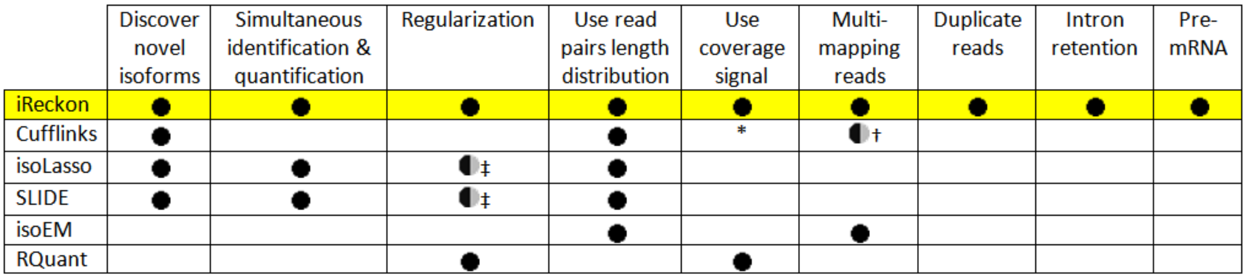
iReckon: Supplementary Material

Aziz M. Mezlini[1,2,3], Eric J. M. Smith[1], Marc Fiume[1], Orion Buske[1], Gleb L. Savich[4], Sohrab Shah[5,6], Sam Aparicio[5,6], Derek Y. Chiang[4], Anna Goldenberg[1,3] and Michael Brudno[1,2,3,7,*]

[1] Department of Computer Science, University of Toronto, Canada
[2] Centre for Computational Medicine, Hospital for Sick Children, Toronto, Canada
[3] Genetics and Genome Biology, Hospital for Sick Children, Toronto, Canada
[4] Department of Genetics, University of North Carolina, USA
[5] Dept of Molecular Oncology, BC Cancer Agency, Vancouver, BC, Canada
[6] Dept of Pathology, University Of British Columbia, Vancouver, BC, Canada
[7] Donnelly Centre, University of Toronto, Canada
[*] To whom correspondence should be addressed: brudno@cs.toronto.edu

1 Comparison of iReckon Features



	Discover novel isoforms	Simultaneous identification & quantification	Regularization	Use read pairs length distribution	Use coverage signal	Multi-mapping reads	Duplicate reads	Intron retention	Pre-mRNA
iReckon	●	●	●	●	●	●	●	●	●
Cufflinks	●			●	*	◐ †			
isoLasso	●	●	◐ ‡	●					
SLIDE	●	●	◐ ‡	●					
isoEM				●		●			
RQuant			●		●				

Fig. S1: Comparison of several methods based on the problems/ features addressed. ∗ Cufflinks does not use coverage signal for abundances estimation (Only for isoforms discovery). † Cufflinks has only one corrective step for reallocating multi-mapped reads. This step is not sufficient for achieving optimal accuracy and it cannot influence isoforms reconstruction outcome. ‡ IsoLasso and SLIDE use LASSO for regularization which is not an adequate regularizer for this problem and both use abundance thresholds to filter out isoforms.

Figure S1 demonstrates key features of iReckon, and for each, whether they are present in three other popular programs for RNA-seq analysis: Cufflinks, IsoLasso, and Slide.

2 Investigation of the Efficacy of iReckon's Features

To isolate the features of iReckon that have the largest impact on performance we have conducted tests of the various features either in isolation, or by removing them from the full iReckon tool. Several features, such as the importance of the use of multi-mapping reads were not tested as they form the core of the iReckon approach, and are not easily deconvolved from the rest of the software. Also we do not test the efficacy of the model for handling PCR duplicates, as we are not aware of a good model for simulating this type of data. In all of these tests we also compare to the results obtained with Cufflinks (the best performing of the other tools we evaluated in the main manuscript) to better illustrate the advantages of the particular feature.

2.1 The Likelihood and EM Algorithm: Abundance Estimation Accuracy

In the first experiment we isolate the quantification aspects of iReckon from the isoforms reconstruction aspect. To do so, we simulated reads from a set of two million read pairs from 2079 known isoforms, all of which are expressed in the sample. We restrict both Cufflinks and iReckon to only consider the known isoforms, and also removed all additional features from iReckon (regularization, pre-mRNA, intron retention, PCR-Duplicates, coherence score, bias correction). The experiment was repeated three times and the results shown are the average of the runs (they were consistent every time). We note that in this case our objective function becomes equivalent to that of [Li et al 2010]. However we also re-align all reads to the transcriptome, rather than only considering mappings generated by Bowtie/Tophat.

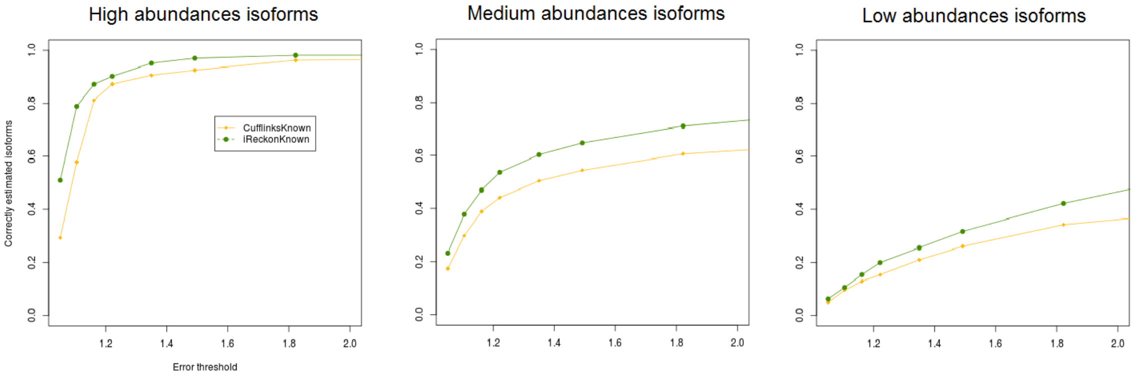


Fig. S2: Isoforms abundances estimation accuracy of both Cufflinks and iReckon when restricted to known isoforms.

The results, shown in Figure S2, demonstrate that iReckon's abundance estimation accuracy is superior to that of Cufflinks. The most likely reason for the performance advantage is that iReckon properly handles reads mapped to multiple isoforms through the EM algorithm, while Cufflinks, in effect, does only one round of EM: it assigns all of the multi-mapped reads to an isoform based on the initial abundance estimates generated by the uniquely mapped reads. Additional reasons may include better quality of mapping achieved by remapping all reads to the transcriptome; however based on the percentage of reads mapped by Cufflinks and iReckon (98.12% and 98.17%, respectively), we do not believe this to be a major influence.

2.2 Regularization in a Simple Setting

Because regularization forms a key element of iReckon we test it twice. In this section we test the performance of regularization in the simplest case, where all of the simulated isoforms are known. Regularization is more important and its effect more significant when we reconstruct novel isoforms and need to consider a large set of possibilities (see Section 3). In these experiments we start with 1545 genes with 2580 known isoforms. 30% of these are held-out, while the rest are used to simulate two millions read pairs. Both Cuffliks and iReckon are provided with all isoforms (simulated and held-out), and restricted to only using these for abundance estimation.

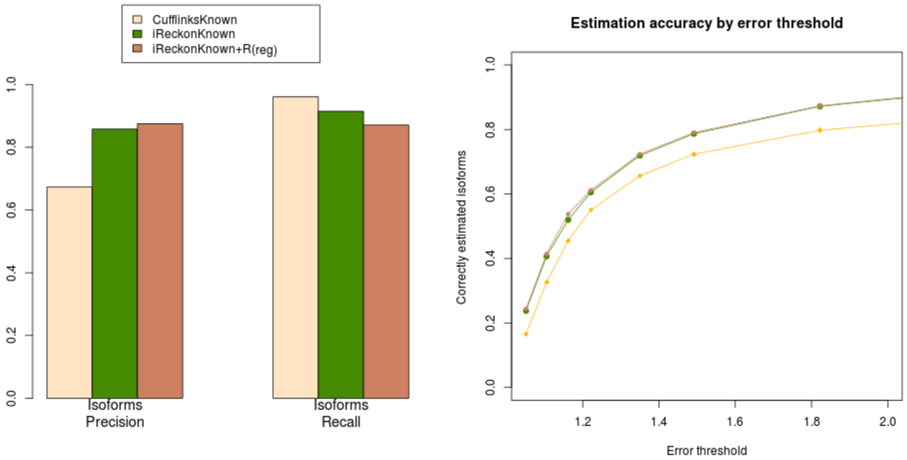


Fig. S3: Isoforms detection power and abundances estimation accuracy of both Cufflinks and iReckon restricted to known isoforms (no novel isoforms reconstruction). iReckon is used with and without regularization.

Figure S3 shows that regularization slightly improves the accuracy of abundances estimation even in the setting where the algorithm only considers known isoforms. The median deviation (the abundance estimation error for the average gene) for iReckon decreases from 18.5% to 15% when we use regularization. As expected, regularization also improves precision and decreases recall. Cufflinks, in comparison, reports a large fraction of the isoforms as being present in the sample. This is likely due to its failure to iterate the EM algorithm and identify the true source of multi-mapped reads.

2.3 Modelling the pre-mRNA

In this experiment we demonstrate the effect of modelling the presence of pre-mRNA within iReckon. Starting with our previous simulation, we also add reads simulated from the pre-mRNA of the expressed isoforms. These are simulated so that pre-mRNA makes up on 10% of the average abundance of known isoforms. Again, we limit Cufflinks and iReckon to only use known isoforms, along with another version of iReckon that models the presence of the pre-mRNA.

Figure S4 demonstrates that the presence of pre-mRNAs biases the abundances estimation, and that directly modelling pre-mRNA greatly improves results. The precison and recall of iReckon both improve by 3%, while the median deviation (the error in abundance

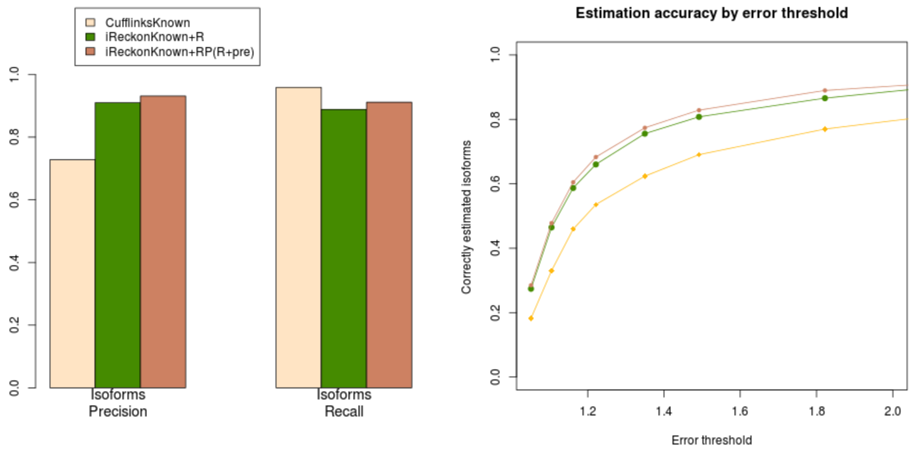


Fig. S4: Isoforms detection power and abundances estimation accuracy of both Cufflinks and iReckon restricted to known isoforms (no novel isoforms reconstruction). iReckon is used with and without direct modelling of the pre-mRNA.

estimation for the average gene) is reduced by 10%. Hence, while none of the results in the manuscript (precision, recall, accuracy) do not account for pre-mRNA directly, the direct modelling of this feature improves our overall results.

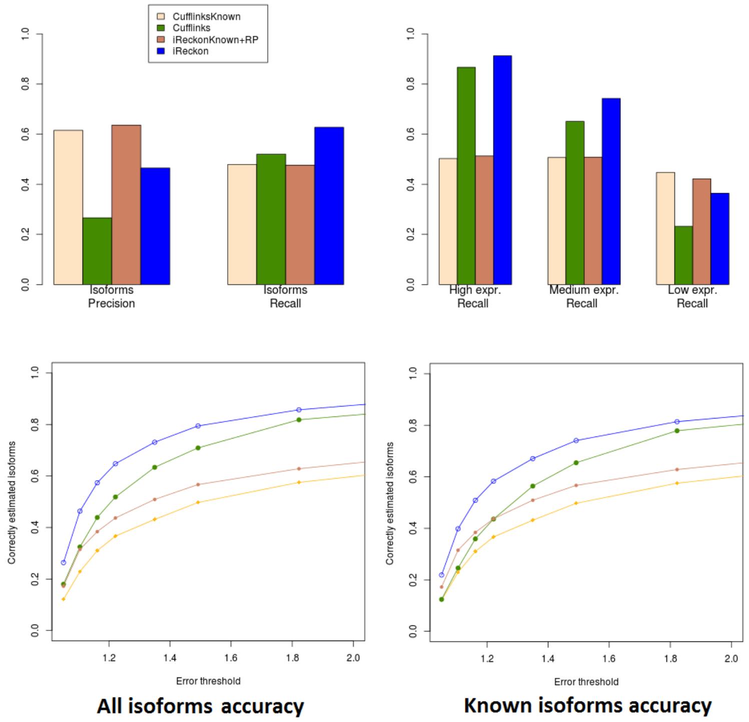
2.4 Introducing novel isoforms

In this section, we introduce the novel isoforms reconstruction problem in its simplest form. We start with 1044 genes with 1471 known isoforms. For each gene we add one (novel) alternative splicing event, and simulate two millions read pairs from both known and novel isoforms. We test both versions of Cufflinks and iReckon restricted to known isoforms, as well as the versions that can discover novel isoforms.

Figure S5 shows the importance of discovering novel isoforms: not doing so not only reduces recall, as expected, but also significantly biases the abundances estimation results, even for known isoforms (reads generated from novel isoforms are incorrectly assigned to known ones, causing overestimation). The modelling of novel isoforms also, expectedly, reduces the precision; however the impact of this is much lower for iReckon then for Cufflinks, even though iReckon considers a larger set of isoforms. This is due to both the iterative EM algorithm (as discussed above) and effective use of regularization.

2.5 Complex isoforms reconstruction

We now increase the complexity of the isoforms reconstruction problem by allowing multiple alternative splicing events per isoform simulated. We use an initial dataset of 1144 isoforms with two millions read pairs simulated. The simulation protocol for isoforms is the one described in the main manuscript. In addition to comparing how well Cufflinks and iReckon perform on this data, we also use the data to evaluate the efficacy of the coherence score used in iReckon (see section 4 of this supplement).



All isoforms accuracy **Known isoforms accuracy**

Fig. S5: Isoforms detection power and abundances estimation accuracy of Cufflinks and iReckon when: restricted to known isoforms (No novel isoforms reconstruction) and allowed to discover novel isoforms

The performance gap between Cufflinks and iReckon in Figure S5 and Figure S6 shows that iReckon can adapt better than Cufflinks to cases with multiple alternative events. Notably, the sensitivity to novel low abundances isoforms of iReckon is four times that of Cufflinks. Simultaneously the coherence score does not appear to play a major role in iReckon's performance, resulting in slightly decreased precision, slightly better recall, and slightly more accurate abundance estimates.

2.6 Impact of low abundances isoforms

To identify the impact of low-abundance isoforms on the results, we conduct the experiment described in the previous section, while excluding low abundance isoforms: our simulation contains only high and medium abundances isoforms. Figure S7 shows that iReckon performance is better both for isoforms detection and quantification even when there are no low abundances isoforms in the data, albeit the difference is smaller than in the full simulation.

2.7 Varying Read Coverage

The total amount of information available from RNA-seq data increases with the number of reads available. In this section, we use an RNA-seq simulation protocol as presented in the main paper, however we vary the total number of reads sampled from each isoform. The

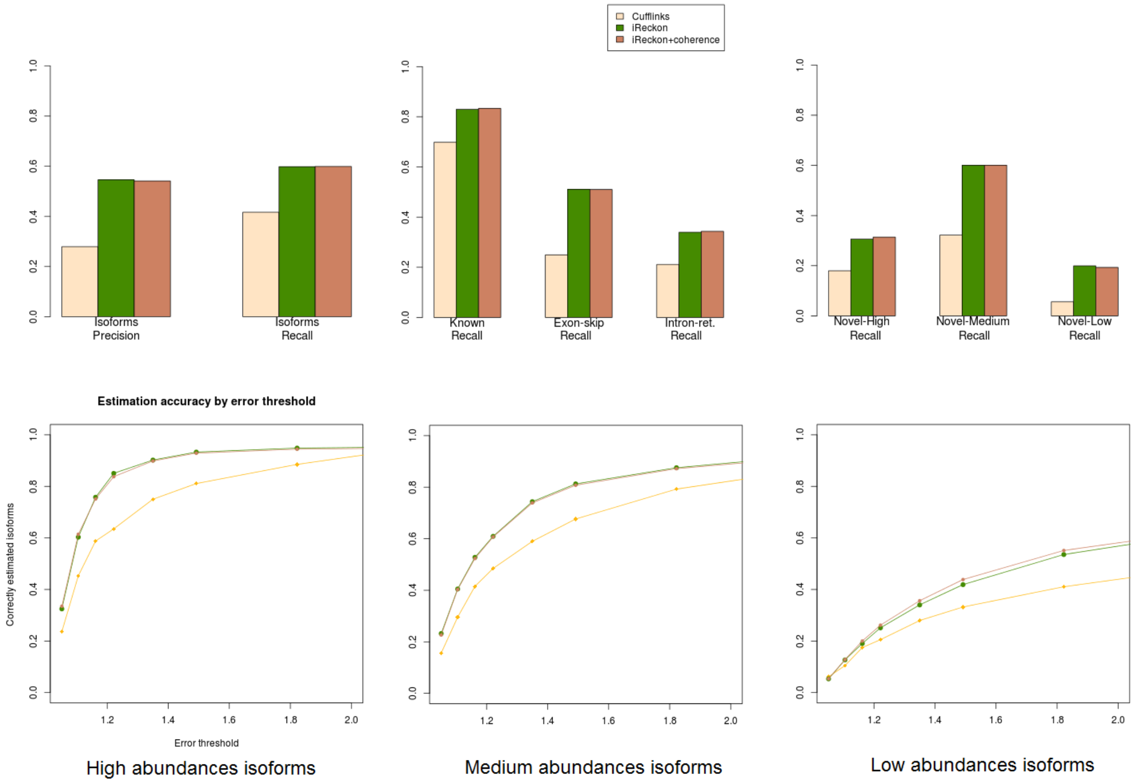


Fig. S6: Isoforms detection power and abundances estimation accuracy of Cufflinks and iReckon. iReckon is used with and without the coherence score.

goal of this experiment is to look at how much the performance of each method increases with increased coverage. While the performance of both iReckon and Cufflinks improves with increased coverage, Figure S8 shows that iReckon is better able to take advantage of extra data. This is especially true for low abundance isoforms, while the performance for high-abundance isoforms does not change as much.

3 Regularization choice

	Recall			**Precision**			**F-measure**		
Nb possible isoforms	$< \mathbf{10}$	$[\mathbf{10}, \mathbf{20}]$	$> \mathbf{20}$	$< \mathbf{10}$	$[\mathbf{10}, \mathbf{20}]$	$> \mathbf{20}$	$< \mathbf{10}$	$[\mathbf{10}, \mathbf{20}]$	$> \mathbf{20}$
Reguralization	0.42	0.25	0.12	0.33	0.22	0.11	0.37	0.23	0.11
No reguralization	0.47	0.32	0.21	0.25	0.18	0.10	0.32	0.23	0.14
LASSO	0.46	0.32	0.21	0.25	0.17	0.10	0.32	0.23	0.14

Table S1: Effect of using iReckon's regularization, no regularization, and the LASSO regularizer. The three grouping based on the number of isoforms (< 10, $[10, 20]$, > 20) correspond to 76.6%, 10.7% and 12.7% of the genes, respectively.

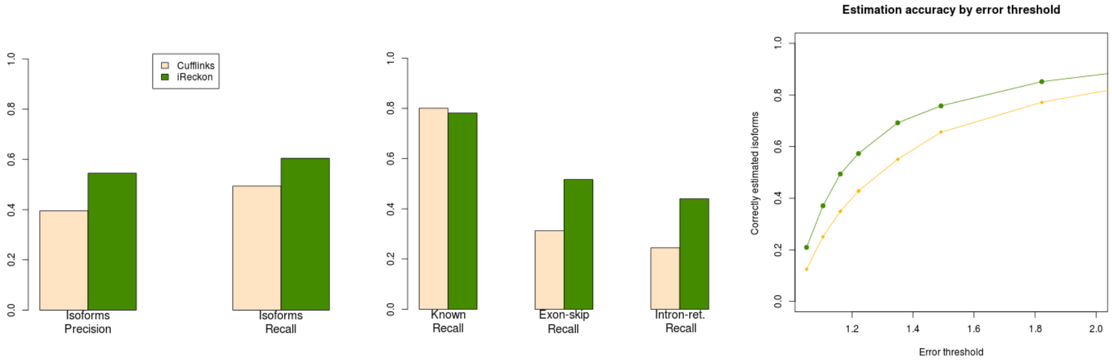


Fig. S7: Isoforms detection power and abundances estimation accuracy of Cufflinks and iReckon in the absence of low abundance isoforms.

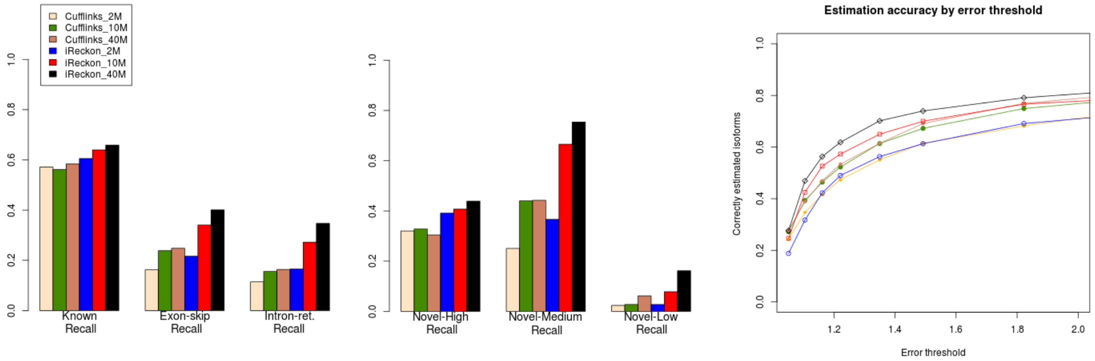


Fig. S8: Isoforms detection power and abundances estimation accuracy of Cufflinks and iReckon when we use two million, ten million or forty million RNA-seq read pairs

To evaluate the effects of regularizing our EM algorithm, as well as the choice of the regularizer in this section we try iReckon with alternate regularization functions using the simulated data described in the main manuscript. In addition to our regularizer (based on the sum of the fourth root of the abundances) we also use standard LASSO (penalizing the sum of the abundances) and no regularization at all. In the results, summarized in Table S1, we separately consider the genes based on the number of isoforms. Regularization reduces the number of isoforms used to explain the data, leading to higher precision and lower recall. The F-measure (harmonic mean of precision and recall) is a standard way of combining these two parameters. Based on the F-measure, the results when using regularization improve when the number of isoforms is less than 20 (87.3% of the genes), while it is less useful when the problem is too complex ($>$ 20 possible isoforms). In such cases it is likely that the model with the smallest number of isoforms may not be the most appropriate.

The choice of the regularization term needs to take into account the external constraints on the parameters being regularized. In the case of abundance estimation, because abundances are similar to frequencies, they have a positivity constraint and a fixed sum. In such cases LASSO is inappropriate as it penalizes the sum of the frequencies, which is fixed and cannot be reduced. This is confirmed by our results in Table S1 where using LASSO produces almost identical results to not using regularization at all.

4 Coherence score computation

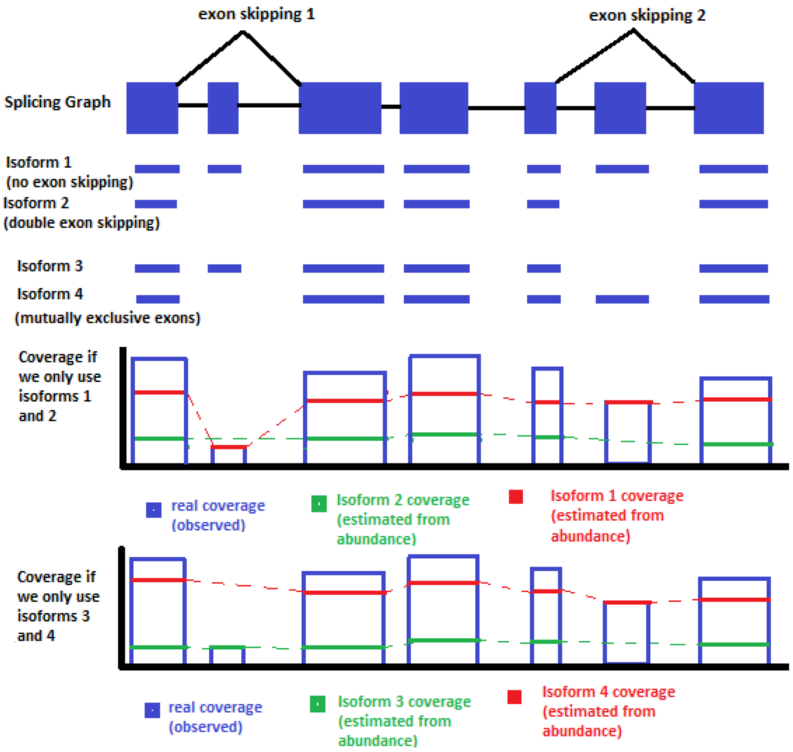

exon skipping 1
exon skipping 2
Splicing Graph
Isoform 1
(no exon skipping)
Isoform 2
(double exon skipping)
Isoform 3
Isoform 4
(mutually exclusive exons)
Coverage if we only use isoforms 1 and 2
real coverage (observed)
Isoform 2 coverage (estimated from abundance)
Isoform 1 coverage (estimated from abundance)
Coverage if we only use isoforms 3 and 4
real coverage (observed)
Isoform 3 coverage (estimated from abundance)
Isoform 4 coverage (estimated from abundance)

Fig. S9: Illustration of coverage coherence for isoform selection. The coverage signal (boxes) can be explained by using either isoforms 1 and 2, or isoforms 3 and 4. While from the perspective of the standard likelihood the two solutions are equally good, the first solution has a non-coherent exon 2 with a much lower abundance in isoform 1 than the other exons. In contrast, the solution using isoforms 3 and 4 has uniform coverage across all exons, and hence is more coherent.

The regularized EM algorithm can converge to many solutions corresponding to local maxima that sometimes have very close log-likelihood. Read pairs (and more specifically pairs of junctions) are helpful in disambiguating some, but not all of the cases. We introduce a novel technique for disambiguation of isoforms based on coherence of the coverage signal. By maximizing the likelihood of all reads individually we do not take into account their distribution across the isoform, while in practice we expect the expression level within the various sub-blocks to be coherent: for example all reads should not be generated from a single exon of a multi-exon isoform. An example of coherence being used to disambiguate isoforms is demonstrated in Figure S9.

To compute the coherence score, we estimate for every isoform the expected coverage over each exon it contains to obtain an isoform-coverage signal (dependent on the estimated abundances). The coherence score is inversely proportional to the sum of the deviations in the signal: the more stable the isoform-coverage is across the mRNA, the higher the coherence score. Because we only use the coherence score to choose from multiple near-optimal solutions it is only considered after we converge to a local optimum using the two first terms of the objective function in Equation 5 of the main manuscript.

5 Constraints implementation

For a studied gene with n isoforms, the abundances $(\theta_1, \ldots, \theta_n)$ obey the following constraints:

$$\sum_{i=1}^{n} \theta_i l_i = C \text{ and } \theta_i \geq 0 \ \forall i \text{ in 1..n} .$$

Where $(l_1, \ldots, l_n)$ are the lengths of the isoforms and C is a constant. This constraints comes from the definition of the RPKM measure for isoforms abundances (Reads Per Kilobase per Million mapped reads):

$$\theta_i = Nb\_Reads_i \cdot \frac{10^6}{Total\_Nb\_Reads} \cdot \frac{10^3}{l_i}$$

Where $Total\_Nb\_Reads$ is the total number of reads that align to the transcriptome and Nb_Reads_i is the number of reads assigned to the i^{th} isoform. When we multiply this definition by the l_i variables and sum over all isoforms in a gene, we have:

$$\begin{aligned}\sum_i \theta_i l_i &= \sum_i \frac{Nb\_Reads_i}{Total\_Nb\_Reads} \cdot 10^9 \\ &= \frac{Gene\_Nb\_Reads}{Total\_Nb\_Reads} \cdot 10^9 \\ &= C\end{aligned}$$

Where $Gene\_Nb\_Reads$ is the number of reads originated from the considered gene. Finally, if we were to sum over all isoforms in the transcriptome then we would have:

$$\sum_i \theta_i l_i = 10^9$$

To implement these constraints during likelihood optimization, we associate the abundances $(\theta_1, \ldots, \theta_n)$ to the variables $(x_1, \ldots, x_n)$ where:

$$\theta_i = \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_k}} \text{ , } \forall i \text{ in 1..n} .$$

And then we optimize over $(x_1, \cdots, x_n)$ without having to worry over constraints. This change of variables is very practical as it is easily derived and results in many simplifications in the calculus. For example if we look at the first term of the objective function corresponding to the log-likelihood, we derive the θ dependent part of it in the M step of the EM algorithm as follows:

$$\begin{aligned}\frac{d(\sum_{n,i}\mathbb{E}[Z_{ni}]\cdot\log(\theta_i))}{dx_k} &= \sum_{n,i}\mathbb{E}[Z_{ni}]\cdot\frac{d\theta_i}{\theta_i dx_k}\\ &= \sum_{n}\mathbb{E}[Z_{nk}] - \theta_k\sum_{n,i}\mathbb{E}[Z_{ni}]\\ &= \sum_{n}\mathbb{E}[Z_{nk}] - \theta_k\cdot n\end{aligned}$$

since we have:

$$\frac{d\theta_i}{dx_k} = \begin{cases} -\theta_i\theta_k & \text{if } i \neq k \\ \theta_i(1-\theta_i) & \text{if } i = k \end{cases}$$

6 Derivation of Expected Number of Natural Duplicates

As described in the main manuscript, the number of occurrences X_f of a particular fragment f is modelled by a binomial distribution $B(w, p_f)$ which can be approximated by the $Poisson(p_f \cdot w)$ distribution since w is usually large (> 20) and p_f is very small (< 0.01). The number of duplicates is then the random variable $Y = max\{0, X_f - 1\}$. (We keep one original read and $X_f - 1$ duplicates)

To compute the mean and variance of Y:

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{k=1}^{\infty} kP(Y = k)\\ &= \sum_{k=1}^{\infty} kP(X_f = k+1)\\ &= \sum_{k=1}^{\infty} k\frac{(p_f\cdot w)^{k+1}}{(k+1)!}e^{-p_f\cdot w}\end{aligned}$$

This sum can be treated as a series of functions $S(x)$ with $x = p_f \cdot w$ and we verify that we have the differential equation:

$$S'(x) = -S(x) + x, S(0) = 0$$

By solving this equation (integral calculus) we find:

$$\mathbb{E}[Y] = p_f\cdot w + e^{-p_f\cdot w} - 1 \tag{1}$$

A similar approach gives the variance:

$$var(Y) = pr + e^{-p_f \cdot w} - e^{-2p_f \cdot w} - 2p_f \cdot w \cdot e^{-p_f \cdot w} \tag{2}$$

7 Abundances estimation accuracy

Here we present additional comparative abundance accuracy results. In all our accuracy estimations and comparisons, we consider only the correctly reconstructed isoforms. As some of the methods are less precise on the exact start and end of isoforms, we allow a margin of error for the beginning and end of the isoforms. Internal exons borders, however, are required to be perfectly reconstructed.

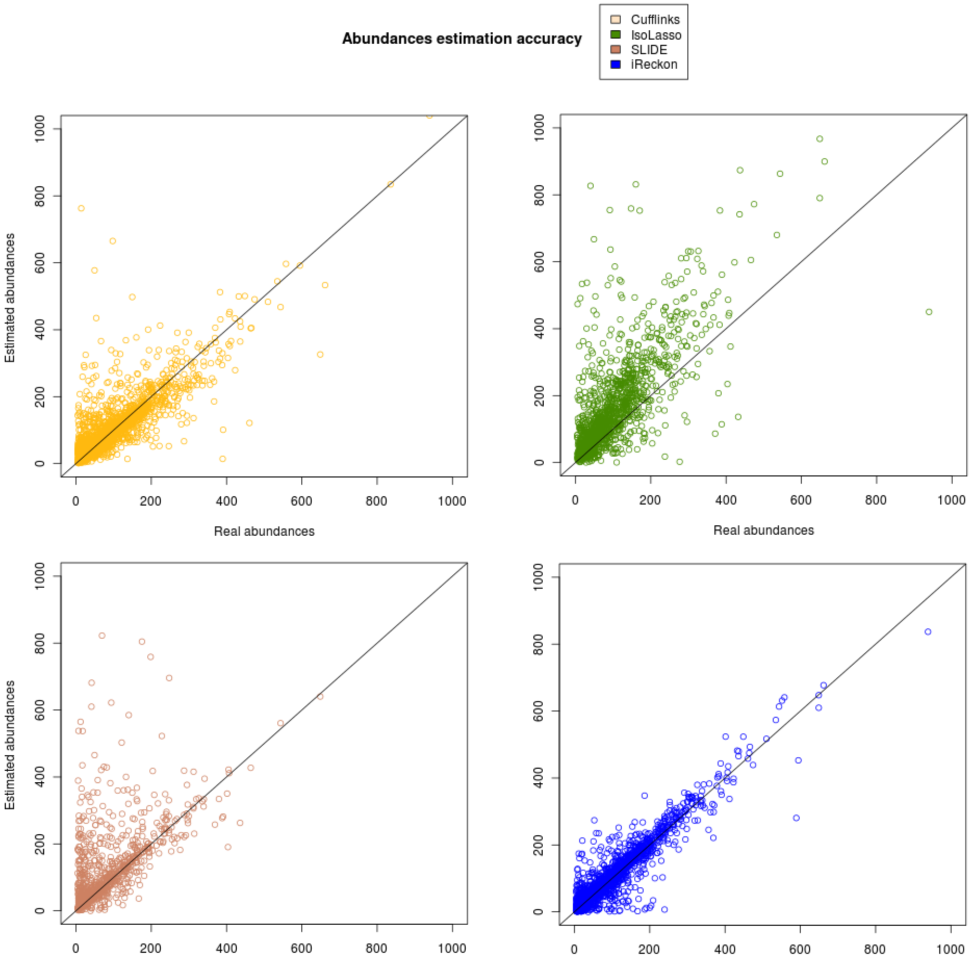


Fig. S10: Correlation between estimated abundances values and real values for the different methods. The line in black is the identity line. Only the strictly positive abundances estimations are displayed.

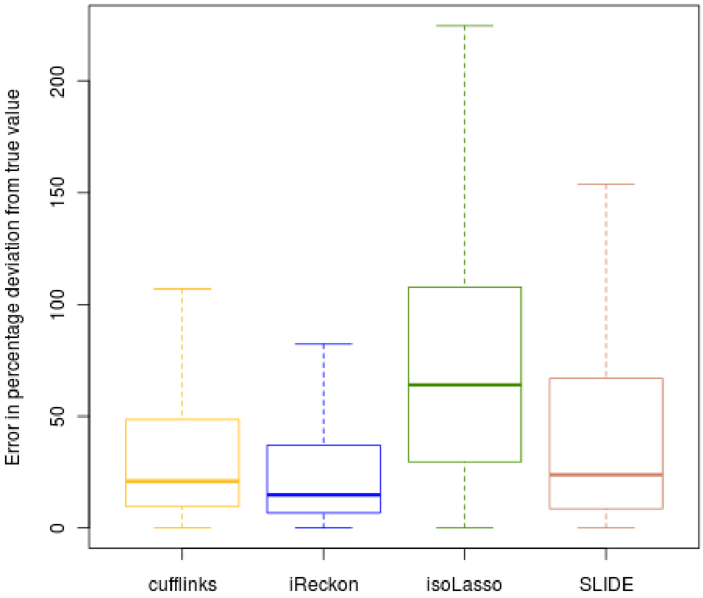


Fig. S11: Box and whiskers plot illustrating the deviation percentage from the true abundances values of the different methods. The thick line indicates the median error, the box indicates the 25-75 percentile range, and the whiskers show the full range of values. iReckon is significantly more accurate than the other methods (p-value = $8.06E^{-18}$ using the wilcoxon test)