positives (Figure 6e-f). In detail, AIDE correctly identified isoforms *FGFR1-238* and *FGFR1-201* for gene *FGFR1*, as well as isoform *ZFAND5-208* for gene *ZFAND5*. On the other hand, Cufflinks only identified *FGFR1-238* and missed the other two isoforms. The validation results of category 2 suggests that AIDE also has good recall performance for isoform discovery in this case study.

AIDE improves isoform discovery accuracy via stepwise selection

We also conduct a proof-of-concept simulation study to verify the efficiency and accuracy of our proposed AIDE method. We also use this study to show why simply performing forward selection is insufficient and how stepwise selection, which combines both forward and backward

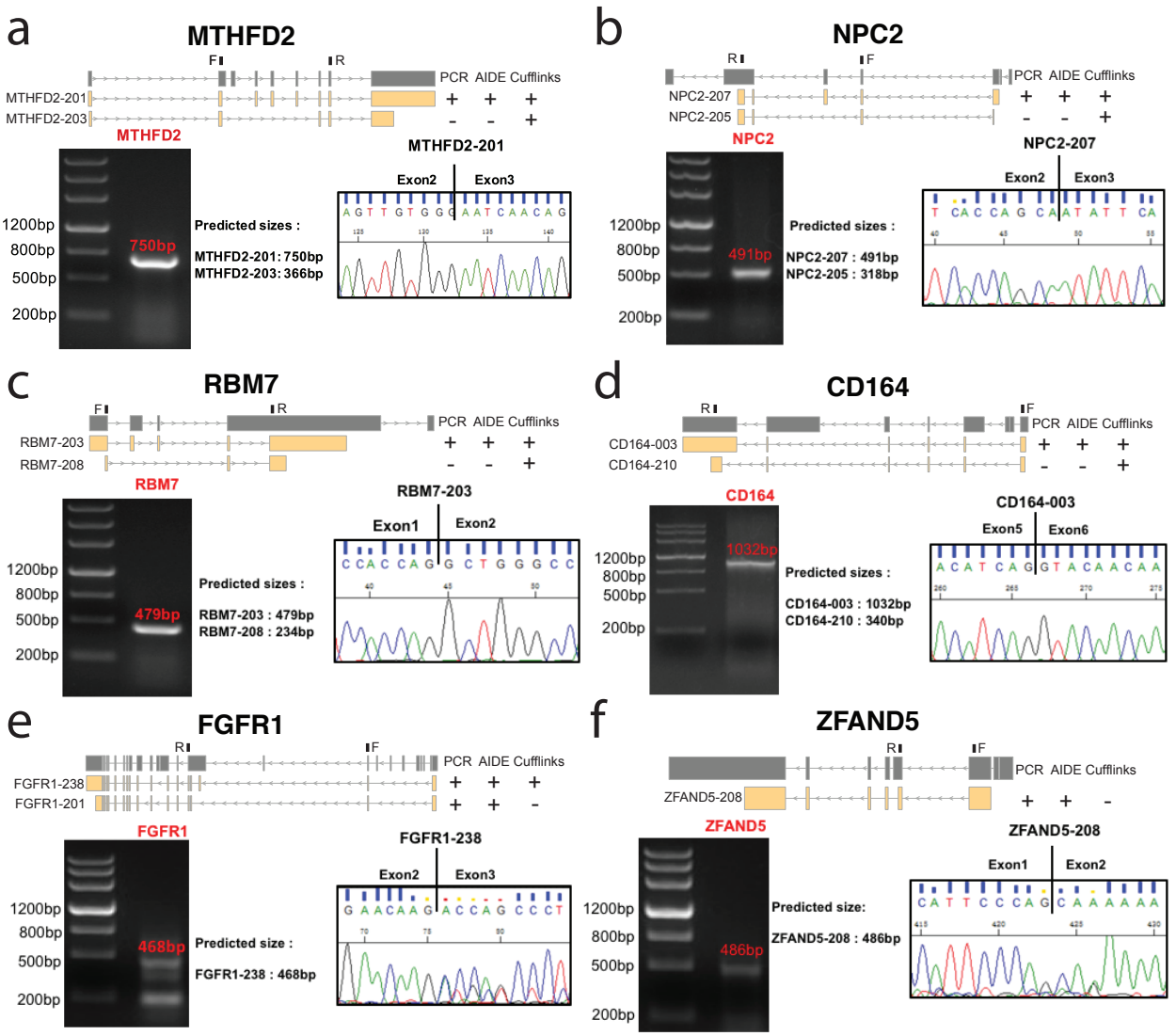

Figure 6: Experimental validation of isoforms predicted by AIDE and Cufflinks. Isoforms of genes *MTHFD2* (**a**), *NPC2* (**b**), *RBM7* (**c**), *CD164* (**d**), *FGFR1* (**e**), and *ZFAND5* (**f**) were validated by PCR and Sanger sequencing. The isoforms to validate (yellow) are listed under each gene (dark gray), with $+$ / $-$ indicating whether an isoform was / was not identified by PCR or a computational method. The forward (F) and reward (R) primers are marked on top of each gene. For each gene, the agarose gel electrophoresis result demonstrates the molecular lengths of PCR products, and the chromatography of Sanger sequencing confirms the sequences around the exon-exon junction unique to the PCR validated isoform.

steps (see Methods), leads to more precise and robust isoform discovery results. Here we consider $2,262$ protein-coding genes from the human GENCODE annotation (version 24) [8]. We treat the annotated isoforms as the true isoforms, and we simulate paired-end RNA-seq reads from those isoforms with pre-determined abundance levels (simulation details in Methods). For every gene, we apply AIDE, which uses stepwise selection, and its counterpart AIDEf, which only uses forward selection, to discover isoforms from the simulated reads. At each forward or backward selection step, we set the significance level of the likelihood ratio test as $4.42 \times 10^{-6} =$

$0.01/2262$ based on the Bonferroni correction. To evaluate the robustness of AIDE to the accuracy of annotation, we consider three types of annotation input into AIDE and AIDEf: (1) "N" (no) annotations: no annotated isoforms are used, and both AIDE and AIDEf are initialized with the candidate isoform compatible with the most reads and directly go into stage 2 (Figure 1); (2) "I" (inaccurate) annotations: the "annotated isoforms" used in stage 1 (Figure 1) consist of half of the randomly selected true isoforms and the same number of false isoforms (not in the GENCODE annotation but randomly assembled from exons); (3) "A" (accurate) annotations: the "annotated isoforms" used in stage 1 (Figure 1) consist of half of the randomly selected true isoforms. Another factor we consider is the RNA-seq read coverage. An "nx" coverage means that a genomic locus is covered by n reads on average. We generate RNA-seq data with different read coverages to compare AIDE and AIDEf with each of the three types of annotations, and we evaluate the performance of AIDE and AIDEf in terms of the precision rate (i.e., the percentage of correctly discovered isoforms among all the discovered isoforms), the recall rate (i.e., the percentage of correctly discovered isoforms among the true isoforms), and the F score, which is the harmonic mean of the precision and recall rates.

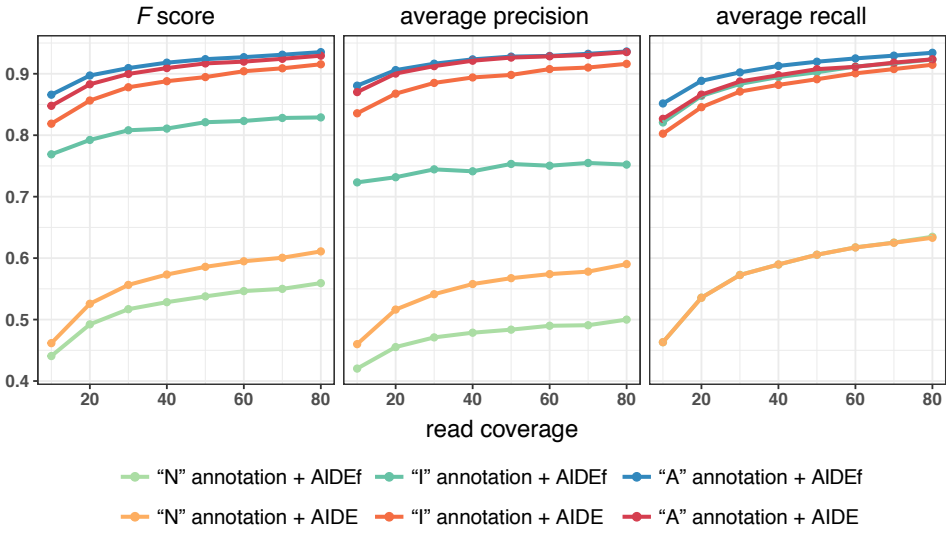

Figure 7: Comparison of AIDE (stepwise selection) and AIDEf (forward selection only) in terms of the average F scores, precision rates, and recall rates (i.e., three measures) across the $2,262$ genes in simulation. For AIDE and AIDEf, each measure is calculated based on three types of annotations and RNA-seq samples with varying read coverages. The vertical axes of the three panels denote the values of the three measures, and the horizontal axis denotes the average per-base coverage of RNA-seq reads.

The simulation results averaged from the $2,262$ genes show that AIDE and AIDEf perform the best when the "A" annotations are supplied, and they have the worst results with the "N" annotations, as expected (Figure 7). Given the "A" annotations, AIDE and AIDEf have similarly good performance. However, when supplied with the "I" annotations and the "N" annotations, AIDE has much better performance than AIDEf. Notably, the performance of AIDE with the "I" annotations is close to that with the "A" annotations, demonstrating the robustness of AIDE to

inaccurate annotations. On the other hand, forward selection has decreased precision rates when the "I" annotations instead of the "A" annotations are supplied because forward selection is incapable of removing non-expressed annotated isoforms from its identified isoform set in stage 1 (Figure 1). Given the "N" annotations, AIDE also has better performance than AIDEf. These results suggest that choosing stepwise selection over forward selection is reasonable for AIDE because perfectly accurate annotations are usually unavailable in real scenarios.

Figure 7 also suggests that both approaches exhibit improved performance with all the three types of annotations as the read coverages increase. Higher read coverages help the most when the "N" annotations are supplied, but its beneficial effects become more negligible with the "A" annotations. When the coverages increase from 10x to 80x, the F scores of AIDE increase by 32.6%, 12.3%, and 9.5% with the "N", "I", and "A" annotations, respectively. Moreover, we observe that the F scores of AIDE with the "N" annotations and the 80x coverage are approximately 30% lower than the F scores with the "A" annotations and the 30x coverage. This suggests that accurate annotations can assist isoform discovery and reduce the costs for deep sequencing depths to a large extent. The results also confirm that AIDE can effectively and selectively leverage useful information from annotations to increase its precision in isoform discovery.

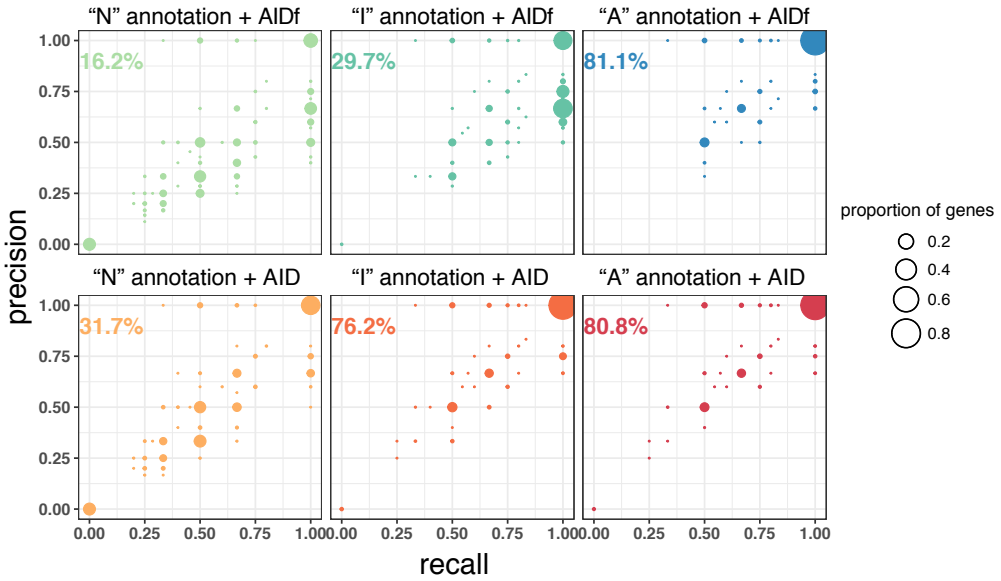

Figure 8: Comparison of AIDE (stepwise selection) and AIDEf (forward selection only) in terms of the per-gene precision and recall rates in simulation, with 80x read coverage and three types of annotations. The circle sizes are proportional to the proportions of genes with the corresponding precision and recall rates, with the top-right circles indicating 100% accuracy.

In addition to the above average performance of all the $2,262$ genes, we also summarize the precision and recall rates of AIDE at the individual gene level. Figure 8 illustrates the distribution of the precision and recall rates across the $2,262$ genes. When the "A" annotations are supplied, both AIDE and AIDEf achieve 100% precision and recall rates for over 80% of the genes. When

the “N” or “I” annotations are supplied, we observe a 2.0- or 2.6- fold increase in the number of genes with 100% precision and recall rates from AIDEf to AIDE. These results again demonstrate the effectiveness of AIDE in removing non-expressed annotated isoforms and identifying novel isoforms with higher accuracy due to its use of statistical model selection principles.

Discussion

We propose a new method AIDE to improve the precision of isoform discovery and the accuracy of isoform quantification from the next-generation RNA-seq data, by selectively borrowing alternative splicing information from annotations. AIDE iteratively identifies isoforms in a stepwise manner while placing priority on the annotated isoforms, and it performs statistical testing to automatically determine which isoforms to retain. We demonstrate the efficiency and superiority of AIDE compared to three state-of-the-art methods, Cufflinks, SLIDE, and StringTie, on multiple synthetic and real RNA-seq datasets followed by an experimental validation through PCR-Sanger sequencing, and the results suggest that AIDE leads to much more precise discovery of full-length RNA isoforms and more accurate isoform abundance estimation.

Due to technical limitations, the isoforms not amplified by PCR may still exist at an extremely low level. We attempted to reduce this possibility by optimizing the design and parameters used in our PCR experiments. First, we only validated the isoforms uniquely predicted by AIDE or Cufflinks if those isoforms have comparable abundance estimates (in FPKM unit). Hence, the PCR amplifications started with similar template amounts. Second, we designed the PCR primers to preferably amplify the isoforms predicted by Cufflinks (Table S3), so that if Cufflinks correctly identifies an isoform, the PCR experiment would capture it with high confidence. Specifically, when PCR primers are compatible with multiple isoforms of different lengths but similar abundance levels, PCR reaction preferentially amplifies the shorter isoform(s). Therefore, we experimentally validated the genes for which the isoform predicted by AIDE is longer than the isoform predicted by Cufflinks, and we designed the primers to be compatible with both isoforms. Third, we performed extensive amplification by setting the PCR cycle number to 50. Therefore, if an isoform is not captured by the PCR, it either does not exist or has extremely low abundance level. In either case, the isoform is not supposed to be biologically functional. Given the above considerations in experimental design and the validation results (Figure 6), we could safely draw the conclusion that AIDE has unique advantages in identifying full-length mRNA isoforms with a high precision and reducing false discoveries.

To the best of our knowledge, AIDE is the first isoform discovery method that identifies isoforms by selectively leveraging information from annotations based on a statistically principled model selection approach. The stepwise likelihood ratio testing procedure in AIDE has multiple advantages. First, AIDE only selects the isoforms that significantly contribute to the explanation of the observed reads, leading to more precise results and reduced false discoveries than those of existing methods. Second, the forward steps allow AIDE to start from and naturally give

priority to the annotated isoforms, which have higher chances to be expressed in a given sample. Meanwhile, the backward steps allow AIDE to adjust its previously selected isoforms given its newly added isoform so that all the selected isoforms together better explain the observed reads. Third, the testing procedure in AIDE allows the users to adjust the conservatism and precision of the discovered isoforms according to their desired level of statistical significance. Because of these advantages, AIDE identifies fewer novel isoforms at a higher precision level than previous methods do, making it easier for biologists to experimentally validate the novel isoforms. In applications where the recall rate of isoform discovery is of great importance (i.e., the primary goal is to discover novel isoforms with a not-too-stringent criterion), users can increase the p-value threshold of AIDE to discover more novel isoforms.

Through the application of AIDE to multiple RNA-seq datasets, we demonstrate that selectively incorporating annotated splicing patterns, in addition to simply obtaining gene and exon boundaries from annotations, greatly helps isoform discovery. The stepwise selection in AIDE also differentiates it from the methods that directly assume the existence of all the annotated isoforms in an RNA-seq sample. The development and application of AIDE has lead us to interesting observations that could benefit both method developers and data users. First, we find that a good annotation can help reduce the need for deep sequencing depths. AIDE has been shown to achieve good accuracy on datasets with low sequencing depths when supplied with accurately annotated isoforms, and its accuracy is comparable to that based on deeply sequenced datasets. Second, we find it more important for an annotation to have high purity than to have high completeness, in order to improve the isoform discovery accuracy of AIDE and the other methods we compared with in our study. Ideally, instead of using all the annotated isoforms in isoform discovery tasks, a better choice is to use a filtered set of annotated isoforms with high confidence. This requires annotated isoforms to have confidence scores, which unfortunately are unavailable in most annotations. Therefore, how to add confidence scores to annotated isoforms becomes an important future research question, and answering this question will help the downstream computational prediction of novel isoforms.

In analysis tasks of discovering differential splicing patterns between RNA-seq samples from different biological conditions, a well-established practice is to first estimate the isoform abundance in each sample by using a method like Cufflinks, and then perform statistical testing to discover differentially expressed isoforms [43, 44]. However, as we have demonstrated in both synthetic and real data studies, existing methods suffer from high risks of predicting false positive isoforms, i.e., estimating non-zero expression levels for unexpressed isoforms in a sample. Such false positive isoforms will severely reduce the accuracy of differential splicing analysis, leading to inaccurate comparison results between samples under two conditions, e.g., healthy and pathological samples. In contrast, AIDE's conservative manner in leveraging the existing annotations allows it to identify truly expressed isoforms at a greater precision and subsequently estimate isoform abundance with a higher accuracy. We expect that the application of AIDE will increase the accuracy of differential splicing analysis, lower the experimental validation costs, and lead to new

biological discoveries at a higher confidence level.

The probabilistic model in AIDE is very flexible and can incorporate reads of varying lengths and generated by different platforms. The non-parametric approach to learning the read generating mechanism makes AIDE a data-driven method and does not depend on specific assumptions of the RNA-seq experiment protocols. Therefore, a natural extension of AIDE is to combine the short but more accurate reads from the next-generation technologies with the longer but more error-prone reads generated by new sequencing technologies such as PacBio [45] and Nanopore [46]. Jointly modeling of the two types of reads using the AIDE method has the potential to greatly improve the overall accuracy of isoform detection [47], since AIDE is shown to have better precision than existing methods, and longer RNA-seq reads capture more splicing junctions and can further improve the recall rate of AIDE. Aside from the stepwise selection procedure used by AIDE, another possible way to incorporate priority on the annotated isoforms in the probabilitic model is to add regularization terms only on the unannotated isoforms. However, this approach is less interpretable compared with AIDE, since the regularization terms lack direct statistical intepretations as the p-value threshold does. Moreover, this approach may lose control of the false discovery rate when the annotation has a low purity. Another future extension of AIDE is to jointly consider multiple RNA-seq samples for more robust and accurate transcript reconstruction. It has been shown that it is often possible to improve the accuracy of isoform quantification by integrating the information in multiple RNA-seq samples [38, 48, 49]. Especially, using our previous method MSIQ we have demonstrated that it is necessary to account for the possible heterogeneity in the quality of different samples to improve the robustness of isoform quantification [38]. Therefore, by extending AIDE to combine the consistent information from multiple technical or biological samples, it is likely to achieve better reconstruction accuracy, and enable researchers to integrate publicly available and new RNA-seq samples for transcriptome studies.

Methods

Isoform discovery and abundance estimation using AIDE

The AIDE method is designed to identify and quantify the mRNA isoforms of an individual gene. Suppose that a gene has m non-overlapping exons and J candidate isoforms. If no filtering steps based on prior knowledge or external information is applied to reduce the set of candidate isoforms, J equals $2^m - 1$, the number of all possible combinations of exons into isoforms. The observed data are the n RNA-seq reads mapped to the gene: $\boldsymbol{R} = \{r_1, \ldots, r_n\}$. The parameters we would like to estimate are the isoform proportions $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)'$, where α_j is the proportion of isoform j among all the isoforms (i.e., the probability that a random RNA-seq read is from isoform j) and $\sum_{j=1}^{J} \alpha_j = 1$. We also introduce hidden variables $\boldsymbol{Z} = \{Z_1, \ldots, Z_n\}$ to denote the isoform origins of the n reads, with $Z_i = j$ indicating that read r_i is from isoform j, and $P(Z_i = j) = \alpha_j$, for $i = 1, \ldots, n$.

The joint probability of read r_i and its isoform origin can be written as:

$$\begin{aligned} P\left(r_i, Z_i|\boldsymbol{\alpha}\right) &= \prod_{j=1}^{J} P\left(r_i, Z_i = j|\boldsymbol{\alpha}\right)^{I\{Z_i=j\}} \\ &= \prod_{j=1}^{J} \left[P\left(r_i|Z_i = j\right)\alpha_j\right]^{I\{Z_i=j\}} \\ &\triangleq \prod_{j=1}^{J} \left(h_{ij}\alpha_j\right)^{I_{ij}}, \end{aligned}$$

where $I_{ij} \triangleq I\{Z_i = j\}$ indicates whether read r_i is from isoform j, and $h_{ij} \triangleq P\left(r_i|Z_i = j\right)$ is the generating probability of read r_i given isoform j, calculated based on the read generating mechanism described in the following subsection.

Read generating mechanism

We have defined h_{ij} as the generating probability of read r_i given isoform j. Specifically, if read r_i is not compatible with isoform j (read r_i contains regions not overlapping with isoform j, or vice versa), then $h_{ij} = 0$; otherwise,

$$\begin{aligned} h_{ij} &= P(\text{starting position of read } r_i \mid \text{isoform } j)P(\text{fragment length of read } r_i \mid \text{isoform } j) \\ &\triangleq P^s_{ij}P^f_{ij}\,. \end{aligned}$$

In the literature, different models have been used to calculate the starting position distribution P^s and the fragment length distribution P^f. Most of these models are built upon a basic model:

$$\begin{aligned} P^s_{ij} &= 1/L_j\,, \\ P^f_{ij} &= \frac{1}{\sqrt{2\pi}\sigma_f}\exp\left\{-\frac{(l_{ij}-\mu_f)^2}{2\sigma_f^2}\right\}, \end{aligned}$$

where L_j is the effective length of isoform j (the isoform length minus the read length), and l_{ij} is the length of fragment i given that read r_i comes from isoform j. However, this basic model does not account for factors like the GC-content bias or the positional bias (see Introduction). Research has shown that these biases affect read coverages differently, depending on different experimental protocols. For example, reverse-transcription with poly-dT oligomers results in an over-representation of reads in the 3' ends of isoforms, while reverse-transcription with random hexamers results in an under-representation of reads in the 3' ends of isoforms [50]. Similarly, different fragmentation protocols have varying effects on the distribution of reads within an isoform [22].

Given these facts, we decide to use a non-parametric method to estimate the distribution P^s of read starting positions, because non-parametric estimation is intrinsically capable of accounting for

the differences in the distribution due to different protocols. We use a multivariate kernel regression to infer P^s from the reads mapped to the annotated single-isoform genes. Suppose there are a total of c_s exons in the single-isoform genes. For $k' = 1, \ldots, c_s$, we use $q_{k'}$ to denote the proportion of reads, whose starting positions are in exon k', among all the reads mapped to the gene containing exon k'. Given any gene with J isoforms, suppose there are c_j exons in its isoform j, $j = 1, \ldots, J$. We estimate the conditional probability that a (random) read r_i starts from exon k $(k = 1, 2, \ldots, c_j)$, given that the read is generated from isoform j, as:

$$P^s_{ij}(b_i = k) \propto \frac{\sum_{k'=1}^{c_s} \prod_{d=1}^{3} \frac{1}{h_d} K\left(\frac{x_{kd} - x_{k'd}}{h_d}\right) q_{k'}}{\sum_{k'=1}^{c_s} \prod_{d=1}^{3} \frac{1}{h_d} K\left(\frac{x_{kd} - x_{k'd}}{h_d}\right)}, \text{ such that } \sum_{k=1}^{c_j} P^s_{ij}(b_i = k) = 1, \tag{1}$$

where b_i denotes the (random) index of the exon containing the starting position of the read r_i. When $b_i = k$, we use x_{k1}, x_{k2}, and x_{k3} to denote the GC content, the relative position, and the length of exon k, respectively. The GC content of exon k, x_{k1}, is defined as the proportion of nucleotides G and C in the sequence of exon k. The relative position of exon k, x_{k2}, is calculated by first linearly mapping the genomic positions of isoform j to $[0, 1]$ (i.e., the start and end positions are mapped to 0 and 1 respectively) and then locating the mapped position of the center position of the exon. For example, if isoform j spans from position 100 to position 1100 in a chromosome, and the center position of the exon is 200 in the same chromosome, then the relative position of this exon is 0.1. The meaning of $x_{k'd}$ ($k' = 1, \ldots, c_s$; $d = 1, \ldots, 3$) are be defined in the same way. The kernel function $K(\cdot)$ is set as the Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2)$. h_d denotes the bandwidth of dimension d and is selected by cross validation. The whole estimation procedure of P^s_{ij} is implemented through the R package `np`.

As for the fragment length distribution P^f, we assume that it follows a truncated log normal distribution. This is because mRNA fragments that are too long or too short are filtered out in the library preparation step before the sequencing step. In addition, the empirical fragment length distribution is usually skewed to right instead of being symmetric (Figure S1). Therefore, a truncated log normal distribution generally fits well the empirical distribution of fragment lengths:

$$P^f_{ij} = \begin{cases} \dfrac{\sqrt{2}\exp\left(-\left[\log\left(\frac{l_{ij}}{m_f}\right)\right]^2 \Big/ 2\sigma_f^2\right)}{\sqrt{\pi}\sigma_f\left[\text{erf}\left(\frac{1}{\sqrt{2}\sigma_f}\log\left(\frac{t_u^f}{m_f}\right)\right) - \text{erf}\left(\frac{1}{\sqrt{2}\sigma_f}\log\left(\frac{t_l^f}{m_f}\right)\right)\right] l_{ij}}, & \text{if } l_{ij} \in [t_l^f, t_u^f] \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

where m_f and σ_f are the median and the shape parameter of the distribution, respectively; t_l^f is the lower truncation threshold, and t_u^f is the upper truncation threshold. The function erf$(\cdot)$ (the "error function" encountered in integrating the normal distribution) is defined as erf$(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$.

The probabilistic model and parameter estimation in AIDE

Given the aforementioned settings, the joint probability of all the observed and hidden data is

$$\begin{aligned} P\left(\boldsymbol{R}, \boldsymbol{Z} | \boldsymbol{\alpha}\right) &= \prod_{i=1}^{n} P\left(r_i, Z_i | \boldsymbol{\alpha}\right) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{J} \left(h_{ij}\alpha_j\right)^{I_{ij}} \\ &= \prod_{i=1}^{n} \prod_{j=1}^{J} \left(P_{ij}^{s} P_{ij}^{f} \alpha_j\right)^{I_{ij}} , \end{aligned}$$

where P_{ij}^{s} and P_{ij}^{f} are defined in Equations (1) and (2), and the complete log-likelihood is

$$\ell\left(\boldsymbol{\alpha} | \boldsymbol{R}, \boldsymbol{Z}\right) = \sum_{i=1}^{n} \sum_{j=1}^{J} I_{ij} \log\left(P_{ij}^{s} P_{ij}^{f} \alpha_j\right) . \tag{3}$$

However, as $\boldsymbol{Z}$ and the resulting I_{ij}'s are unobservable, the problem of isoform discovery becomes to estimate $\boldsymbol{\alpha}$ via maximizing the log-likelihood based on the observed data:

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= \underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, \ell\left(\boldsymbol{\alpha} | \boldsymbol{R}\right) \\ &= \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \log P(\boldsymbol{R} | \boldsymbol{\alpha}) \\ &= \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \log \left[\prod_{i=1}^{n} \left(\sum_{j=1}^{J} P_{ij}^{s} P_{ij}^{f} \alpha_j\right)\right] \\ &= \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{J} P_{ij}^{s} P_{ij}^{f} \alpha_j\right) , \end{aligned} \tag{4}$$

subject to $\alpha_j \geq 0$ and $\sum_{j=1}^{J} \alpha_j = 1$. To directly solve (4) is not easy, so we use on the expectation-maximization (EM) algorithm along with the complete log-likelihood (3), and it follows that we can iteratively update the estimated isoform proportions as

$$\alpha_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \frac{P_{ij}^{s} P_{ij}^{f} \alpha_j^{(t)}}{\sum_{j'=1}^{J} P_{ij'}^{s} P_{ij'}^{f} \alpha_{j'}^{(t)}} .$$

As the algorithm converges, we obtain the estimated isoform proportion $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_J)'$.

Stepwise selection in AIDE

If we directly consider all the $J = 2^m - 1$ candidate isoforms in formula (4) and calculate $\hat{\boldsymbol{\alpha}}$, the problem is unidentifiable when $J > n$. Even when $J \leq n$, this may lead to many falsely discovered

isoforms whose $\hat{\alpha}_j > 0$, especially for complex genes, because the most complex model with all the possible candidate isoforms would best explain the observed reads. Therefore, instead of directly using the EM algorithm to maximize the log-likelihood with all the possible candidate isoforms, we perform a stepwise selection of isoforms based on the likelihood ratio test (LRT). This approach has two advantages. On the one hand, we can start from a set of candidate isoforms with high confidence based on prior knowledge, and then we can sequentially add new isoforms to account for reads that cannot be fully explained by existing candidate isoforms. For example, a common case is to start with annotated isoforms. On the other hand, the stepwise selection by LRT intrinsically introduce sparsity into the isoform discovery process. Even though the candidate isoform pool can be huge when a gene has a large number of exons, the set of expressed isoforms is usually much smaller in a specific biological sample. LRT can assist us in deciding a termination point where adding more isoforms does not further improve the likelihood.

The stepwise selection consists of steps with two opposite directions: the forward step and the backward step (Figure 1). The forward step aims at finding a new isoform to best explain the RNA-seq reads and significantly improve the likelihood given the already selected isoforms. The backward step aims at rectifying the isoform set by removing the isoform with the most trivial contribution among the selected isoforms. Since stepwise selection is in a greedy-search manner, some forward steps, especially those taken in the early iterations, may not be the globally optimal options. Therefore, backward steps are necessary to correct the search process and result in a better solution path for the purpose of isoform discovery.

We separate the search process into two stages. We use stepwise selection to update the identified isoforms at both stages, but the initial isoform sets and the candidate sets are different in the two stages. Stage 1 starts with a single annotated isoforms that explains the most number of reads, and it considers all the annotated isoforms as the candidate isoforms. Stage 1 stops when the the forward step can no longer finds an isoform to add to the identified isoform set, i.e., the LRT does not reject the null hypothesis given the p-value threshold. Stage 2 starts with the isoforms identified in stage 1, and considers all the possible isoforms, including the annotated isoforms not chosen in stage 1, as the candidate isoforms (Figure 1). The initial isoform set is denoted as $S_1^{(0)}$ (stage 1) or $S_2^{(0)}$ (stage 2), the candidate isoform set is denoted as C_1 (stage 1) or C_2 (stage 2), and the annotation set is denoted as A. At stage 1, the initial set and candidate set are respectively defined as

$$S_1^{(0)} = \left\{ \underset{j \in A}{\operatorname{argmax}} (\text{number of reads compatible with isoform } j) \right\},$$
$$C_1 = A\,.$$

Suppose the stepwise selection completes after t_1 steps in stage 1, and the estimated isoform proportions after step t $(t = 1, 2, \ldots, t_1)$ are denoted as $\hat{\boldsymbol{\alpha}}^{(t)} = \left(\hat{\alpha}_1^{(t)}, \ldots, \hat{\alpha}_J^{(t)}\right)'$. Note that $\forall j \notin C_1$,

$\hat{\alpha}_j^{(t)} = 0$ in stage 1. At stage 2, the initial set and the candidate set are respectively defined as

$$S_2^{(0)} = \left\{ j : \hat{\alpha}_j^{(t_1)} > 0 \right\} ,$$
$$C_2 = \{1, 2, \ldots, J = 2^m - 1\} .$$

Here we introduce how to perform forward and backward selection based on a defined initial isoform set $S^{(0)}$ and a candidate set C. We ignore the stage number subscripts for notation simplicity. At both stages, we first estimate the expression levels of the initial isoform set $S^{(0)}$:

$$\hat{\boldsymbol{\alpha}}^{(0)} = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}}\, \ell\left(\boldsymbol{\alpha}|\boldsymbol{R}\right) = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(0)}} P_{ij}^s P_{ij}^f \alpha_j \right) ,$$

subject to $\alpha_j \geq 0$ if $j \in S^{(0)}$, $\alpha_j = 0$ if $j \notin S^{(0)}$, and $\sum_{j \in S^{(0)}} \alpha_j = 1$, based on the EM algorithm.

Forward step

The identified isoform set at step t is denoted as $S^{(t)} = \{j : \hat{\alpha}_j^{(t)} > 0\}$. The log-likelihood at step t is

$$\ell^{(t)} = \ell\left(\hat{\boldsymbol{\alpha}}^{(t)}|\boldsymbol{R}\right) = \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t)}} P_{ij}^s P_{ij}^f \hat{\alpha}_j^{(t)} \right) .$$

At step $(t+1)$, we consider adding one isoform $k \in C \backslash S^{(t)}$ into $S^{(t)}$ as a forward step. Given $S^{(t)}$ and k, we estimate the corresponding isoform proportions as

$$\hat{\boldsymbol{\alpha}}^{(t,k)} = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t)} \cup \{k\}} P_{ij}^s P_{ij}^f \alpha_j \right) ,$$

subject to $\alpha_j \geq 0$ if $j \in S^{(t)} \cup \{k\}$, and $\alpha_j = 0$ otherwise. Then we choose the isoform

$$k^* = \underset{k}{\operatorname{argmax}} \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t)} \cup \{k\}} P_{ij}^s P_{ij}^f \hat{\alpha}_j^{(t,k)} \right) ,$$

which maximizes the likelihood among all the newly added isoforms. Then the log-likelihood with the addition of this isoform k^* becomes

$$\ell^* = \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t)} \cup \{k^*\}} P_{ij}^s P_{ij}^f \hat{\alpha}_j^{(t,k^*)} \right) .$$

To decide whether to follow the forward step and add isoform k^* the identified isoform set, we use a likelihood ratio test (LRT) to test the null hypothesis ($H_0 : S^{(t)}$ is the true isoform set from which the RNA-seq reads were generated) against the alternative hypothesis ($H_a : S^{(t)} \cup \{k^*\}$ is the true isoform set). Under H_0 we asymptotically have

$$-2(\ell^{(t)} - \ell^*) \sim \chi^2(1)\,.$$

If the null hypothesis is rejected at a pre-specified significance level (i.e., p-value threshold), then $S^{(t+1)} = S^{(t)} \cup \{k^*\}$, $\hat{\boldsymbol{\alpha}}^{(t+1)} = \hat{\boldsymbol{\alpha}}^{(t,k^*)}$, and the log-likelihood is updated as

$$\ell^{(t+1)} = \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t+1)}} P_{ij}^s P_{ij}^f \hat{\alpha}_j^{(t+1)} \right) .$$

Otherwise, $S^{(t+1)} = S^{(t)}$, $\hat{\boldsymbol{\alpha}}^{(t+1)} = \hat{\boldsymbol{\alpha}}^{(t)}$, and $\ell^{(t+1)} = \ell^{(t)}$.

Backward step

Every time we add an isoform to the identified isoform set in a forward step (say the updated isoform set is $S^{(t+1)}$), we subsequently consider possibly removing one isoform $k \in S^{(t+1)}$ from $S^{(t+1)}$ in a backward step. Given $S^{(t+1)}$ and k, we estimate the corresponding isoform proportions as

$$\hat{\boldsymbol{\alpha}}^{(t+1,-k)} = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t+1)} \setminus \{k\}} P_{ij}^s P_{ij}^f \alpha_j \right) ,$$

subject to $\alpha_j \geq 0$ if $j \in S^{(t+1)} \setminus \{k\}$, and $\alpha_j = 0$ otherwise. Then we choose the isoform

$$k^- = \underset{k}{\operatorname{argmax}} \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t+1)} \setminus \{k\}} P_{ij}^s P_{ij}^f \hat{\alpha}_j^{(t+1,-k)} \right) ,$$

which maximizes the likelihood among all the isoforms in $S^{(t+1)}$. Then the log-likelihood with the removal of this isoform k^- becomes

$$\ell^- = \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t+1)} \setminus \{k^-\}} P_{ij}^s P_{ij}^f \hat{\alpha}_j^{(t+1,-k^-)} \right) .$$

To decide whether to follow the forward step and remove isoform k^- from the identified isoform set, we use a likelihood ratio test (LRT) to test the null hypothesis ($H_0 : S^{(t+1)} \setminus \{k^-\}$ is the true isoform set from which the RNA-seq reads were generated) against the alternative hypothesis ($H_a : S^{(t+1)}$

is the true isoform set). Under H_0 we asymptotically have

$$-2(l^- - l^{(t+1)}) \sim \chi^2(1)\,.$$

If the null hypothesis is not rejected at a pre-specified significance level (i.e., p-value threshold), then $S^{(t+2)} = S^{(t+1)} \backslash \{k^-\}$, $\hat{\boldsymbol{\alpha}}^{(t+2)} = \hat{\boldsymbol{\alpha}}^{(t+1,-k^-)}$, and the log-likelihood is updated as

$$\ell^{(t+2)} = \sum_{i=1}^{n} \log \left(\sum_{j \in S^{(t+2)}} P_{ij}^s P_{ij}^f \hat{\alpha}_j^{(t+2)} \right) .$$

Otherwise, $S^{(t+2)} = S^{(t+1)}$, $\hat{\boldsymbol{\alpha}}^{(t+2)} = \hat{\boldsymbol{\alpha}}^{(t+1)}$, and $\ell^{(t+2)} = \ell^{(t+1)}$.

In both stage 1 and stage 2, we iteratively consider the forward step and backward step and stop the algorithm at the first time when a forward step no longer adds an isoform to the identified set (Figure 1). To determine whether to reject a null hypothesis in a LRT, we set a threshold on the p-value. The default threshold is $\frac{0.01}{\text{number of genes considered for isoform discovery}}$. Unlike the thresholds set on the FPKM values or isoform proportions in other methods, this threshold on p-values allow users to tune the AIDE method based on their desired level of statistical significance. A larger threshold generally leads to more discovered isoforms and a better recall rate, while a smaller threshold leads to fewer discovered isoforms that are more precise.

Transcriptome profiling by next-generation sequencing

Breast cancer biopsies were collected from Clinical Research Center for Breast of West China Hospital under the supervision of China Association for Ethical Studies. The whole-transcriptome RNA sequencing was performed with the next-generation sequencing standard protocol using the sequencer Solexa (Illumina). Long RNA sequences were first converted into a library of cDNA fragments, and adaptors were subsequently added to each cDNA fragment. Solexa relies on the attachment of small DNA fragments to a platform, optically transparent surface, and solid-phase amplification to create an ultrahigh-density sequencing flow cell with > 10 million clusters, each containing $\sim 1,000$ copies of template/cm^2. These templates were sequenced by a robust four-color DNA sequencing-by-synthesis technology that employs reversible terminators with removable fluorescence. The high-sensitivity fluorescence was then detected by laser excitation. The typical output from a single reaction was approximately 2GB containing 20-40 million reads. The Solexa-sequenced reads were assembled using the CLC bio software. Before the assembly, the reads were preprocessed by masking the polyA tails and removing the adapters.

Validation of transcripts by PCR and Sanger sequencing

Biopsy was collected freshly and the total RNA was extracted with the RiboPure Kit (Ambion). The reverse transcription reactions were performed using the RevertAid First-Strand cDNA Synthesis System kit (ThermoFisher) . With the cDNAs as templates and primers from TSINGKE, the PCR procedure ($95°$C 5min, $95°$C 30s, $55°$C 30s, $72°$C 1min to 2min, 40 to 50 cycles, $72°$C 5min for extension) was conducted using a ThermoFisher PCR system. PCR products were purified using Gel Extraction Kit (OMEGA) followed by Sanger sequencing with their special forward primers. Finally, sequencing result was analyzed using the Sequence Scanner Software.

Proof-of-concept simulation strategy

We consider $2,262$ protein-coding genes from the human GENCODE annotation (version 24) [8]. The exon numbers of these genes range from 2 to 10. We treat the annotated isoforms in this GENCODE annotation as the true isoforms, randomly assign isoform abundance, and simulate paired-end RNA-seq reads from those isoforms with a 100 bp read length. As we know the underlying true isoforms and their abundance in this simulation, we can use them to evaluate the accuracy of AIDE.

Suppose that gene i has J_i annotated isoforms in the GENCODE annotation. We assume these J_i isoforms are the true isoforms, and we randomly select from them $\lceil \frac{J_i}{2} \rceil$ isoforms to be used in the "A" annotations or the "I" annotations. The isoform proportions of the true isoforms are simulated from a symmetric Dirichlet distribution with parameters $(1/\lceil \frac{J_i}{2} \rceil, \ldots, 1/\lceil \frac{J_i}{2} \rceil)'$. The total number of reads generated from gene i depends on the gene length and the read coverage. Within each transcript, the starting positions of reads are uniformly distributed, and the fragment length follows a log-normal distribution with mean 180 bp and standard deviation 65 bp. Once we simulate the reads, we find the chromosome coordinates of the two ends of each read, and those coordinates are the only observed data input into AIDE and AIDEf.

Acknowledgements

This work was supported by the following grants: UCLA Dissertation Year Fellowship (to W.V.L), National Science Foundation DMS-1613338 (to X.T. and J.J.L.), NIH/NIGMS R01GM120507, PhRMA Foundation Research Starter Grant in Informatics, Johnson & Johnson WiSTEM2D Award, and Sloan Research Fellowship (to J.J.L), and National Key Research and Development Program of China (No. 2016YFC0906000 [2016YFC0906003]), National Natural Science Foundation of China (No. 81773752), Key Program of the Science and Technology Bureau of Sichuan (No. 2017SZ00005), the Recruitment Program of Global Young Experts (known as "the Thousand Young Talents Plan") (to H.S.). The authors would like to thank the insightful feedbacks from Dr. Alexandar Hoffmann at UCLA.

Competing interests

The authors declare that they have no competing interests.

Declarations

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

References

[1] Claudia Ghigna, Cristina Valacca, and Giuseppe Biamonti. Alternative splicing and tumor progression. *Current genomics*, 9(8):556–570, 2008.

[2] Joan E Hooper. A survey of software for genome-wide discovery of differential splicing in rna-seq data. *Human genomics*, 8(1):3, 2014.

[3] J Adams. Transcriptome: connecting the genome to gene function. *Nat Educ*, 1(1):195, 2008.

[4] Guey-Shin Wang and Thomas A Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, 8(10):749–761, 2007.

[5] Daniel Mordes, Xiaoyan Luo, Amar Kar, David Kuo, Lili Xu, Kazuo Fushimi, Guowu Yu, Paul Sternberg Jr, and Jane Y Wu. Pre-mrna splicing and retinitis pigmentosa. *Molecular vision*, 12:1259, 2006.

[6] Natalia N Singh and Ravindra N Singh. Alternative splicing in spinal muscular atrophy underscores the role of an intron definition model. *RNA biology*, 8(4):600–606, 2011.

[7] Sagar Chhangawala, Gabe Rudy, Christopher E Mason, and Jeffrey A Rosenfeld. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome biology*, 16(1):131, 2015.

[8] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012.

[9] Yuting Ye and Jingyi Jessica Li. Nmfp: a non-negative matrix factorization based preselection method to increase accuracy of identifying mrna isoforms from rna-seq data. *BMC Genomics*, 17(1):127, 2016.

[10] Stefan Canzar, Sandro Andreotti, David Weese, Knut Reinert, and Gunnar W Klau. Cidane: Comprehensive isoform discovery and abundance estimation. *Genome biology*, 17(1):1, 2016.

[11] Aziz M Mezlini, Eric JM Smith, Marc Fiume, Orion Buske, Gleb L Savich, Sohrab Shah, Sam Aparicio, Derek Y Chiang, Anna Goldenberg, and Michael Brudno. ireckon: Simultaneous isoform discovery and abundance estimation from rna-seq data. *Genome research*, 23(3):519–529, 2013.

[12] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[13] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature biotechnology*, 28(5):503–510, 2010.

[14] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature biotechnology*, 33(3):290–295, 2015.

[15] Wei Li, Jianxing Feng, and Tao Jiang. Isolasso: a lasso regression approach to rna-seq based transcriptome assembly. *Journal of Computational Biology*, 18(11):1693–1707, 2011.

[16] Jingyi Jessica Li, Ci-Ren Jiang, James B Brown, Haiyan Huang, and Peter J Bickel. Sparse linear modeling of next-generation mrna sequencing (rna-seq) data for isoform discovery and abundance estimation. *Proceedings of the National Academy of Sciences*, 108(50):19867–19872, 2011.

[17] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[18] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[19] Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic acids research*, 36(16):e105–e105, 2008.

[20] Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. Gc-content normalization for rna-seq data. *BMC bioinformatics*, 12(1):480, 2011.

[21] Jun Li, Hui Jiang, and Wing Hung Wong. Modeling non-uniformity in short-read rates in rna-seq data. *Genome biology*, 11(5):1, 2010.

[22] Alyssa C Frazee, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Polyester: simulating rna-seq datasets with differential transcript expression. *Bioinformatics*, 31(17):2778–2784, 2015.

[23] Rob Patro, Geet Duggal, and Carl Kingsford. Salmon: accurate, versatile and ultrafast quantification from rna-seq data using lightweight-alignment. *bioRxiv*, page 021592, 2015.

[24] Wei Zheng, Lisa M Chung, and Hongyu Zhao. Bias detection and correction in rna-sequencing data. *BMC bioinformatics*, 12(1):1, 2011.

[25] Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38(12):e131–e131, 2010.

[26] Regina Bohnert and Gunnar Rätsch. rquant. web: a tool for rna-seq-based transcript quantitation. *Nucleic acids research*, 38(suppl 2):W348–W351, 2010.

[27] Adam Roberts, Cole Trapnell, Julie Donaghey, John L Rinn, and Lior Pachter. Improving rna-seq expression estimates by correcting for fragment bias. *Genome biology*, 12(3):1, 2011.

[28] Hui Jiang and Julia Salzman. A penalized likelihood approach for robust estimation of isoform expression. *Statistics and its interface*, 8(4):437, 2015.

[29] Michael I Love, John B Hogenesch, and Rafael A Irizarry. Modeling of rna-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature Biotechnology*, 34(12):1287–1291, 2016.

[30] Tamara Steijger, Josep F Abril, Pär G Engström, Felix Kokocinski, Tim J Hubbard, Roderic Guigó, Jennifer Harrow, Paul Bertone, RGASP Consortium, et al. Assessment of transcript reconstruction methods for rna-seq. *Nature methods*, 10(12):1177–1184, 2013.

[31] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):1, 2016.

[32] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, et al. Ensembl 2018. *Nucleic acids research*, 46(D1):D754–D761, 2017.

[33] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[34] Kate R Rosenbloom, Joel Armstrong, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, et al. The ucsc genome browser database: 2015 update. *Nucleic acids research*, 43(D1):D670–D681, 2015.

[35] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. Identification of novel transcripts in annotated genomes using rna-seq. *Bioinformatics*, 27(17):2325–2329, 2011.

[36] Bronwen L Aken, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos García Girón, Thibaut Hourlier, et al. The ensembl gene annotation system. *Database*, 2016, 2016.

[37] Pierre-Luc Germain, Alessandro Vitriolo, Antonio Adamo, Pasquale Laise, Vivek Das, and Giuseppe Testa. Rnaonthebench: computational and empirical resources for benchmarking rnaseq quantification and differential expression methods. *Nucleic acids research*, 44(11):5054–5067, 2016.

[38] Wei Vivian Li, Anqi Zhao, Shihua Zhang, Jingyi Jessica Li, et al. Msiq: Joint modeling of multiple rna-seq samples for accurate isoform quantification. *The Annals of Applied Statistics*, 12(1):510–539, 2018.

[39] Meghana M Kulkarni. Digital multiplexed gene expression analysis using the nanostring ncounter system. *Current Protocols in Molecular Biology*, pages 25B–10, 2011.

[40] Anne Limbourg, Johann von Felden, Kumaravelu Jagavelu, Kashyap Krishnasamy, L Christian Napp, Piyushkumar R Kapopara, Matthias Gaestel, Bernhard Schieffer, Johann Bauersachs, Florian P Limbourg, et al. Map-kinase activated protein kinase 2 links endothelial activation and monocyte/macrophage recruitment in arteriogenesis. *PloS one*, 10(10):e0138542, 2015.

[41] Liping Zhang, Limei Ran, Gabriela E Garcia, Xiaonan H Wang, Shuhua Han, Jie Du, and William E Mitch. Chemokine cxcl16 regulates neutrophil and macrophage infiltration into injured muscle, promoting muscle regeneration. *The American journal of pathology*, 175(6):2518–2527, 2009.

[42] Joachim L Schultze and Susanne V Schmidt. Molecular features of macrophage activation. In *Seminars in immunology*, volume 27, pages 416–423. Elsevier, 2015.

[43] Fatemeh Seyednasrollah, Asta Laiho, and Laura L Elo. Comparison of software packages for detecting differential expression in rna-seq studies. *Briefings in bioinformatics*, 16(1):59–70, 2013.

[44] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, 7(3):562, 2012.

[45] Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289, 2015.

[46] Ashley Byrne, Anna E Beaudin, Hugh E Olsen, Miten Jain, Charles Cole, Theron Palmer, Rebecca M DuBois, E Camilla Forsberg, Mark Akeson, and Christopher Vollmers. Nanopore long-read rnaseq reveals widespread transcriptional variation among the surface receptors of individual b cells. *Nature Communications*, 8:16027, 2017.

[47] Shuhua Fu, Yingke Ma, Hui Yao, Zhichao Xu, Shilin Chen, Jingyuan Song, Kin Fai Au, and Bonnie Berger. Idp-denovo: de novo transcriptome assembly and isoform annotation by hybrid sequencing. *Bioinformatics*, 1:9, 2018.

[48] Yen-Yi Lin, Phuong Dao, Faraz Hach, Marzieh Bakhshi, Fan Mo, Anna Lapuk, Colin Collins, and S Cenk Sahinalp. Cliiq: Accurate comparative detection and quantification of expressed isoforms in a population. In *Algorithms in Bioinformatics*, pages 178–189. Springer, 2012.

[49] Jonas Behr, André Kahles, Yi Zhong, Vipin T Sreedharan, Philipp Drewe, and Gunnar Rätsch. Mitie: Simultaneous rna-seq-based transcript identification and quantification in multiple samples. *Bioinformatics*, 29(20):2529–2538, 2013.

[50] Francesca Finotello, Enrico Lavezzo, Luca Bianco, Luisa Barzon, Paolo Mazzon, Paolo Fontana, Stefano Toppo, and Barbara Di Camillo. Reducing bias in rna sequencing data: a novel approach to compute counts. *BMC bioinformatics*, 15(1):S7, 2014.

Supplementary Materials

Calculation of precision and recall rate of isoform reconstruction

We denote the true isoform set in a RNA-seq sample as S and the discovered isoform set as D. We use $B(s)$ and $E(s)$ to denote the number of bases and the number of exons in isoform s, respectively. We calculate the isoform-level precision and recall rate of isoform discovery by comparing the two sets S and D. When compare the isoform s_1 in S and s_2 in D, we allow a small difference by requiring the number of mismatch bases to be smaller than $0.99B(s_1)$. The isoform-level precision and recall rates are calculated as

$$\text{precision}^{\text{isoform}} = \frac{|D \cap S|}{|D|}, \quad \text{recall}^{\text{isoform}} = \frac{|D \cap S|}{|S|}.$$

The exon-level precision and recall rates are calculated as

$$\text{precision}^{\text{exon}} = \frac{\sum_{s \in D \cap S} E(s)}{\sum_{s \in D} E(s)}, \quad \text{recall}^{\text{exon}} = \frac{\sum_{s \in D \cap S} E(s)}{\sum_{s \in S} E(s)}.$$

The base-level precision and recall rates are calculated as

$$\text{precision}^{\text{base}} = \frac{\sum_{s \in D \cap S} B(s)}{\sum_{s \in D} B(s)}, \quad \text{recall}^{\text{base}} = \frac{\sum_{s \in D \cap S} B(s)}{\sum_{s \in S} B(s)}.$$

Table S1: Description of real RNA-seq data sets used in the article

.

sample	cell type	read length	ENCODE accession number
1	HepG2	50×2	ENCFF084JYA
2	HepG2	50×2	ENCFF790CFB
3	HepG2	38×2	ENCFF916YZY, ENCFF800YJR
4	HepG2	38×2	ENCFF179TFY, ENCFF782TAX
5	HepG2	50×2	ENCFF168NGI
6	HepG2	50×2	ENCFF711DJN
sample	cell type	read length	GEO accession number
1	human ESC	76×2	GSE90225
2	human ESC	76×2	GSE33480
3	human ESC	101×2	GSE47626

Table S2: Experimental validation of identified transcripts

.

	Transcript	GENCODE v24	AIDE	Cufflinks	validated by PCR-Sanger sequencing
categoty 1	MTHFD2-203	+	-	+	No
	NPC2-205	+	-	+	No
	RBM7-208	+	-	+	No
	CD164-210	+	-	+	No
	XBP1-205	+	-	+	Yes
	SYNGR2-new	-	-	+	Yes
category 2	FGFR1-238	+	+	-	Yes
	ZFAND5-208	+	+	-	Yes
	BRCA2-207	+	+	-	No
	PDCD5-201	+	+	-	No

+, Isoform was predicted; -, isoform was not predicted

Table S3: Primers designed for PCR validation of identified transcripts

.

primer name	sequence	predicted size (bp)	PCR product size (bp)
MTHFD2-203-F	TTCTGGAAGGAAACTGGCCC	366	750
MTHFD2-203-R	ACCAACTTGGGTTTGGCAGT		
NPC2-205-F	AGTGAATGTGAGCCCATGCC	318	491
NPC2-205-R	TCTGCTACAGAGCACCTCCT		
RBM7-208-F	GAAGCGGATCGCACTCTCTT	234	479
RBM7-208-R	TGATCCAGAGGTGAAGAACCA		
CD164-210-F	ATCTCCAACGTAACCTCGGC	340	1032
CD164-210-R	TGTAGTTCCTTGTGTGGCATCT		
XBP1-205-F	CGACGGGACCCCTAAAGTTC	389	389
XBP1-205-R	AGGGGCTGAAACAACTTGGG		
SYNGR2-new-F	CTTCCCACCCGAAGTTGAGG	653	653
SYNGR2-new-R	GAAGGCTCCAGAGGTGCTG		
FGFR1-238-F	ACTGCAGAACTGGGATGTGG	468	468
FGFR1-238-R	CTACGGGCATACGGTTTGGT		
ZFAND5-208-F	TCGGGGAAGGGTCGGATTAT	486	486
ZFAND5-208-R	CGGGGTCTGGTTAGTCTCCT		
BRCA2-207-F	ACAAAGGCAACGCGTCTTTC	427	419
BRCA2-207-R	AGGCACATTCCATAGCTGCC		
PDCD5-201-F	ACGAGGAGCTTGAGGCGCTGA	408	413
PDCD5-201-R	TAGACTTGTTCCGTTAAGTTC		

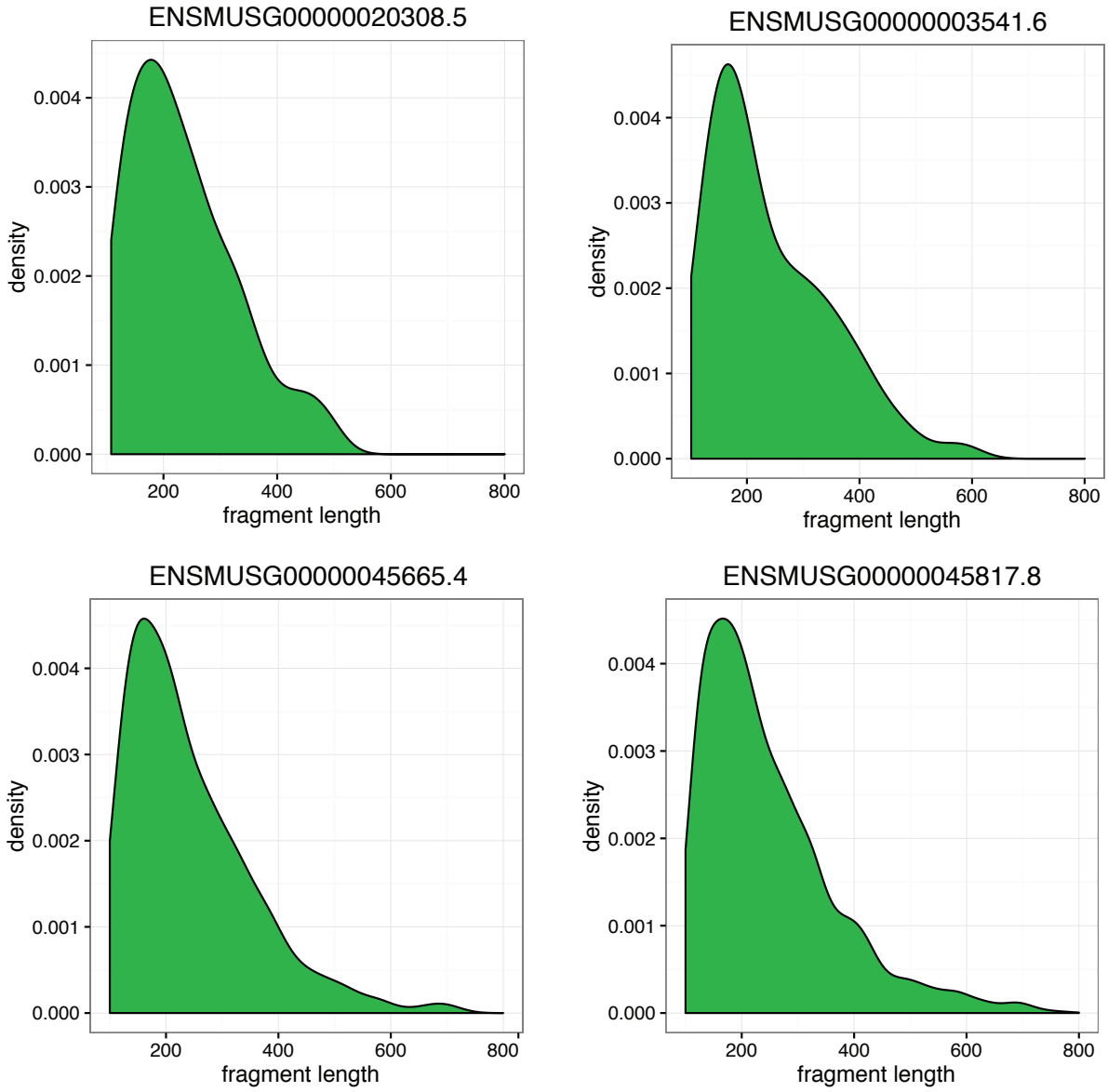

Figure S1: The distribution of fragment lengths in real RNA-seq data. The empirical fragment length distribution of four example mouse genes in the mouse bone marrow-derived macrophage dataset (Table S1).

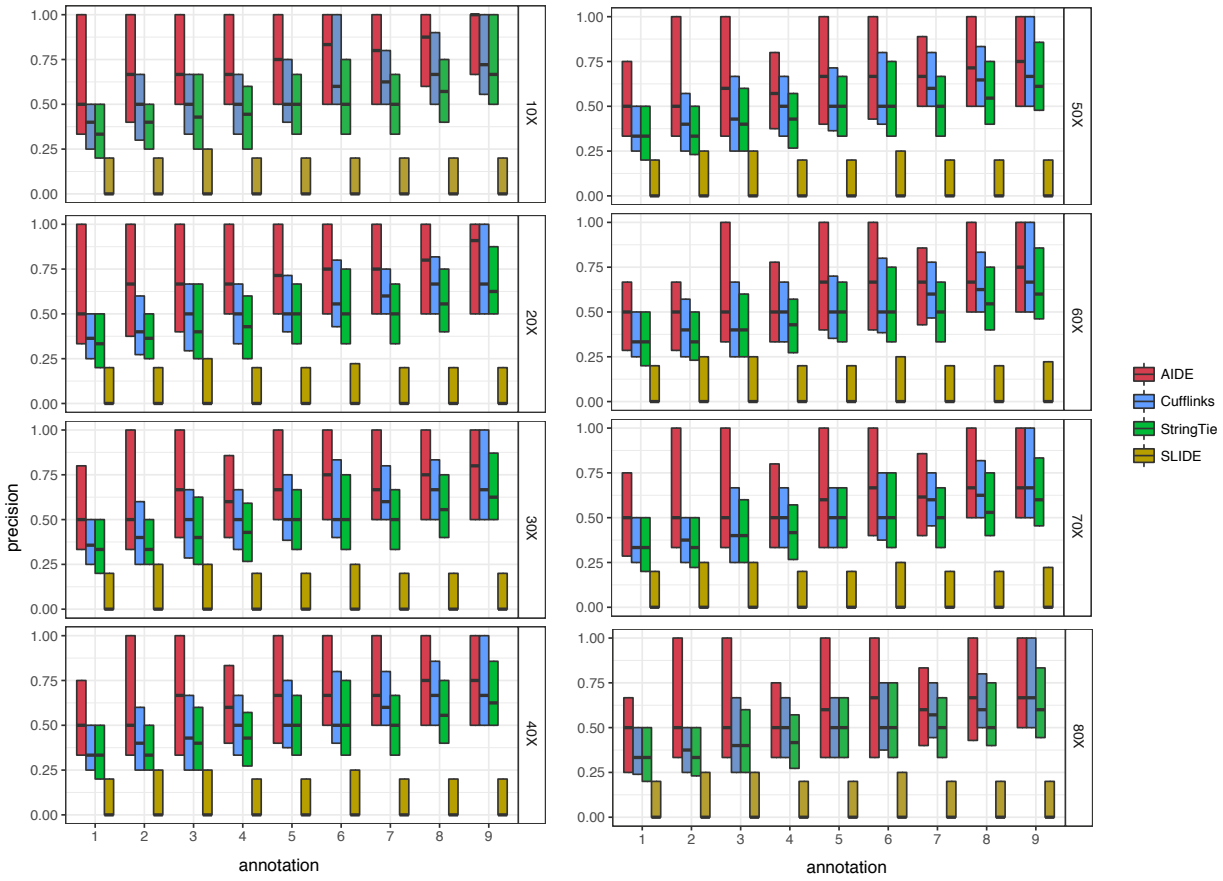

Figure S2: Distributions of per-gene precision rates of AIDE and the other three isoform discovery methods in simulation. Each box gives the 1st quantile, median, and 3rd quantile of the per-gene precision rates given the corresponding synthetic annotation set and read coverage.

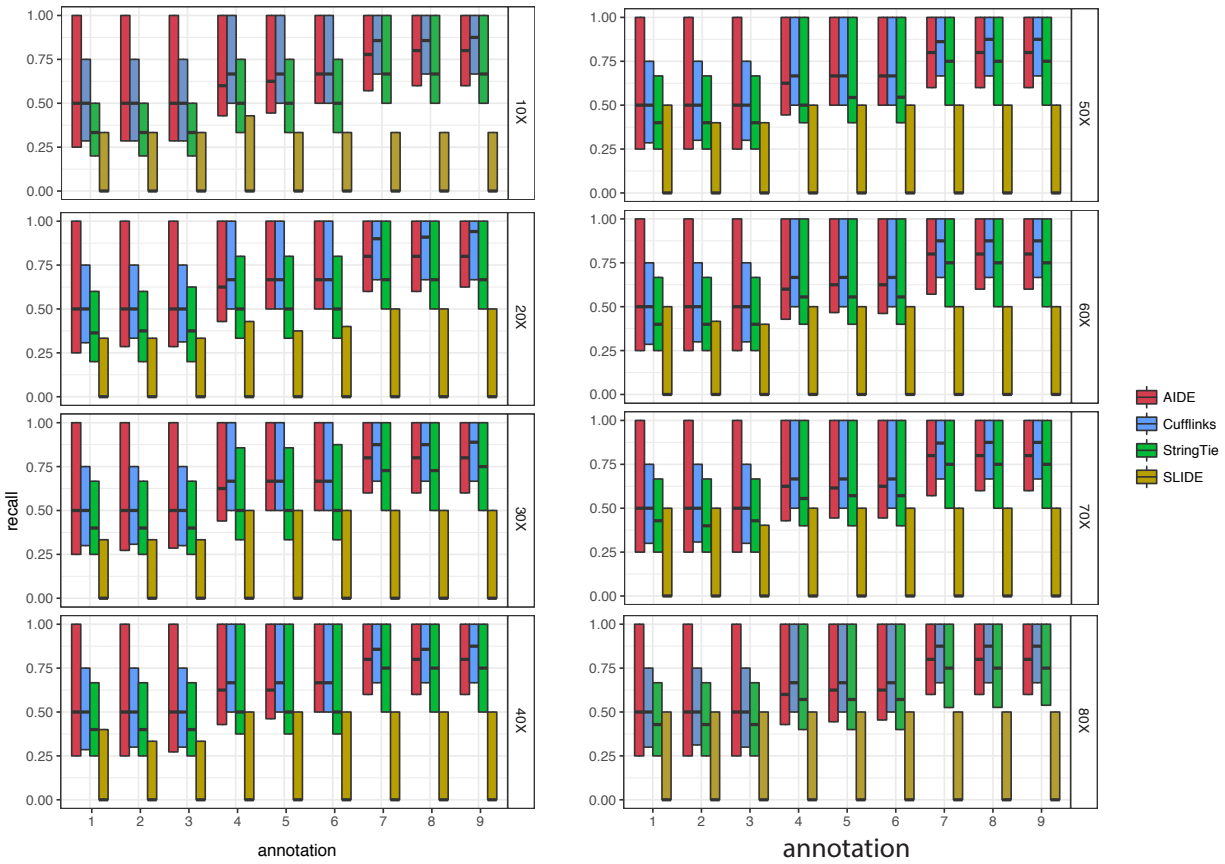

Figure S3: Distributions of per-gene recall rates of AIDE and the other three isoform discovery methods in simulation. Each box gives the 1st quantile, median, and 3rd quantile of the per-gene recall rates given the corresponding synthetic annotation set and read coverage.

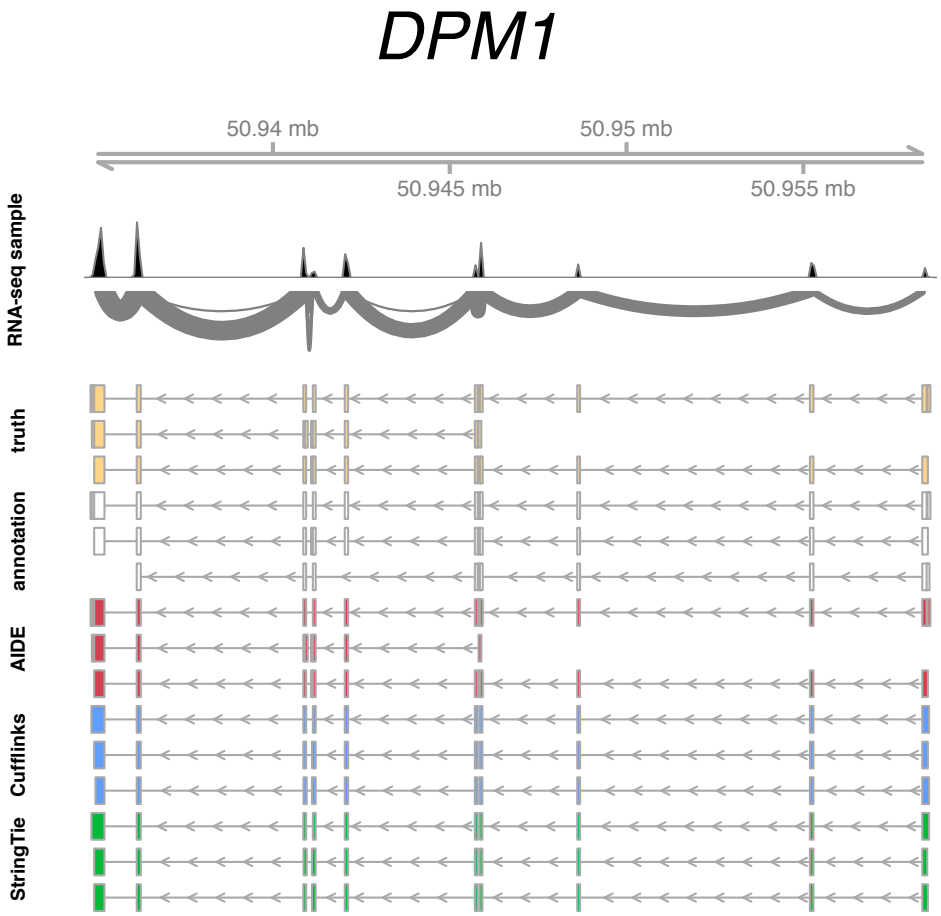

Figure S4: Isoform discovery for the human gene *DPM1* with the syethetic annotation set 1. The histogram and the sashimi plot denote the RNA-seq reads mapped to the gene *DPM1* in the RNA-seq data simulated by the `polyester` R package. The annotation (white) for this gene has a 67% purity and a 67% completeness, compared with the truly expressed isoforms (yellow). AIDE, Cufflinks, and StringTie each discovers three isoforms, but only AIDE is able to identify the shortest isoform missing in the annotation. SLIDE reports 17 isoforms, which are not displayed in the plot.

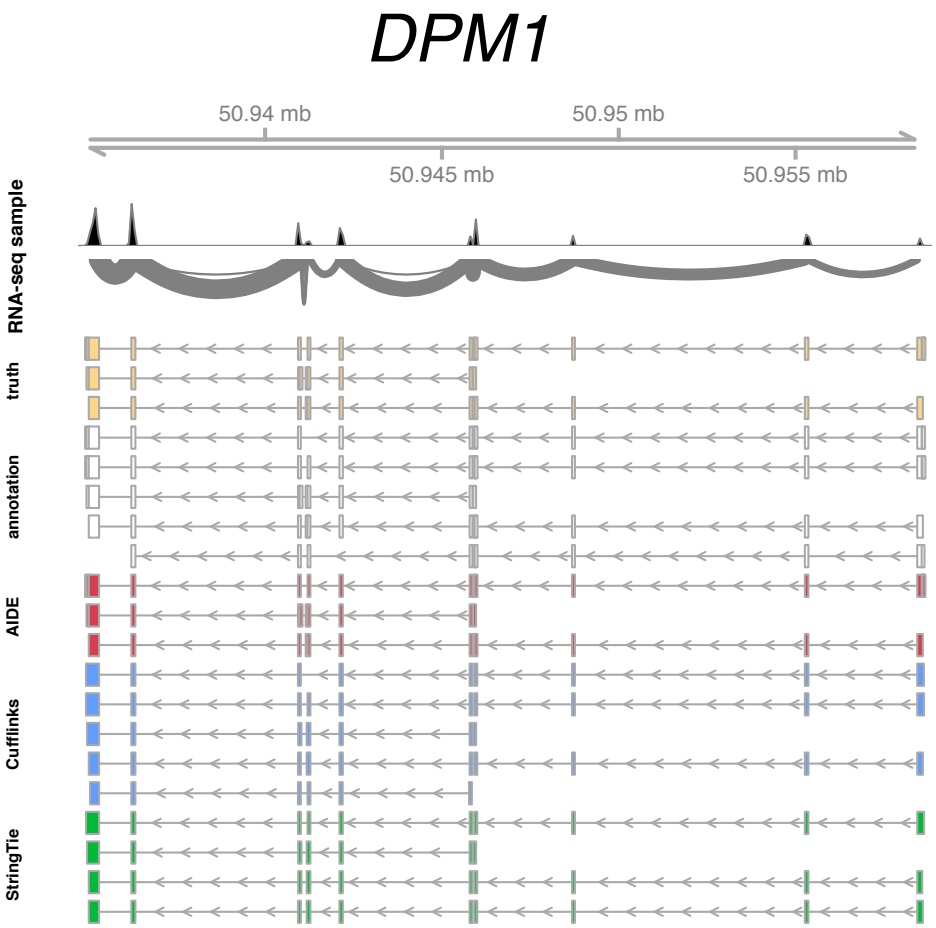

Figure S5: Isoform discovery for the human gene *DPM1* with the syethetic annotation set 9. The histogram and the sashimi plot denote the RNA-seq reads mapped to the gene *DPM1* in the RNA-seq data simulated by the `polyester` R package. The annotation (white) for this gene has a 60% purity and a 100% completeness, compared with the truly expressed isoforms (yellow). AIDE, Cufflinks, and StringTie respectively discovers three, five, and four isoforms, and only AIDE is able to identify the three true isoforms with 100% accuracy. SLIDE reports 20 isoforms, which are not displayed in the plot.

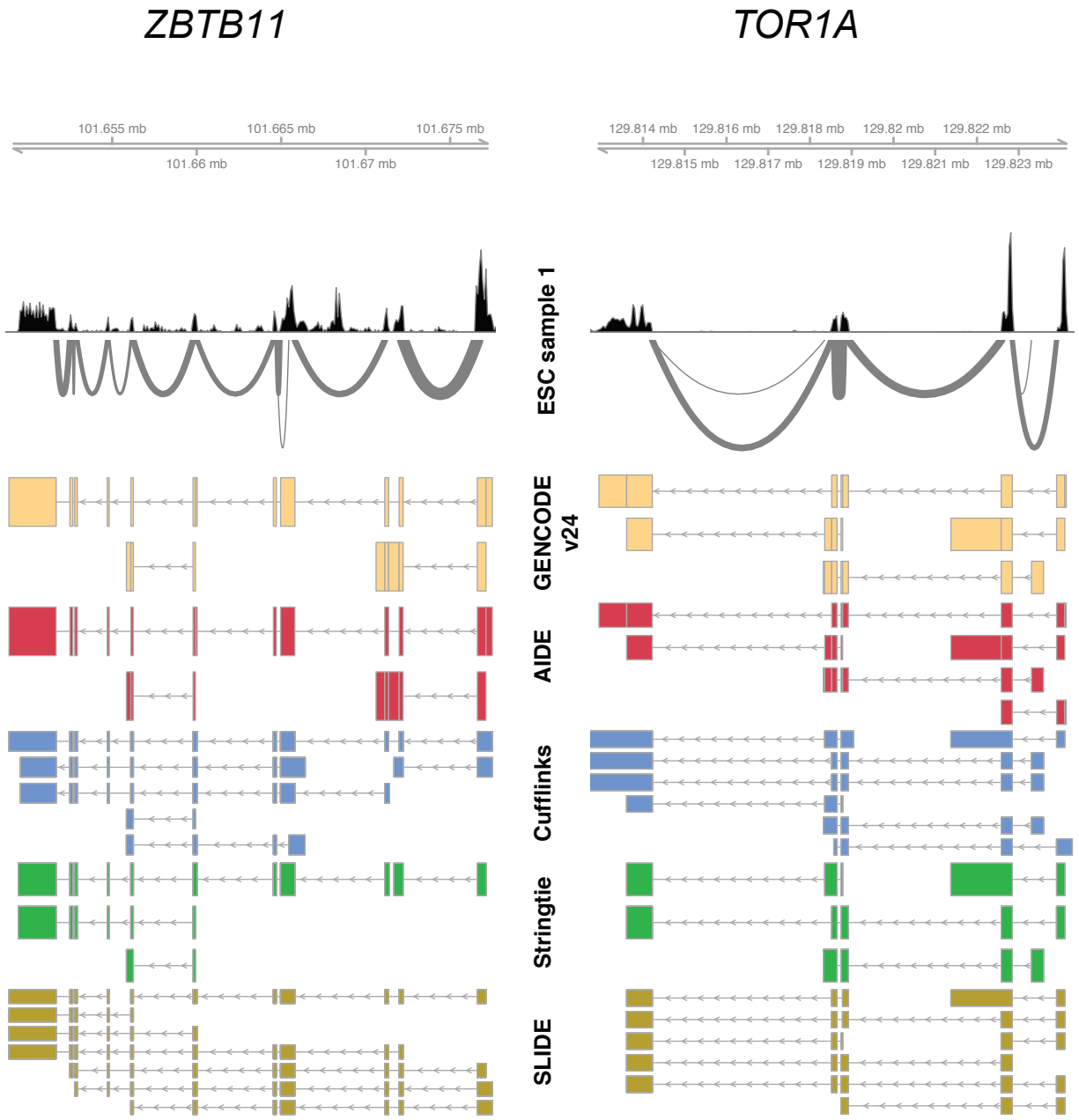

Figure S6: Isoform discovery for the human genes *ZBTB11* and *TOR1A* based on real data. The histogram and the sashimi plot denote the RNA-seq reads mapped to the two genes in the human ESC sample 1 (Table S1). Isoform discovery is based on the GENCODE human annotation version 24. AIDE achieves the best accuracy among the four methods.

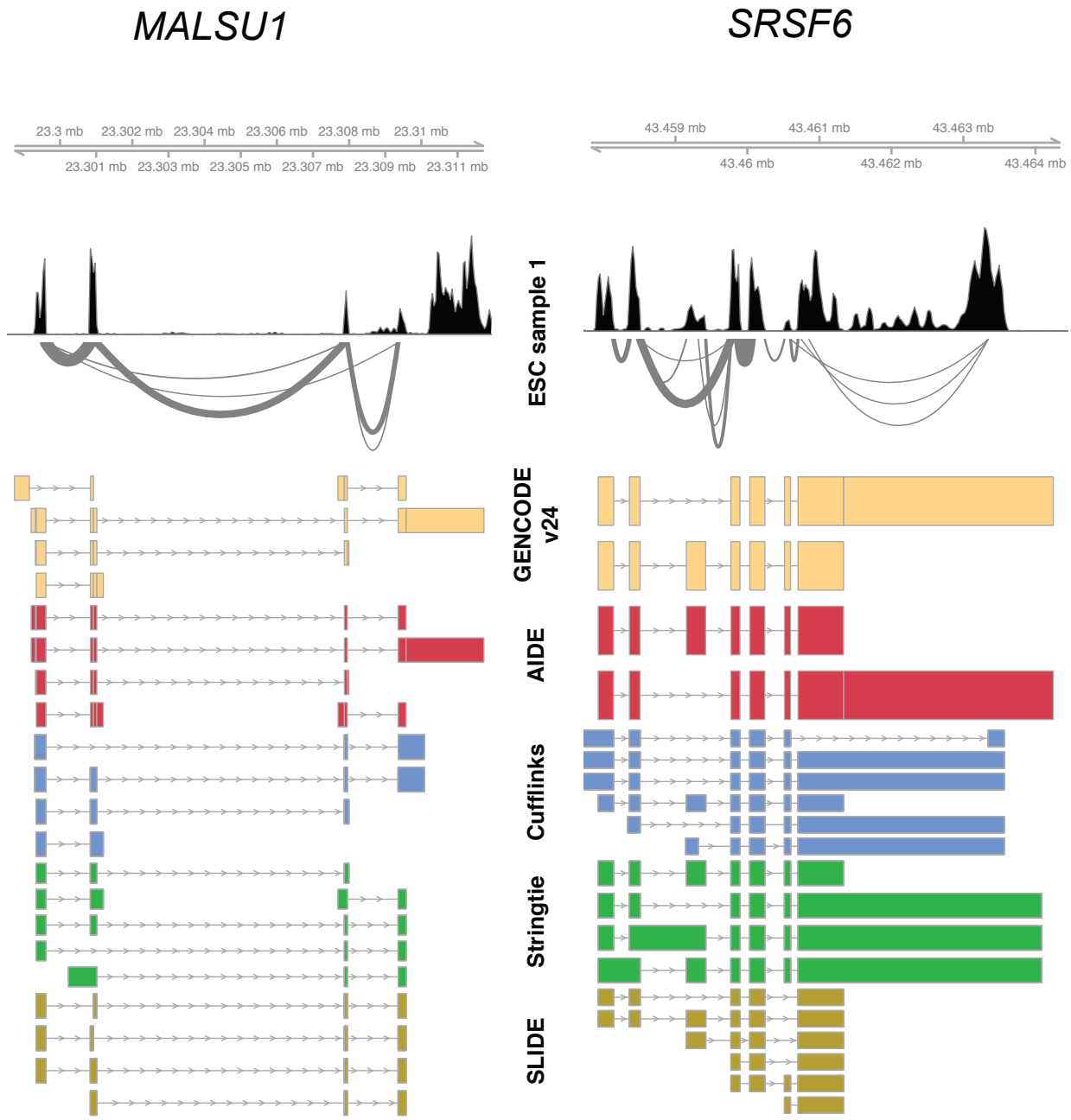

Figure S7: Isoform discovery for the human genes *MALSU1* and *SRSF6* based on real data. The histogram and the sashimi plot denote the RNA-seq reads mapped to the two genes in the human ESC sample 1 (Table S1). Isoform discovery is based on the GENCODE human annotation version 24. AIDE achieves the best accuracy among the four methods.

Figure S8: Isoform discovery for the human genes *ZBTB11* and *TOR1A* given different p-value thresholds. The histogram and the sashimi plot denote the RNA-seq reads mapped to the two genes in the human ESC sample 1 (Table S1). Isoform discovery is based on the GENCODE human annotation version 24. The threshold on the p-values resulted from the likelihood ratio tests decreases from 10^{-2} to 10^{-10}.

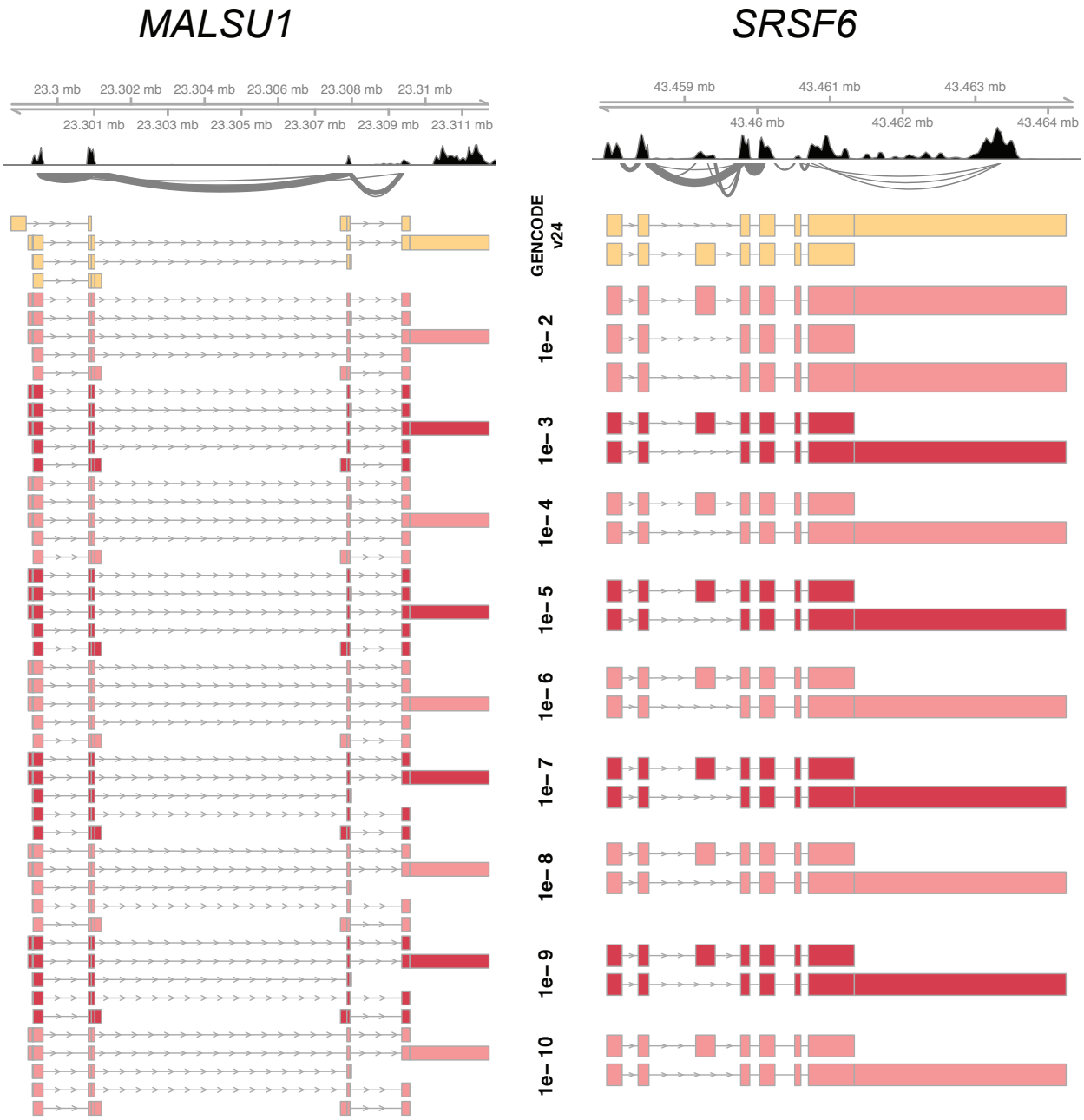

Figure S9: Isoform discovery for the human genes *MALSU1* and *SRSF6* given different p-value thresholds. The histogram and the sashimi plot denote the RNA-seq reads mapped to the two genes in the human ESC sample 1 (Table S1). Isoform discovery is based on the GENCODE human annotation version 24. The threshold on the p-values resulted from the likelihood ratio tests decreases from 10^{-2} to 10^{-10}.