Analysis and design of RNA sequencing experiments for identifying isoform regulation

Yarden Katz[1,2], Eric T Wang[2,3], Edoardo M Airoldi[4] & Christopher B Burge[2,5]

Through alternative splicing, most human genes express multiple isoforms that often differ in function. To infer isoform regulation from high-throughput sequencing of cDNA fragments (RNA-seq), we developed the mixture-of-isoforms (MISO) model, a statistical model that estimates expression of alternatively spliced exons and isoforms and assesses confidence in these estimates. Incorporation of mRNA fragment length distribution in paired-end RNA-seq greatly improved estimation of alternative-splicing levels. MISO also detects differentially regulated exons or isoforms. Application of MISO implicated the RNA splicing factor hnRNP H1 in the regulation of alternative cleavage and polyadenylation, a role that was supported by UV cross-linking–immunoprecipitation sequencing (CLIP-seq) analysis in human cells. Our results provide a probabilistic framework for RNA-seq analysis, give functional insights into pre-mRNA processing and yield guidelines for the optimal design of RNA-seq experiments for studies of gene and isoform expression.

The distinct isoforms expressed from metazoan genes through alternative splicing can be important in development, differentiation and disease[1]. For example, the pyruvate kinase gene produces two distinct tissue-specific spliced isoforms that differ in their enzymatic activity, allosteric regulation and ability to support tumor growth[2]. Conservative estimates predict 2–12 mRNA isoforms for most mammalian genes (**Supplementary Fig. 1**), though some genes, including neurexins, may express more than 1,000 isoforms each[3].

Recently, high-throughput sequencing of short cDNA fragments, RNA-seq, has emerged as a powerful approach to characterizing the transcriptome. RNA-seq data have recently been used to show that the vast majority of human genes are alternatively spliced and that most alternative exons show tissue-specific regulation[4]. To date, RNA-seq analysis methods have focused mostly on estimation of gene expression levels and discovery of novel exons and genes[4–6], assembly and annotation of mRNA transcripts[5,7], and estimation of the expression levels of alternative exons[4]. Two recent methods, Cufflinks and Scripture, can produce *de novo* annotations of transcripts in metazoan genomes using RNA-seq data alone[8–10].

Accurate quantification of alternative-exon abundance and detection of differentially regulated exons and isoforms remain challenging. Paired-end RNA-seq protocols, in which both ends of a cDNA fragment are sequenced, are paving the way for isoform-centric rather than exon-centric analyses. Here we have developed the MISO model, a probabilistic framework that uses information in single-end or paired-end RNA-seq data to enable more comprehensive and accurate analysis of alternative splicing, at either the exon or isoform level. MISO provides confidence intervals (CIs) for estimates of exon and isoform abundance, detects differential expression and uses latent information to improve accuracy. We applied MISO to analyze isoform regulation by the splicing factor hnRNP H. Using MISO, we showed how the mean and variance of the library insert length affects the information obtained about splicing events in paired-end RNA-seq data, yielding guidelines for the design of RNA-seq experiments.

RESULTS

Quantifying alternative splicing with MISO

To detect alternative splicing using RNA-seq data, MISO and other methods use sequence reads aligned to splice-junction sequences that are either precomputed from known or predicted exon-intron boundaries, or discovered *de novo* by spliced alignment to the genome (Online Methods). In the most common type of alternative splicing in mammals, an exon is included or excluded from the mature mRNA; 'percentage spliced in' (PSI or Ψ)[11] denotes the fraction of mRNAs that represent the inclusion isoform. Reads aligning to the alternative exon or to its junctions with adjacent constitutive exons provide support for the inclusion isoform, whereas reads aligning to the junction between the adjacent constitutive exons support the exclusion isoform; the relative read density of these two sets forms the standard estimate of Ψ, denoted $\hat{\Psi}_{SJ}$ (**Fig. 1** and **Supplementary Fig. 2**)[4].

This estimate ignores reads that align to the bodies of the flanking constitutive exons, which could have derived from either isoform. Nevertheless, these constitutive reads contain latent

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA. [2]Department of Biology, MIT, Cambridge, Massachusetts, USA. [3]Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts, USA. [4]Department of Statistics and FAS Center for Systems Biology, Harvard University, Cambridge, Massachusetts, USA. [5]Department of Biological Engineering, MIT, Cambridge, Massachusetts, USA. Correspondence should be addressed to E.M.A. (airoldi@fas.harvard.edu) or C.B.B. (cburge@mit.edu).

Figure 1 | More accurate inference of splicing levels using MISO. (**a**) Generative process for MISO model. White, alternatively spliced exon; gray and black, flanking constitutive exons. RNA-seq reads aligning to the alternative exon body (white) or to splice junctions involving this exon support the inclusive isoform, whereas reads joining the two constitutive exons (black-gray exon junction) support the exclusive isoform. Reads aligning to the constitutive exons are common to both isoforms. (**b**) The $\hat{\Psi}_{SJ}$ estimate uses splice-junction and alternative exon–body reads only. (**c**) The MISO estimate, $\hat{\Psi}_{MISO}$ (derived here analytically), also uses constitutive reads and paired-end read information; orange lines connect reads in a pair; the insert length distribution is shown at right. (**d**) Comparison of $\hat{\Psi}_{SJ}$ and $\hat{\Psi}_{MISO}$ estimates from simulated data. Reads were sampled at varying coverage, measured in RPK, from the gene structure shown at top right, with underlying true $\Psi = 0.5$. Mean values from 3,000 simulations are shown ($\pm$s.d.) for each coverage value. Percentiles of gene expression values are shown for a data set assuming 25 million mapped paired-end (PE) read pairs (25M PE; blue, extrapolating from an Illumina GA2 run that yielded 15 million mapped read pairs) and for a data set of 78 million mapped read pairs from an Illumina HiSeq 2000 instrument (78M PE; red), both obtained from human heart tissue.
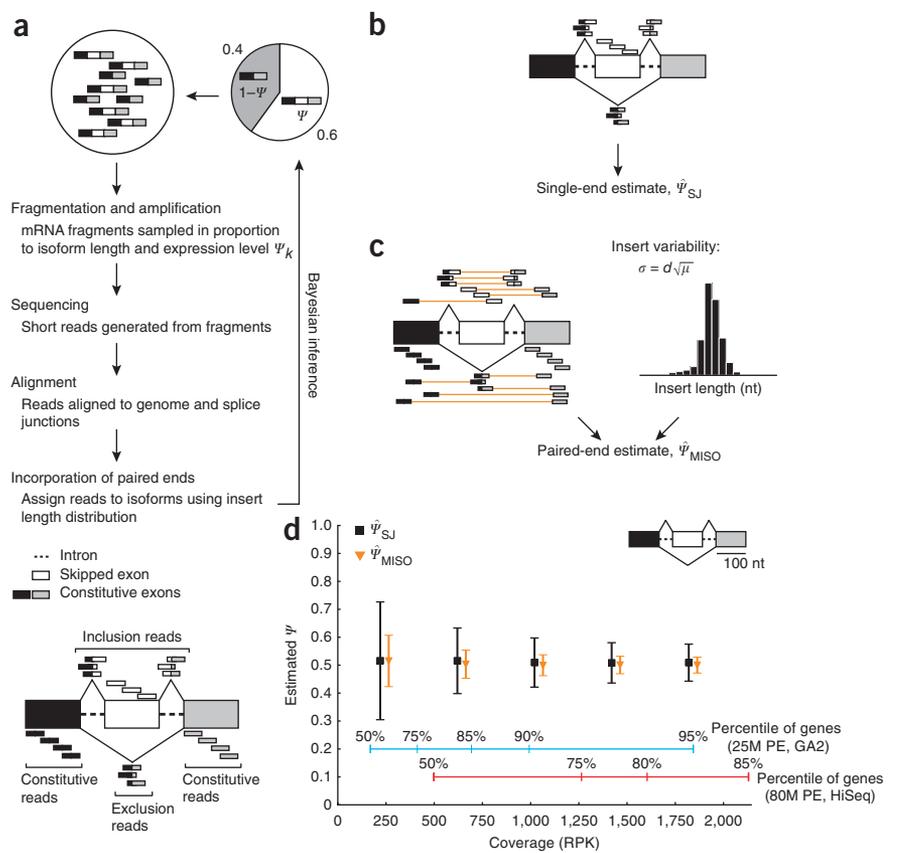
information about the splicing of the alternative exon, as higher expression of the exclusion isoform will generally increase the density of reads in the flanking exons relative to the alternative exon, and lower expression of the exclusion isoform will decrease this ratio of densities. MISO captures this, as well as the information in the lengths of library inserts in paired-end data, by recasting the analysis of isoforms as a Bayesian inference problem. Our approach is related to the alternative-splicing quantification method[12], which does not use paired-end information.

MISO samples reads uniformly from the chosen isoform, then recovers the underlying abundances of isoforms (Ψ and $1 - \Psi$ in the case of a single alternative exon) using the short read data (**Fig. 1a** and **Supplementary Fig. 3**). As a result of mRNA fragmentation in library preparation, mRNA abundance and length contribute roughly linearly to read sampling in RNA-seq. This effect is treated by rescaling the abundances Ψ and $1 - \Psi$ of the two isoforms by the number of possible reads that could be generated from each isoform, respectively. In the model, reads from a gene locus are produced by a generative process in which an isoform is first chosen according to its rescaled abundance, and a sequence read is then sampled uniformly from possible read positions along the mRNA (Online Methods). For the exon-centric analyses involving a single alternative exon we derived an analytic solution to the inference problem, whereas for isoform-centric analyses and estimation using CIs we developed an efficient inference technique based on Monte Carlo sampling (Online Methods). Our new estimator, $\hat{\Psi}_{MISO}$, uses all of the read positions used in $\hat{\Psi}_{SJ}$, plus reads aligning to the adjacent exons (**Fig. 1b,c**) and information about the library insert length distribution in paired-end RNA-seq. Both $\hat{\Psi}_{SJ}$ and $\hat{\Psi}_{MISO}$ are unbiased estimators of Ψ.

An improved measure of exon expression

Simulating read generation from an alternatively spliced gene, we observed that the $\hat{\Psi}_{MISO}$ estimate had consistently much lower variance and error than $\hat{\Psi}_{SJ}$ (**Fig. 1d**). For reference, the distribution of read-coverage values at depths typically obtained from one lane of sequencing on an Illumina Genome Analyzer 2 (GA2) and on a HiSeq 2000 are shown, in units of reads per kilobase of exon model (RPK). For a gene with median coverage in the GA2 data set (~220 RPK), the s.d. of the estimated Ψ value was reduced more than twofold, from 0.21 for $\hat{\Psi}_{SJ}$ to 0.09 for $\hat{\Psi}_{MISO}$.

Validation of MISO estimates

To assess the uncertainty in the splicing estimates for each exon, we calculated CIs for Ψ (Online Methods) from moderate-depth breast cancer RNA-seq data (**Supplementary Table 1**; examples are shown in **Fig. 2a,b**). Comparing $\hat{\Psi}_{MISO}$ estimates for 52 alternative exons to corresponding quantitative reverse-transcription PCR (qRT-PCR) values[11,13] yielded a Pearson correlation $r = 0.87$ (**Fig. 2c** and **Supplementary Table 2**; a bias in the RT-PCR data was analyzed in **Supplementary Figs. 4–6**). Restricting the analysis to exons with 95% CI width <0.25 increased the correlation with qRT-PCR data considerably, to $r = 0.96$ (**Fig. 2d**). Thus, MISO CIs identify exons whose RNA-seq–based Ψ-value estimates are more reliable.

Detection of differentially expressed isoforms

Differential splicing of alternative exons entails a difference in Ψ values, $\Delta\Psi$, and can be evaluated statistically using the Bayes factor (BF), which quantifies the odds of differential regulation

Figure 2 | MISO CIs for Ψ values and qRT-PCR validation. qRT-PCR measurements from ref. 13 for a set of 52 alternatively skipped exons were compared to MISO posterior mean estimates of Ψ, denoted $\hat{\Psi}_{MISO}$. Full listing of events is given in **Supplementary Table 1**. (**a,b**) The Ψ posterior distributions obtained by sampling and 95% CIs are shown for two representative exons, one with a wide (*NFYA*, exon 3) and one with a narrower (*ZNF207*, exon 6) CI. qRT-PCR Ψ measurements are indicated in red. (**c**) Scatterplot of MISO and qRT-PCR Ψ estimates for the full set of 52 events. (**d**) Scatterplot of MISO and qRT-PCR estimates for the subset of 23 high-confidence events, for which CI width <0.25. One exon was excluded from this plot because of expressed sequence tag (EST) evidence of an alternative isoform expected to confound the qRT-PCR analysis (**Supplementary Fig. 6**).

occurring. MISO is used to calculate the posterior probability distributions of Ψ and $\Delta\Psi$ for the two samples. The latter distribution is used to calculate the BF, defined as the ratio of the posterior probability of the alternative hypothesis, $\Delta\Psi \neq 0$, to that of the null hypothesis, $\Delta\Psi = 0$ (Online Methods); thus, higher values of the BF indicate increased confidence in differential regulation.

In a recent study we used RNA-seq to characterize transcriptome changes after RNA-interference knockdown of the splicing factor hnRNP H in cultured human cells[14]. This factor is known to bind polyguanine (poly(G)) runs, typically activating splicing when binding in introns flanking an exon and repressing splicing when binding in exons (**Fig. 3a,b**). An example of BF calculation for a gene with moderately high read coverage is shown in **Figure 3c**. When we compared RNA-seq to qRT-PCR data, we found that 100% of exons (6 of 6) with BF ≥ 20 were detected as differentially
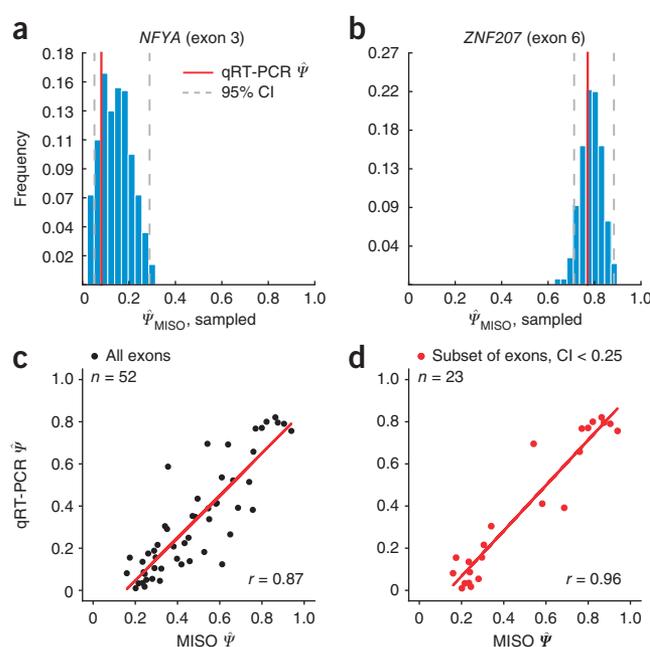
regulated by qRT-PCR, compared to 21% of exons (4 of 19) with BF < 20 ($P < 0.0004$, Fisher's exact test), and the magnitude of $\Delta\Psi$ showed good agreement (**Supplementary Fig. 7**). Overall, 15% of alternative exons changed with BF ≥ 20 (**Fig. 3d**); similarly widespread changes in splicing have been observed by all-exon microarray analysis[14].

Genome-wide validation of isoform regulation by CLIP-seq
To identify events directly regulated by hnRNP H and further validate the BF analysis, we performed CLIP-seq analysis of hnRNP H1 under the same conditions as in ref. 14 to identify RNA binding sites of hnRNP H transcriptome-wide. Notably, the percentage of exons with CLIP tags in their flanking introns whose splicing was enhanced by hnRNP H ($\Delta\Psi > 0$ between control and knockdown conditions) increased from 60% to over 90% as the BF threshold was increased, approaching a plateau at a BF = 5 (**Fig. 3e**), corresponding to 5:1 odds of regulation. This effect was stronger for hnRNP H binding in the downstream intron
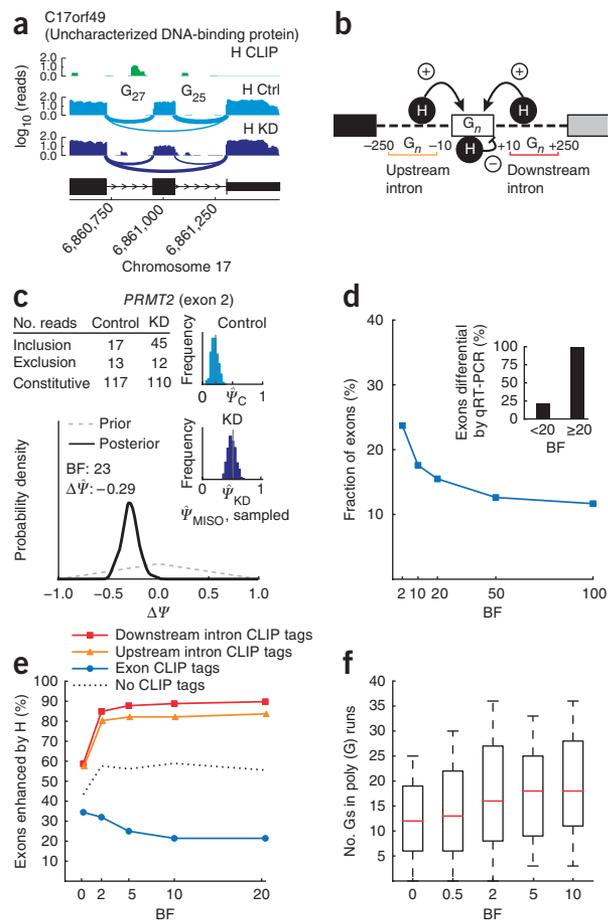

Figure 3 | Bayes factor analysis of hnRNP H regulation of exon splicing. (**a**) CLIP tag density (H CLIP; green) and RNA-seq read densities in hnRNP H–knockdown and control conditions (H KD and H Ctrl; light and dark blue, respectively) for an alternative exon in human C17orf49. Number of guanines in poly(G) runs in upstream and downstream introns is shown. (**b**) Model of hnRNP H function in splicing regulation: binding of poly(G) runs (G_n) adjacent to an exon enhances the exon's splicing (+ arrows); binding in exon body represses splicing (− arrow). A 250-nt window in flanking introns was used to count CLIP tags in analyses. (**c**) BF for exon 2 of *PRMT2* gene. Gray dashed line, distribution over $\Delta\Psi$ under the null hypothesis; black solid line, posterior distribution. (**d**) Cumulative distribution of BFs using hnRNP H RNA-seq data for exons with sufficiently high read coverage. Inset, fraction of differentially regulated exons ($\Delta\Psi \geq 0.15$ by qRT-PCR), grouping exons by BF ($n = 25$ exons). (**e**) Percentage of exons enhanced by hnRNP H ($\Delta\Psi > 0$), plotted against increasing BF thresholds, for exons with CLIP tags in downstream or upstream introns but not in exon body (red and orange curves), for exons with CLIP tags in exon body but not in flanking introns (blue curve) and for exons with no CLIP tags (dotted black line). (**f**) Guanines in poly(G) runs in downstream intron, plotted against increasing BFs.

Figure 4 | Bayes factor analysis implicates hnRNP H in alternative cleavage and polyadenylation. (**a**) CLIP tag density (H CLIP; green) and RNA-seq read densities in hnRNP H control and knockdown conditions (H Ctrl and H KD; light and dark blue, respectively) along the 3′ UTR of the *NFATC4* gene. Core and extension poly(A) sites for *NFATC4* are shown, with a model illustrating the effect of hnRNP H effect on poly(A) site selection. (**b**) Number of CLIP tags per kilobase normalized by expression (RPKM) for exons with shortened and lengthened UTRs between hnRNP H control and knockdown conditions (red and blue curves, respectively). Values plotted are averages of subsampled mean densities (*n* = 100 subsamplings) where exons were matched for expression (RPKM). Error bars show s.e.m. CLIP tag density for UTRs not differentially regulated (BF < 1), as shown by dotted gray line.
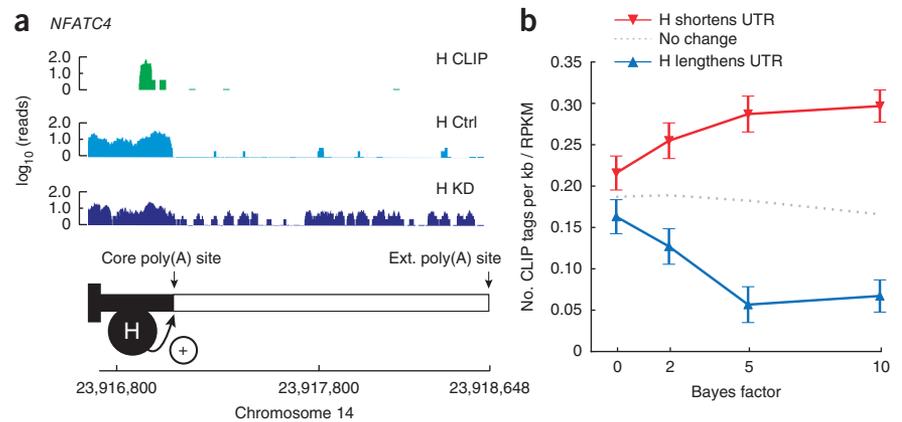
and was reversed for events with exonic CLIP tags, consistent with previous studies (for example, ref. 14 and references therein); virtually no bias was detected, on average, for exons not associated with CLIP tags. Further evidence that BF values reflect regulated exons came from the observation that exons with larger BFs had more guanines in poly(G) runs in their downstream introns (**Fig. 3f**).

A possible role for hnRNP H in alternative polyadenylation
We used a similar approach to examine whether hnRNP H also has a role in regulating tandem alternative cleavage and polyadenylation (APA), in which cleavage at distinct polyadenylation sites (PASs), without intervening splicing, results in mRNAs with longer or shorter 3′ untranslated regions (UTRs), often affecting mRNA stability, localization or translation[15]. Evidence that hnRNP H1 and its paralogs hnRNPs F and H2 affect the efficiency of constitutive cleavage and polyadenylation has been described[16,17], but regulation of alternative 3′ UTR events by this factor has not previously been reported. Notably, we observed that increased density of CLIP tags just upstream of the core (5′) PAS correlated with greater use of this site in control conditions than in the hnRNP H knockdown, suggesting a role for hnRNP H in promoting core PAS use.

For example, a high density of hnRNP H CLIP tags was observed upstream of the core PAS of the *NFATC4* gene, and RNA-seq data indicated greater use of this site in control conditions than in knockdown conditions (**Fig. 4a**). Because MISO encodes isoforms in a general way as lists of exon coordinates, APA events can be analyzed similarly to alternative splicing events (Online Methods). Applying MISO to RNA-seq data from control and hnRNP H knockdown cells, we observed that genes with higher expression of the shorter 3′ UTR isoform in the presence of hnRNP H—particularly those with large BF values—had higher CLIP tag density near the core PAS (**Fig. 4b**). Together, these analyses implicate hnRNP H1 in widespread regulation of APA in human genes by activation of the core PAS when bound nearby. Elevated levels of hnRNP H1 have been observed in certain cancers[18], and it would be of interest to determine whether hnRNP H1 contributes to the widespread '3′ UTR shortening' (preferential expression of upstream PASs) that occurs in cancer cells[19,20].

**RNA-seq design: paired-end reads and insert length**
A size-selection step is used in RNA-seq library preparation to control the mean length of inserted cDNA fragments. In paired-end sequencing, the full distribution of the lengths of these inserts can be measured precisely from read pairs that map to large constitutive regions such as 3′ UTRs, which are typically intronless. This length distribution can then be used to make qualitatively new types of inferences about alternative isoforms. For example, when the reads in a pair map upstream and downstream of an alternatively spliced exon, the inclusion and exclusion isoforms will typically imply different intervening insert lengths, often enabling the isoform from which the read was generated to be inferred with high confidence.

These considerations led us to compare the fraction of reads that are 'assignable'—that is, consistent with only one of the two isoforms—in simulations of paired-end and single-end sequencing, varying the mean, μ, of the insert length distribution (**Fig. 5**). To assess the amount of splicing information present in the length distribution, we considered read pairs that were 20 times more likely to have derived from one isoform than the other under the insert length distribution to be 'probabilistically assignable', with a 'false read assignment' (FRA) frequency of 1/20 = 5%. In **Figure 5d**, the insert length distribution has a mean ~260 ± 10 nucleotides (nt), making it far more likely that the read pair shown derived from the inclusion isoform.

Variability in the insert length distribution influences the confidence with which read pairs can be assigned to isoforms. Varying the s.d., σ, of the insert length distribution by a dispersion factor, d (where $\sigma = d\sqrt{\mu}$), we observed that even for a relatively broad insert length distribution ($d = 2$), inclusion of the 5% FRA reads substantially increased the fraction of assignable reads for a gene containing a (typically sized) 100-nt alternative exon (**Fig. 5b**). For tighter length distributions ($d = 1$ or $d = 0.5$), the fraction of assignable reads increased markedly, from ~15% when ignoring insert length information to >50% when considering insert length for large mean lengths, indicating that paired-end data with low-dispersion length distributions can potentially increase the yield of information about splicing by threefold or more at a given sequencing depth. Obtaining a length distribution with d near 1 requires care in library preparation but is achievable in practice (the libraries used in this study had d values between 0.6 and 1.5).

Figure 5 | Improved estimation of isoform abundance using paired-end reads. (**a**) Representative gene model with 100-nt first exon, 100-nt skipped exon (exon 5, in white), 150-nt constitutive exons and 600-nt last exon. (**b**) We simulated reads from the two-isoform gene model shown in **a** while varying the mean, μ, of the insert length distribution, setting the s.d. $\sigma = \sqrt{\mu}$ to adjust for the higher variability expected in the size selection for longer fragments. Fraction of 1-bit (assignable to only one isoform) paired and single-end reads is plotted ($\pm$s.d.). (**c**) Distribution of errors for paired-end and single-end estimation as coverage increases (measured in RPK). (**d**) Histogram shows library insert length distribution computed from read pairs mapped to long constitutive 3′ UTRs in a human testes RNA-seq data set. In the example exon trio shown (similar to that in **Fig. 1d**), the insert length distribution assigns a higher probability to the top (inclusion) isoform than to the bottom (exclusion) isoform, for which the inferred insert length is improbably small. (**e**) Fraction of assignable 2-bit and 1-bit reads ($\pm$s.d.) for paired-end and single-end reads as a function of the number of intervening constitutive exons, k.
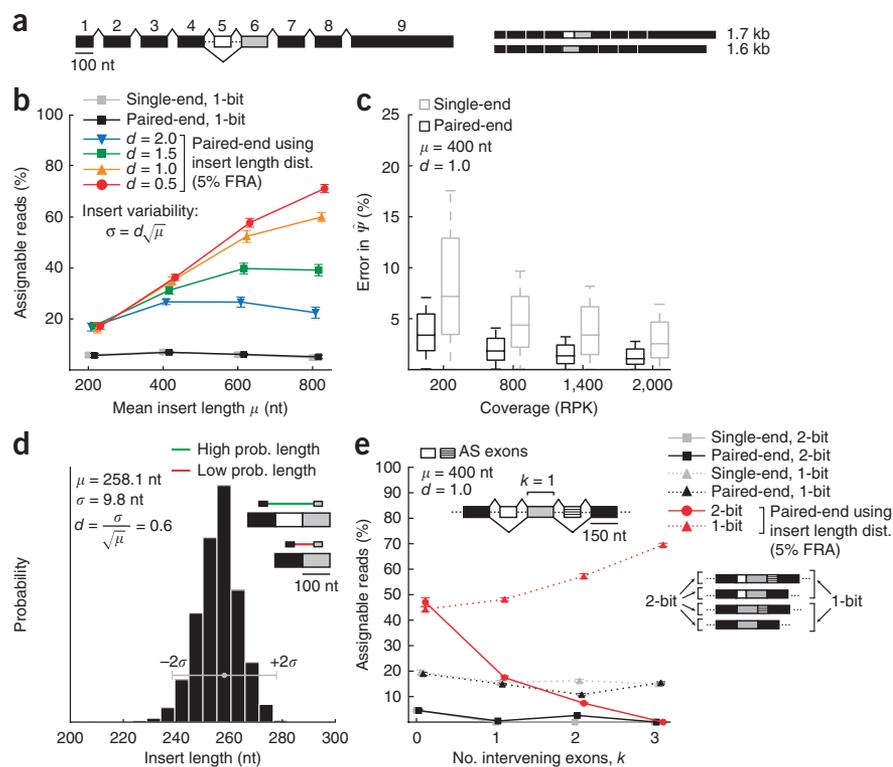
For $d < 1.5$, the proportion of assignable reads increased steadily with insert length (**Fig. 5a**), as larger inserts make it more likely that reads from a pair will fall on opposite sides of an alternative exon and be probabilistically assignable. Thus, if dispersion is kept near or below 1, use of longer insert lengths should yield more information about splicing. However, changing mRNA fragment size can have other effects on RNA-seq experiments, potentially affecting the priming and reverse-transcription steps and the sampling of mRNAs of different lengths.

To assess the nature and extent of these effects, we generated libraries with mean insert lengths of ~100 nt and ~280 nt from the same RNA sample, derived from control mouse myoblasts, and generated similar libraries from myoblasts depleted of the splicing factor CUGBP1 (**Supplementary Fig. 8a**). Gene expression estimates were relatively unaffected by insert length for mRNAs 1 kilobase (kb) or longer, but, as expected, read coverage of very short mRNAs only a few hundred bases in length was reduced by ~20–40% in the longer-insert libraries (**Supplementary Fig. 8b**). The precise pattern of fluctuations in read coverage along constitutive regions differed between libraries with different insert sizes but was highly correlated between libraries generated with similar insert sizes (**Supplementary Fig. 8c**). The reproducibility of the patterns of local fluctuations indicated that they are primarily determined by fragment size[21]—which could affect RNA secondary structure and therefore the priming and reverse-transcription steps—rather than by technical noise. Because such fluctuations could affect analysis of alternative splicing, comparisons made between RNA-seq data sets prepared using similar library insert lengths will be most accurate. Changes in gene expression resulting from the knockdown of CUGBP1 were detected highly reproducibly at the two different library insert sizes ($r \approx$ 0.9; **Supplementary Fig. 8d**), indicating that library insert size

can be varied at least over this range without affecting the ability to detect changes in expression. The overall magnitude of read-coverage fluctuations was only modestly greater for the 100-nt-insert library than for the library with 280-nt inserts (**Supplementary Fig. 8e**), but further tests of longer insert libraries will be needed to determine the magnitude and impact of the expected increases in local read-coverage fluctuations. Overall, the optimal insert size to use in an RNA-seq experiment will depend on the importance one places on outputs such as detection of splicing changes relative to efficient capture of short mRNAs.

More accurate Ψ values using insert length information

Insert length information is incorporated in MISO by probabilistic assignment of read pairs to isoforms that are consistent with both individual reads, weighting the assignment of read pairs by the relative probability of observing the given insert length, according to the structure of each isoform. To quantify the impact of the increased assignability of reads on accuracy of Ψ estimates, we simulated paired-end reads from a typical gene model containing an alternative exon (**Fig. 5a**). Use of paired-end reads with insert length information markedly increased the accuracy of estimates of Ψ in simulations, reducing the error by a factor of ~2–5 (**Fig. 5c**). With a typical gene model containing a typically sized alternative exon, applying the $\hat{\Psi}_{\text{MISO}}$ estimation method that makes use of paired-end length information, rather than the standard $\hat{\Psi}_{\text{MISO}}$ estimate, reduced the error in estimated Ψ from about 8% to ~4% for a gene with RPK of 200, and the error was further reduced to ~2% at higher coverage values.

Applications to complex alternative splicing

Paired-end data can also be used to make inferences about isoform levels for genes that contain multiple alternative splicing

events. To assess how much information can be gained about splicing by paired-end sequencing in these cases, we simulated reads from a gene model containing a pair of alternative exons while varying the number of exons, k, separating the two alternative exons (**Fig. 5e**). In this gene model, 2 bits of information are required to uniquely specify an isoform: 1 bit to indicate whether the first alternative exon was included or excluded, and 1 bit to describe the splicing of the second alternative exon. Reads that can be uniquely assigned to one of the four isoforms are therefore considered '2-bit reads', whereas reads that are assignable to exactly two of the four isoforms are considered '1-bit reads' (**Fig. 5e**). When $k = 0$, a single read may overlap the junction of the two alternative exons or the junction between the flanking constitutive exons, providing 2 bits of information. For $k \geq 1$, no 2-bit reads occurred for the typical read and exon lengths used in the simulation, but read pairs can sometimes provide 2 bits of information—for example, if the two reads derive from the two alternative exons or from junctions that are informative about the splicing of these exons, though this is fairly rare. When insert length information is used and probabilistically assignable reads are considered, far more read pairs yield 1 or even 2 bits of information (**Fig. 5c** and Online Methods), indicating that short-read data has some potential to address more complex alternative splicing events.

The MISO model generalizes to the isoform-centric case in which genes express arbitrarily many isoforms through alternative splicing (**Supplementary Note** and **Supplementary Figs. 9–11**); an application of MISO to estimate the abundance of four isoforms from the *GRIN1* gene is shown in **Supplementary Figure 12**. However, sequencing methods involving longer reads, longer library insert lengths or both are needed to quantify isoforms in genes with multiple distant alternative splicing events.

DISCUSSION

Alternative splicing is highly regulated during development and differentiation, and misregulation of RNA processing underlies a variety of human diseases[2,22]. Because individual alternative exons typically represent only a few percent of the length of the mRNA, analysis of splicing requires greater sequencing depth and more powerful statistical methods than are needed to study gene expression. The MISO model introduced here represents a detailed probabilistic model of RNA-seq, and it has a variety of advantages, including improved accuracy and the ability to analyze all major types of alternative pre-mRNA processing at either the exon level or the isoform level.

This study also has important implications for the design of RNA-seq experiments. Our analyses indicate that paired-end sequencing yields far more information about alternative exons and isoforms than single-end sequencing does. This information derives primarily from cases in which the reads in a pair flank an alternative exon, so that the inclusion and exclusion isoforms imply different intervening mRNA lengths. Use of somewhat longer mRNA fragments, of 300 bases or more, in library preparation should generally enhance isoform inference by increasing the occurrence of such read pairs, with tradeoffs related to the capture of very short mRNAs and changes in the pattern and extent of local fluctuations in read coverage along exons. Our analyses of read-coverage fluctuations strongly imply that RNA-seq–based comparisons of expression and splicing will be most accurate when the insert lengths of the libraries being compared are similar. In some cases a mixed experimental design involving use of different library insert sizes from a single sample may be appropriate—for example, combining one lane of paired-end sequencing from a longer-insert RNA-seq library for inference of mRNA isoform abundance together with a lane of shorter-insert single-end sequencing for analysis of gene expression.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

AUTHOR CONTRIBUTIONS
Y.K., development of MISO model and software, analyses involving MISO, writing of main text and methods; E.T.W., hnRNP H CLIP-seq experiments and associated computational analyses, CUGBP1 knockdown RNA-seq experiments and associated computational analyses; E.M.A., development of model and statistical analysis, writing of methods; C.B.B., development of MISO model, contributions to computational analyses, writing of main text.

1. Matlin, A.J., Clark, F. & Smith, C.W.J. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* **6**, 386–398 (2005).
2. Christofk, H.R. *et al.* The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* **452**, 230–233 (2008).
3. Rowen, L. *et al.* Analysis of the human neurexin genes: alternative splicing and the generation of protein diversity. *Genomics* **79**, 587–597 (2002).
4. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
5. Mortazavi, A., Williams, B.A.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* **5**, 621–628 (2008).
6. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
7. Yassour, M. *et al.* Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 3264–3269 (2009).
8. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
9. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).

10. Griffith, M., Griffith, O.L., Mwenifumbo, J., Goya, R. & Morrissy, A.S. Alternative expression analysis by RNA sequencing. *Nat. Methods* **7**, 843–847 (2010).
11. Venables, J.P. *et al.* Identification of alternative splicing markers for breast cancer. *Cancer Res.* **68**, 9525–9531 (2008).
12. Jiang, H. & Wong, W.H. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25**, 1026–1032 (2009).
13. Venables, J.P. *et al.* Cancer-associated regulation of alternative splicing. *Nat. Struct. Mol. Biol.* **16**, 670–676 (2009).
14. Xiao, X. *et al.* Splice site strength-dependent activity and genetic buffering by poly-G runs. *Nat. Struct. Mol. Biol.* **16**, 1094–1100 (2009).
15. Millevoi, S. & Vagner, S. Molecular mechanisms of eukaryotic pre-mRNA 3′ end processing regulation. *Nucleic Acids Res.* **38**, 2757–2774 (2010).
16. Alkan, S.A., Martincic, K. & Milcarek, C. The hnRNPs F and H2 bind to similar sequences to influence gene expression. *Biochem. J.* **393**, 361–371 (2006).
17. Millevoi, S. *et al.* A physical and functional link between splicing factors promotes pre-mRNA 3′ end processing. *Nucleic Acids Res.* **37**, 4672–4683 (2009).
18. Honoré, B., Baandrup, U. & Vorum, H. Heterogeneous nuclear ribonucleoproteins F and H/H′ show differential expression in normal and selected cancer tissues. *Exp. Cell Res.* **294**, 199–209 (2004).
19. Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A. & Burge, C.B. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* **320**, 1643–1647 (2008).
20. Mayr, C. & Bartel, D.P. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673–684 (2009).
21. Li, J., Jiang, H. & Wong, W.H. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.* **11**, R50 (2010).
22. Cooper, T.A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* **136**, 777–793 (2009).

ONLINE METHODS

Software availability. All software implementations used in this paper are available at http://genes.mit.edu/burgelab/miso.

Cell culture and CUG-BP1 RNA-seq experiments. Control and CUG-BP1 knockdown C2C12 myoblasts[23] were a generous gift of Carol Wilusz (Colorado State University). Cells were cultured at 37 °C, 5% CO_2 in Dulbecco's modified Eagle's Medium containing 20% (v/v) FBS and maintained in subconfluency conditions. Total RNA was isolated from cells at 50–60% confluency by lysis with Trizol, followed by chloroform extraction, precipitation and cleanup plus DNase treatment on RNeasy columns (Qiagen). Total RNA was poly(A)-selected using poly(T) Dynabeads (Invitrogen) and prepared for Illumina sequencing. After adaptor ligation, libraries were agarose gel purified; two 1-mm-thick bands centered at ~250 and ~400 nt were excised with razor blades. These gel-purified products were amplified by 11 cycles of PCR using adaptor primers and subjected to a final gel purification to maintain a tight size distribution.

Mapping and processing of RNA-seq reads. Reads were aligned to the genome and to a precomputed set of splice junctions (as described[4]) using the Bowtie alignment program[24]. Reads were required to map with two mismatches or fewer, and junction reads were required to include four bases or more from each spanning exon. For simplicity, we considered only reads that map uniquely to the union of the genome and splice junctions, and correct for differing uniqueness of different regions by excluding non-unique positions from the analysis; for an alternative treatment, see ref. 25. Alternative exons used in all analyses were derived as described[4], by considering an exon as skipped if it is supported by one or more annotated ESTs or cDNAs. Alternative poly(A) sites were compiled from PolyA DB[26]. Here we used a precomputed junction database, but an alternative is to discover splice junctions *de novo* using one of several available tools for junction discovery and transcript annotation (for example, refs. 8,9,27–32). The human heart and testis RNA-seq data listed in **Supplementary Table 1** were provided by G. Schroth (Illumina) and are available upon request. The breast cancer RNA-seq data were generated by R. Nutiu (MIT) from tissue provided by S. Abou Elela (University of Sherbrooke). Read data relevant to **Figure 2** will be provided upon request.

Notation and model. After aligning reads to the genome and splice junctions, we considered only reads that map uniquely to genes for both exon-centric and isoform-centric analyses. Assuming N reads that align to a given gene with K isoforms, each read R_n (where $1 \leq n \leq N$) is associated with a vector with components R_n^1 through R_n^K indicating its compatibility with the K different isoforms: if the nth read maps to the kth isoform of the gene, then R_n^k is set to 1, and 0 otherwise. Given a set of uniquely aligning reads, we seek to infer the 'percentage spliced isoform' values $\vec{\Psi}$, representing the relative abundances of the gene's isoforms. Here, Ψ_k denotes the fraction of mRNAs corresponding to the kth isoform (and thus $\sum_k \Psi_k = 1$). Faithfully modeling the physical process of fragmentation and subsequent size selection is not yet feasible, but the general effect of these processes on the data is that the probability of sampling a read from an mRNA increases approximately linearly with the mRNA's length. To capture this effect, we rescale the isoform abundances Ψ_1 through Ψ_k by the numbers c_1 through c_k of reads of length RL that could be generated from these isoforms: if l_k is the length of the kth isoform, then $c_k = l_k - \text{RL} + 1$. In the two-isoform case, this rescaling yields the values

$$\Psi_{f1} = \frac{c_1 \Psi_1}{c_1 \Psi_1 + c_2 \Psi_2} \text{ and } \Psi_{f2} = 1 - \Psi_{f1} \qquad [1]$$

corresponding to the expected proportion of reads generated from each isoform. Let I_n be a variable representing the isoform from which the nth read was sequenced, where $1 \leq n \leq N$. The rescaled abundances then correspond to the probability that the nth read was generated from the kth isoform, denoted $P(I_n | \vec{\Psi})$, accounting for fragmentation. Given the assignment of the nth read to the kth isoform, we then define the probability of observing a specific read from that isoform. This probability will depend on a set of fixed parameters Θ of the experiment, such as the length of a read (RL) and the minimum overhang length for reads that span splice junctions. To account for uniqueness of reads in the alignment, let $m(\text{RL}, I_n)$ be the number of mappable read positions in some isoform I_n for an experiment where the read length is RL. For convenience, only reads and read positions that satisfy the overhang constraint are considered mappable. The probability of observing read R_n from the kth isoform, denoted $P(R_n | I_n = k, \Theta)$, is defined to be uniform over the number of reads observable from isoform k—that is, R_n^k equals 1 with probability $1/m(\text{RL}, I_n)$, and is 0 otherwise. Our goal is to invert the generative process by which reads are produced and infer the underlying isoform abundances that best explain the observed reads. Formally, this is achieved by computing a probability distribution, called the 'posterior', over the unobserved random variable ($\vec{\Psi}$), given the RNA-seq data. This is done using Bayes' rule, which states how the posterior distribution can be computed in terms of two quantities: (i) the probability of the data given a setting of the variable, referred to as the 'likelihood' of the data, and (ii) our *a priori* expectation about the values of this variable, referred to as the 'prior' distribution. The relationships between the prior, likelihood and posterior distributions are depicted in **Supplementary Figure 3**. In our case, the prior specifies our expectation about the value of $\vec{\Psi}$ before observation of reads (we use a prior that is unbiased, not favoring any particular abundance value), and the likelihood specifies the probability of observing a set of reads given a $\vec{\Psi}$ value. The posterior describes the probability of $\vec{\Psi}$ given a set of reads. Given a set of N reads $R_{1:N}$, the posterior distribution denoted $P(\vec{\Psi} | R_{1:N})$, Bayes' rule gives

$$P(\vec{\Psi}) \propto P(R_{1:N} | \vec{\Psi}) P(\vec{\Psi})$$

Intuitively, this equation states that the probability of a set of abundances given the reads are proportional to our *a priori* expectation about the values of these abundances (the prior), weighted by how likely reads we observed are to have been produced from these abundances (the likelihood). To compute the posterior, we need to consider all possible assignments of every read to each isoform and use the probabilities defined above to score these assignments:

$$P(\vec{\Psi} | R_{1:N}) \propto \sum_{I_1 = 1}^{K} \cdots \sum_{I_N = 1}^{K} \prod_{n=1}^{N} P(R_n | I_{1:N}, \Theta) P(I_{1:N} | \vec{\Psi}) P(\vec{\Psi})$$

where I_n indicates the isoform from which the nth read was generated. For exon-centric analyses where there are only two isoforms, we use a prior distribution uniform over [0, 1], which is a special case of the Beta distribution commonly used as a prior in Bayesian statistics[33]. In this case, our model is a variant of the well-studied 'Beta-Bernoulli' model[33]. Whereas in the general case of many isoforms inference is performed using approximate techniques[34,35], an analytic solution can be obtained in the two-isoform case under certain assumptions about the prior distribution (**Supplementary Note**). We can extend the model to isoform-centric analyses (where there are many isoforms) using the Dirichlet-Multinomial distributions[33], which are the multivariate generalization of the Beta-Bernoulli distributions used in the exon-centric case. The general model is specified as follows:

$$\vec{\Psi} \sim \text{Dirichlet}(\vec{\alpha}) \text{ once for every gene } g,$$

$$I_n \mid \vec{\Psi} \sim \text{Multinomial}(1, \vec{\Psi}) \text{ for every read } n \text{ that maps to gene } g,$$

$$R_n \sim P(R_n \mid I_n, \Theta) \text{ for every read } n \text{ that maps to gene } g$$

where the corrected abundances that account for the lengths of multiple isoforms are now defined as a vector $\vec{\Psi}_f$, and where an entry

$$\Psi_{fk} = (c_k \Psi_k) \Big/ \left(\sum_{j=1}^{K} c_j \Psi_j \right)$$

corresponds to the probability of sampling a read from the kth isoform. As above, we consider a symmetric Dirichlet distribution with all parameters equal to encode a uniform prior over $\vec{\Psi}$, not favoring any particular distribution of isoforms. As in the exon-centric case, we seek the posterior distribution $P(\vec{\Psi} \mid R_{1:N})$, which can be obtained by Bayes' rule. A graphical model representation of MISO summarizing the relations between all these variables is shown in **Supplementary Figure 9**, where variables are indicated by nodes and probabilistic dependencies are indicated as edges between the nodes.

Quantitation of diverse classes of alternative pre-mRNA processing events. By representing alternative pre-mRNA processing events generically as a mixture of isoforms, with each isoform defined by a list of exon coordinates, it is possible to quantify diverse classes of events can, including alternative 5′ and 3′ splice sites, alternative first and last exons, tandem APA sites, mutually exclusive exons and retained introns. Different event types will be supported by distinct types of reads—for example, tandem APA events are currently quantified using reads that are unique to the extended isoform and reads that map to the core region shared by both isoforms. The Ψ value in this case is defined as the ratio of the abundance of the long isoform relative to the sum of the abundances of long and short isoforms. Intuitively, MISO uses the density of reads in the extended region relative to that in the core region to estimate this quantity. Similarly, the absolute and relative sizes of alternative regions will affect the read coverage and the power to reliably quantify isoform abundances (for example, alternative 3′ splice sites can differ by as few as three bases, whereas tandem APA events typically differ by ~1 kb).

Incorporation of paired-end information. In single-end sequencing, it is sufficient to represent a read by the set of isoforms it could

have been derived from. For paired-end reads, it is also necessary to incorporate information about the insert lengths that are consistent with each read. We represent a paired-end read with a pair of parameters, (R_n, λ_n), where the first element, R_n, is a binary vector representing the alignment of reads to isoforms as in the single-end case: $R_n^k = 1$ if the nth read aligns to the kth isoform; $R_n^k = 0$ otherwise. The second element, λ_n, is a vector of observed inserted fragment lengths, where λ_n^k is set to the length of the insert implied by isoform k for the nth read pair, assuming the read was consistent with isoform k and is undefined otherwise. To score how likely observed insert lengths are and use this information to assign reads to isoforms, we modeled the distribution of isoform lengths in the mRNA-seq sample. This distribution is computed empirically by mapping read pairs to long constitutive 3′ UTRs, whose size is much larger than the expected insert length selected during the sample library preparation. The mean (μ) and variance (σ^2) of this distribution are then used to compute the probability that a read pair came from each isoform, if it is consistent with more than one. The probability of assigning the nth read pair to the kth isoform given $\vec{\Psi}$ depends on both the lengths of the isoforms l_1 through l_k and the mean of the insert length distribution, μ,

$$P(I_n = k \mid \vec{\Psi}, \mu) = \frac{(l_k - \mu + 1)\Psi_k}{\sum_{j=1}^{K} (l_j - \mu + 1)\Psi_j}$$

The probability of observing a paired-end read (R_n, λ_n) given its assignment to an isoform k and the experiment's parameters Θ, denoted with $P(R_n, \lambda_n \mid I_n = k, \Theta)$, is assumed to be uniform over the number of fragments of the relevant size that can be generated from isoform k; that is, $R_n^k = 1$ with probability $1/m(\lambda_n^k, I_n)$, and it is 0 otherwise. We modeled the empirical fragment length distribution $P(\lambda_n \mid \mu, \sigma)$ as a discretized normal distribution, with mean μ and s.d. σ. The two parameters μ, σ can be set to fit the empirical distribution of insert lengths in any RNA-seq sample before the MISO inference procedure is run. All paired-end simulations were conducted with two 36-nt reads, where insert lengths were sampled from a discretized normal distribution whose mean μ and dispersion d were varied as described in the main text.

Analytic estimates of Ψ in exon-centric analyses. In exon-centric analyses we have two isoforms. In this case, we derive the maximum *a posteriori* estimate of Ψ, $\vec{\Psi}_{\text{MISO}}$, using single-end reads (a complete derivation is provided in the **Supplementary Note**). $\vec{\Psi}_{\text{MISO}}$ is a function of five main parameters: the numbers of inclusion, exclusion and common reads (N_I, N_E, N_C) and the fixed conditional probabilities of a read given its assignment to the first and second isoforms (p_1 and p_2, respectively). Under the assumption of a uniform prior on Ψ, it is sufficient to define an estimate for Ψ_f and transform to get an estimate of Ψ, using the inverse of equation (1). The derivation of $\hat{\Psi}_f$ is then reduced to solving a quadratic equation, whose relevant solution is

$$\hat{\Psi}_f = \frac{A - \sqrt{B + C}}{D}$$

(described fully in **Supplementary Note**). We then obtain an estimator for $\hat{\Psi}_{\text{MISO}}$ by plugging $\hat{\Psi}_f$ into the inverse of equation (1). Our estimate is compared to $\hat{\Psi}_{\text{A3SS}}$ and the $\hat{\Psi}_{\text{SJ}}$ estimates from

ref. 4 in **Supplementary Figure 2**. Calculation of these measures and proofs of unbiasedness are given in the **Supplementary Note**.

Estimates of Ψ in isoform-centric analyses. To estimate the full posterior distribution over abundances in either exon- or isoform-centric analyses, analytic solutions are not available, and approximate inference techniques must be used instead[35]. The correction of isoform abundances needed to account for the fragmentation step leads to violations of the mathematically convenient conjugacy properties of traditional Dirichlet-Multinomial mixture models, which are required for standard methods of performing inference in such models, such as Gibbs sampling[33]. To perform efficient inference in our model, we devised a Markov chain Monte Carlo (MCMC) inference scheme based on a novel proposal distribution. We use a hybrid MCMC sampler that combines the Metropolis-Hastings (MH) algorithm with a Gibbs sampler[34]. In MH, a proposal distribution Q is used to estimate the target distribution $P(\vec{x})$, where P can be evaluated up to proportionality on any set of states but cannot easily be sampled from. Transitions to different states of P are repeatedly proposed from Q, and these are stochastically accepted or rejected according to the MH ratio, α:

$$\alpha = \min\left(\frac{P(x_{t+1})Q(x_t; x_{t+1})}{P(x_t)Q(x_{t+1}; x_t)}, 1 \right)$$

In our case, P is the joint distribution on $\vec{\Psi}$ and the latent assignment of reads to isoforms $I_{1:N}$. A substantial challenge for inference in our model is to construct a proposal distribution Q that efficiently proposes high-probability $\vec{\Psi}$ values under P while respecting the constraint that $\vec{\Psi}$ must sum to 1. To achieve this, we use the logistic-normal distribution[36] to construct a random walk in the simplex space by drifting over the parameters of the Beta distributions from which $\vec{\Psi}$ values are drawn. See **Supplementary Note** for a full derivation of the inference scheme and **Supplementary Figure 11** for the resulting algorithm.

Computation of Bayesian confidence intervals. Given a posterior distribution over $\vec{\Psi}$ obtained with the proposed MCMC sampler, a Bayesian CI for Ψ_k is computed using the method described[37]. The $100(1 - \alpha)\%$ Bayesian CI is an interval (a, b) where the probability of a value for Ψ_k being contained in (a, b) is $100(1 - \alpha)\%$. Let $S = \{\psi_k^i\}_{i=1}^n$ be a set of n posterior samples for a given Ψ_k. The $100(1 - \alpha)\%$ interval (a, b) is computed as: $\psi_k^{(\alpha/2)n}$, $\psi_k^{(1-\alpha/2)n}$), where $\psi_k^{(\alpha/2)n}$ is the $(\alpha/2)n$th smallest sample in S, and $\psi_k^{(1-\alpha/2)n}$ is the $(1 - \alpha/2)n$th smallest sample in S. Such an interval is a consistent estimator of the Bayesian CI[37].

Statistical test for differential isoform expression using Bayes factors. To detect the differential expression of an isoform between two samples A and B, we use a two-sided point null hypothesis test. Let $\delta = \Psi_A - \Psi_B$, where Ψ_A, Ψ_B correspond to the expression levels of the isoform in samples A, B, respectively. The null hypothesis (H_0) states that $\delta = 0$, and the alternative hypothesis (H_1) that $\delta \neq 0$. To choose between the two competing hypotheses, we compute the BF[38], which can be interpreted as the weight of the evidence in the data D in support of H_1 over H_0:

$$BF = \frac{P(D|H_1)P(H_1)}{P(D|H_0)P(H_0)}$$

The BF can be accurately estimated using the Savage-Dickey density ratio[33]—that is, by calculating it as a ratio of the posterior density at $\delta = 0$ under H_1 and the prior density under H_1 at the same point:

$$BF \approx \frac{P(\delta = 0 | H_1)}{P(\delta = 0 | D, H_1)}$$

We assume a uniform prior over Ψ_A and Ψ_B, which yields a prior distribution that peaks where $\delta = 0$, corresponding to the case of no differential regulation between the conditions (that is, a 'triangular prior' where $P(\delta = 0 | H_1) = 1$). This reduces the BF calculation to $1/P(\delta = 0 | H_1, D) = 1$.

Analysis of quantitative reverse-transcription PCR data. Only alternative exons meeting the coverage criteria outlined above were used. To ensure detectable alternative splicing of the exon in the breast cancer sample, we required that the qRT-PCR value be greater than 0 and smaller than 1. To correct for the length bias in the qRT-PCR data when computing the overlap between qRT-PCR data and MISO CIs, we used an out-of-sample cross-validation scheme to calculate an adjusted qRT-PCR value (as described in **Supplementary Note**).

hnRNP H CLIP-seq experiment and data analysis. CLIP-seq for hnRNP H was performed as described[14]. Read fragments of size 15–30 nt were aligned to the human genome (hg18) and a pre-computed set of splice junctions. CLIP tag densities were normalized by RPKM values estimated from the hnRNP H control condition. For analyses of alternative and constitutive exons, only tags in the exon body and in the intronic region upstream and downstream of exons (using at most 250 nt of half the intron proximal to the exon) were considered. For analysis of tandem 3′ UTRs, a window of −250 to 500 nt relative to the core poly(A) site (based on PolyA DB annotation) was used, excluding regions that are less than 500 nt away from the extension poly(A) site. Plotted values in **Figure 4b** are means of mean CLIP tag densities from 100 subsamplings of exons with corresponding BF values, matched for their gene's RPKM to control for the inherent correlation between BF and expression level. Error bars are means of standard errors from subsamplings.

Simulations of single-end and paired-end reads. All single-end read simulations were performed with reads 36 nt long having an overhang constraint of 4. Paired-end simulations were peformed with two 36-nt reads, with varying mean insert lengths (μ) and dispersion (d), as described in the main text. Coverage was measured in reads per kilobase (RPK) of constitutive exons of a gene model. For *GRIN1*, reads were simulated from the four described isoforms at 1,000-RPK coverage, using the exon sizes given in the UCSC Known Genes table for the mouse genome (mm9). All sampler results were run for 10,000 iterations using a burn-in of 500 iterations and a 10:1 thinning ratio. Posterior marginal distributions were averaged across 50 independent chains and runs of the sampler.

23. Zhang, L., Lee, J.E., Wilusz, J. & Wilusz, C.J. The RNA-binding protein CUGBP1 regulates stability of tumor necrosis factor mRNA in muscle cells: implications for myotonic dystrophy. *J. Biol. Chem.* **283**, 22457–22463 (2008).
24. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

25. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500 (2010).
26. Zhang, H., Hu, J., Recce, M. & Tian, B. PolyADB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.* **33**, D116–D120 (2005).
27. Wang, L. *et al.* A statistical method for the detection of alternative splicing using rna-seq. *PLoS ONE* **5**, e8529 (2010).
28. Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9**, R175 (2008).
29. Trapnell, C., Pachter, L. & Salzberg, S.L. Tophat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
30. Ameur, A., Wetterbom, A., Feuk, L. & Gyllensten, U. Global and unbiased detection of splice junctions from rna-seq data. *Genome Biol.* **11**, R34 (2010).
31. Au, K.F., Jiang, H., Lin, L., Xing, Y. & Wong, W.H. Detection of splice junctions from paired-end rna-seq data by splicemap. *Nucleic Acids Res.* **38**, 4570–4578 (2010).
32. Wu, T.D. & Nacu, S. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
33. O'Hagan, A. & Forster, J. Kendall's advanced theory of statistics, vol. 2b: Bayesian inference. (2nd edn.) *J. Am. Stat. Assoc.* **100**, 1465–1466 (2005).
34. Liu, J.S. *Monte Carlo Strategies in Scientific Computing (Springer Series in Statistics)* (Springer, 2008).
35. Airoldi, E.M. Getting started in probabilistic graphical models. *PLOS Comput. Biol.* **3**, e252 (2007).
36. Aitchison, J. & Shen, S.M. Logistic-normal distributions: some properties and uses. *Biometrika* **67**, 261–272 (1980).
37. Chen, M. & Man Shao, Q. Monte Carlo estimation of Bayesian credible and HPD intervals. *J. Comput. Graph. Statist.* **8**, 69–92 (1998).
38. Kass, R.E. & Raftery, A.E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).