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).

RECEIVED 18 AUGUST; ACCEPTED 8 OCTOBER; PUBLISHED ONLINE 7 NOVEMBER 2010; DOI:10.1038/NMETH.1528

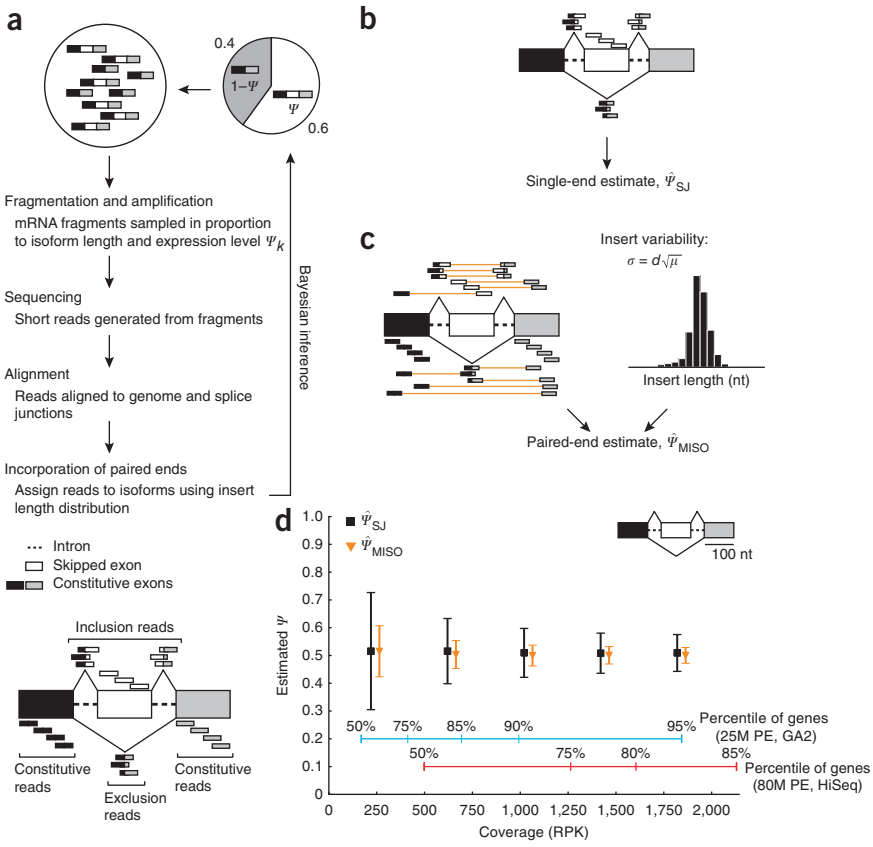


Figure 1 | More accurate inference of splicing levels using MISO. (**a**) Generative process for MISO model. White, alternatively spliced exon; gray and black, flanking constitutive exons. RNA-seq reads aligning to the alternative exon body (white) or to splice junctions involving this exon support the 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 (±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

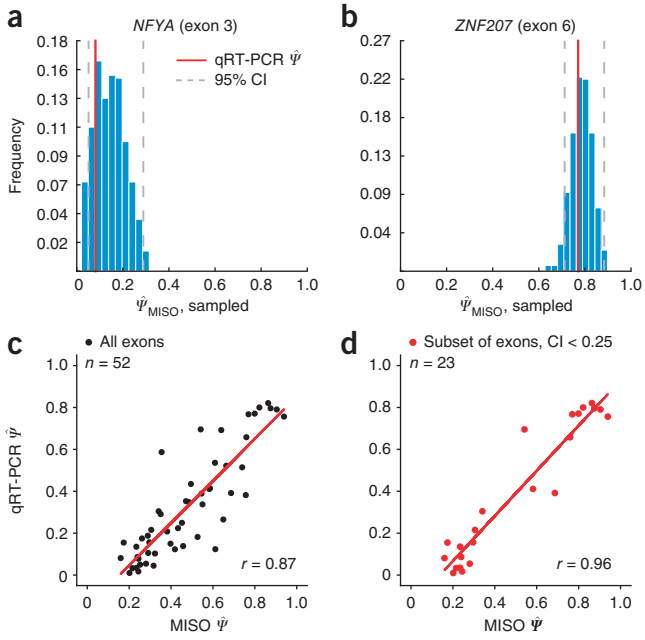


Figure 2 | MISO CIs for Ψ values and qRT-PCR validation. qRT-PCR measurements from ref. 13 for a set of 52 alternatively skipped exons were compared to MISO posterior mean estimates of Ψ, denoted $\hat{\Psi}_{MISO}$. Full listing of events is given in **Supplementary Table 1**. (**a**,**b**) The Ψ posterior 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].

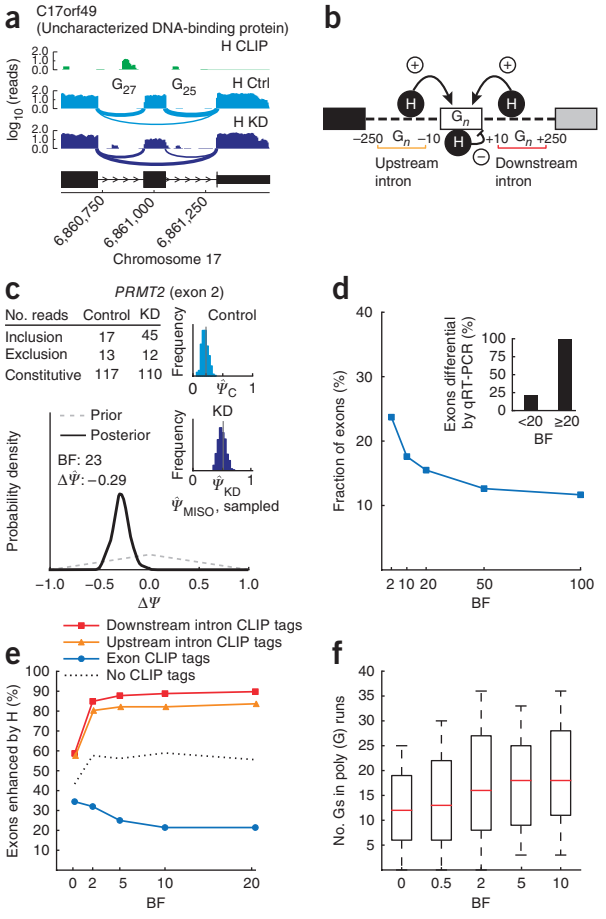


Figure 3 | Bayes factor analysis of hnRNP H regulation of exon splicing. (**a**) CLIP tag density (H CLIP; green) and RNA-seq read densities in hnRNP H–knockdown and control conditions (H KD and H Ctrl; light and dark blue, respectively) for an alternative exon in human C17orf49. Number of guanines in 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.

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

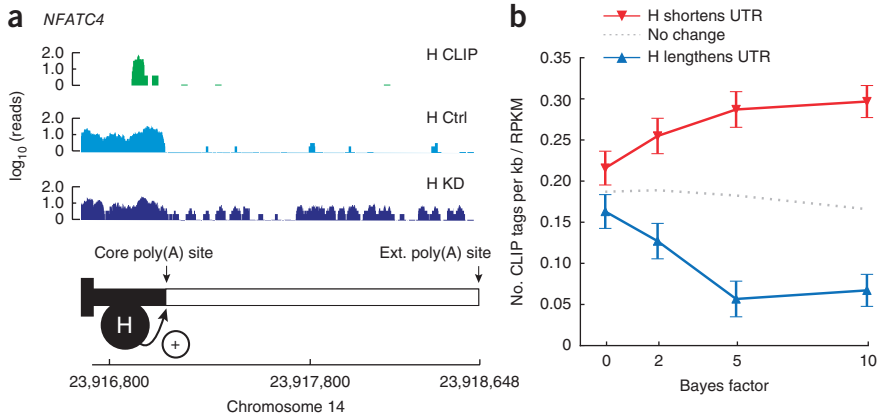


Figure 4 | Bayes factor analysis implicates hnRNP H in alternative cleavage and polyadenylation. (**a**) CLIP tag density (H CLIP; green) and RNA-seq read densities in hnRNP H control and knockdown conditions (H Ctrl and H KD; light and dark blue, respectively) along the 3′ UTR of the *NFATC4* gene. 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).

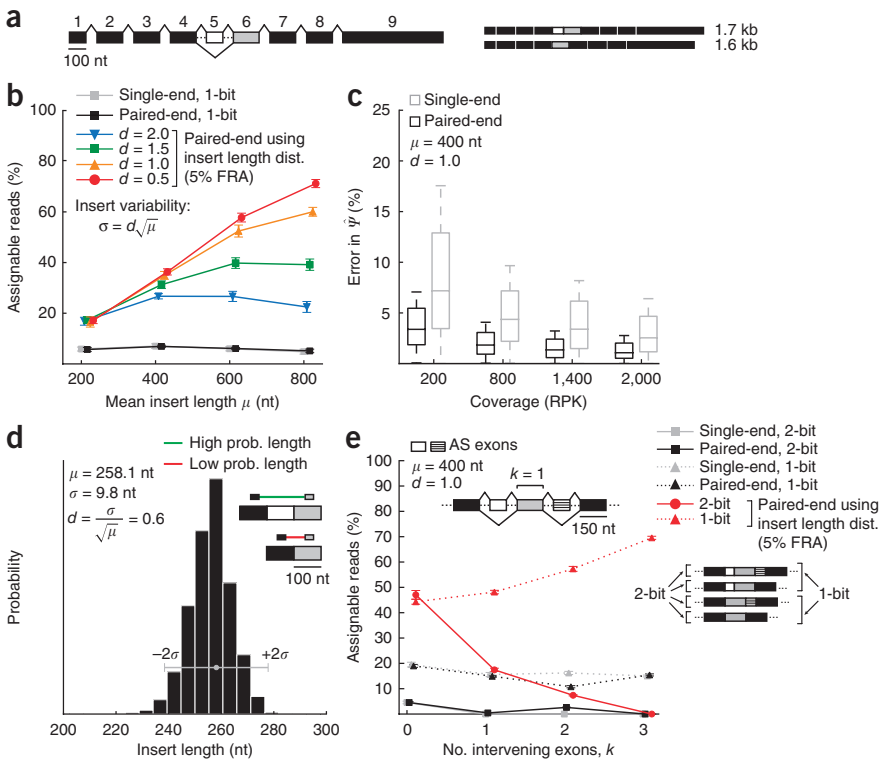


Figure 5 | Improved estimation of isoform abundance using paired-end reads. (**a**) Representative gene model with 100-nt first exon, 100-nt skipped exon (exon 5, in white), 150-nt constitutive exons and 600-nt last exon. (**b**) We simulated reads from the two-isoform gene model shown in **a** while 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 (±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 (±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 $\hat{\Psi}_{\mathrm{MISO}}$ estimation method that makes use of paired-end length information, rather than the standard $\hat{\Psi}_{\mathrm{MISO}}$ 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/.

Accession codes. Gene Expression Omnibus: GSE23694.

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

ACKNOWLEDGMENTS

We thank C. Wilusz (Colorado State University) for the gift of the CUGBP1-knockdown and control C2C12 cells; R. Darnell for advice regarding CLIP-seq protocols; S. Abou Elela, V. Butty, R. Nutiu and G. Schroth for sharing RNA-seq data; and J. Ernst, D. Gresham, M. Guttman, F. Jäkel, E. Jonas, F. Markowetz, D. Roy, R. Sandberg, T. Velho, X. Xiao and members of the Burge lab for insightful discussions and comments on the manuscript. This work was supported by grants from the US National Science Foundation (E.M.A.) and the US National Institutes of Health (E.M.A. and C.B.B.).

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.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

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 $1 \le n \le N$) 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 $c_k = l_k - \mathrm{RL} + 1$. In the two-isoform case, this rescaling yields the values

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

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 $1 \le n \le N$. The rescaled abundances then correspond to the probability that the nth read was generated from the kth isoform, denoted $P(I_n \mid \vec{\Psi})$, 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 $m(\mathrm{RL}, I_n)$ 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 $P(R_n \mid I_n = k, \Theta)$, is defined to be uniform over the number of reads observable from isoform k—that is, R_n^k equals 1 with probability $1/m(\mathrm{RL}, I_n)$, 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 $P(\vec{\Psi} \mid R_{1:N})$, Bayes' rule gives

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

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:

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

doi:10.1038/nmeth.1528

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:

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

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

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

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

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

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, $\hat{\vec{\Psi}}_{\text{MISO}}$, using single-end reads (a complete derivation is provided in the **Supplementary Note**). $\hat{\vec{\Psi}}_{\text{MISO}}$ 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:

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

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:

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

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 $P(\delta = 0 \mid H_1) = 1$). This reduces the BF calculation to $1/P(\delta = 0 \mid H_1, D) = 1$.

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 precomputed 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).

doi:10.1038/nmeth.1528

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).

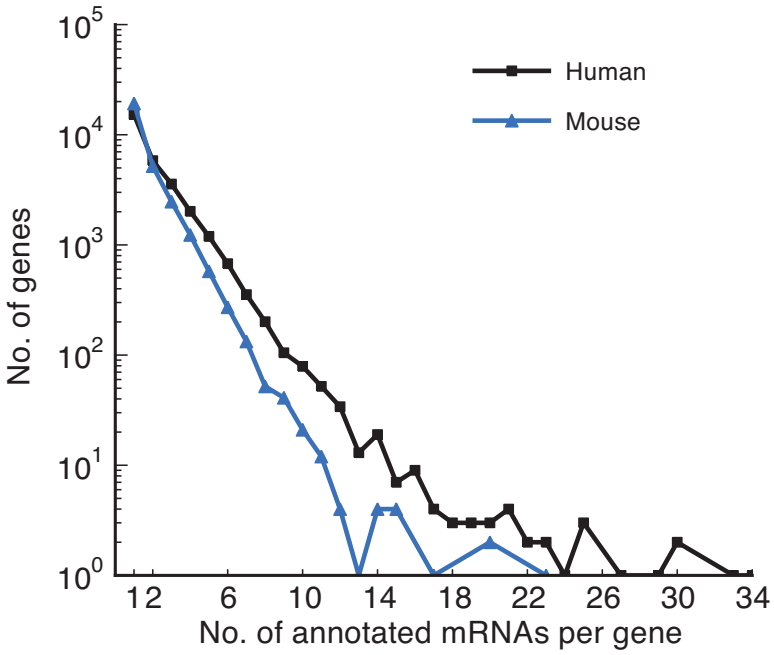
Nature Methods

Analysis and design of RNA sequencing experiments for identifying mRNA isoform regulation

Yarden Katz, Eric T Wang, Edoardo M Airoldi & Christopher B Burge

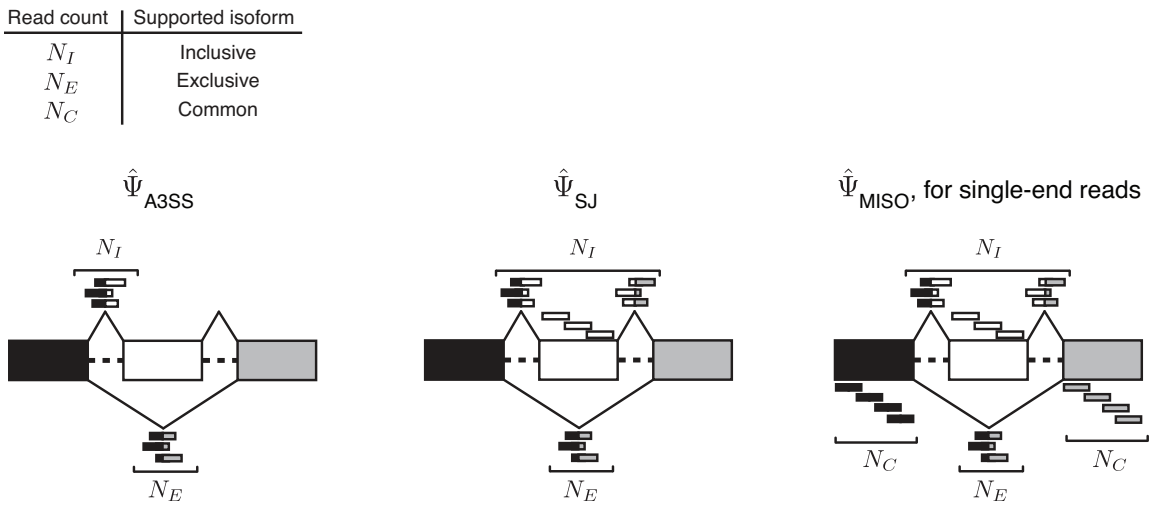
Supplementary Figure 1	Annotated isoforms in human and mouse genomes.
Supplementary Figure 2	Estimates of Ψ using single-end reads.
Supplementary Figure 3	Steps of MISO statistical inference procedure
Supplementary Figure 4	Evidence for exon size-dependent qRT-PCR bias.
Supplementary Figure 5	MISO Ψestimation for 52 alternative exons in breast cancer tissue.
Supplementary Figure 6	mRNA-Seq data for hnRNPAB in breast cancer tissue.
Supplementary Figure 7	Comparison of MISO $\Delta\Psi$ and qRT-PCR $\Delta\Psi$ values for hnRNP H knockdown dataset
Supplementary Figure 8	Comparison of read coverage fluctuations and gene expression for technical replicate libraries with short and long insert lengths.
Supplementary Figure 9	Graphical model representation of MISO for single-end reads.
Supplementary Figure 10	Random walk sampling scheme for inference in MISO.
Supplementary Figure 11	Sampling-based MISO inference algorithm.
Supplementary Figure 12	Isoform abundance estimation for four isoforms of *GRIN1* gene.
Supplementary Table 1	Short read datasets used in study
Supplementary Table 2	Events used in qRT-PCR validation of MISO on breast cancer data set
Supplementary Note	

Supplementary Figures



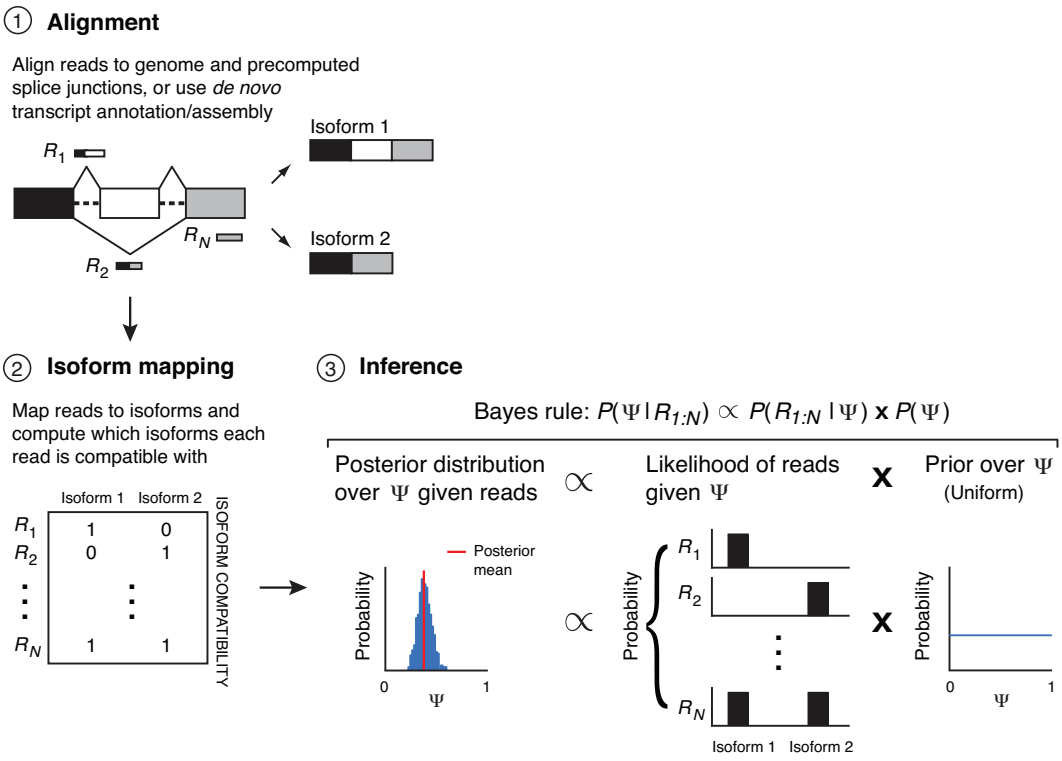
Supplementary Figure 1. Annotated isoforms in human and mouse genomes. Number of UCSC annotated mRNAs per gene, showing the large number of multi-isoform genes, even by conservative estimates that do not take into account RNA-Seq data.

Nature Methods: doi: 10.1038/nmeth.1528



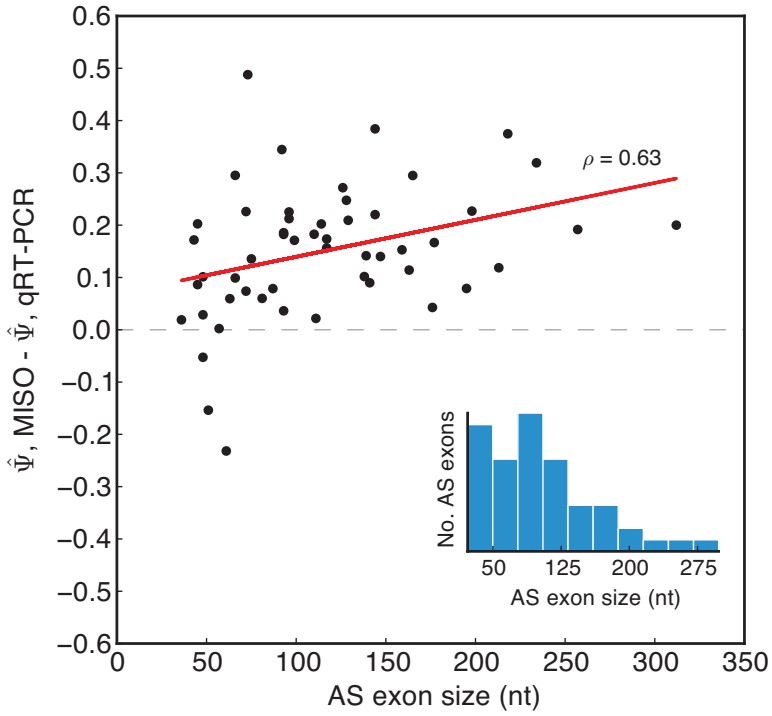
Supplementary Figure 2. Estimates of Ψ using single-end reads. The three Ψ estimates and the reads used in each estimate. N_I, N_E correspond to the number of reads supporting the inclusive isoform, respectively, while N_C corresponds to the number of reads supporting both isoforms (constitutive reads). The $\hat{\Psi}_{\text{SJ}}$ estimate shown corresponds to the estimate used in the majority of the analyses in[1], and $\hat{\Psi}_{\text{A3SS}}$ was also used in the same study for a subset of exons with an alternative splice site (see Supplementary Note for a proof of unbiasedness of these estimates.) The $\hat{\Psi}_{\text{MISO}}$ estimate shown corresponds to the analytic estimate from the MISO model (full derivation described in Supplementary Note), which is only obtained for single-end data. Estimates incorporate increasing amounts of information present in reads, with $\hat{\Psi}_{\text{A3SS}}$ using the least amount of information and $\hat{\Psi}_{\text{MISO}}$ using the most.

Nature Methods: doi: 10.1038/nmeth.1528



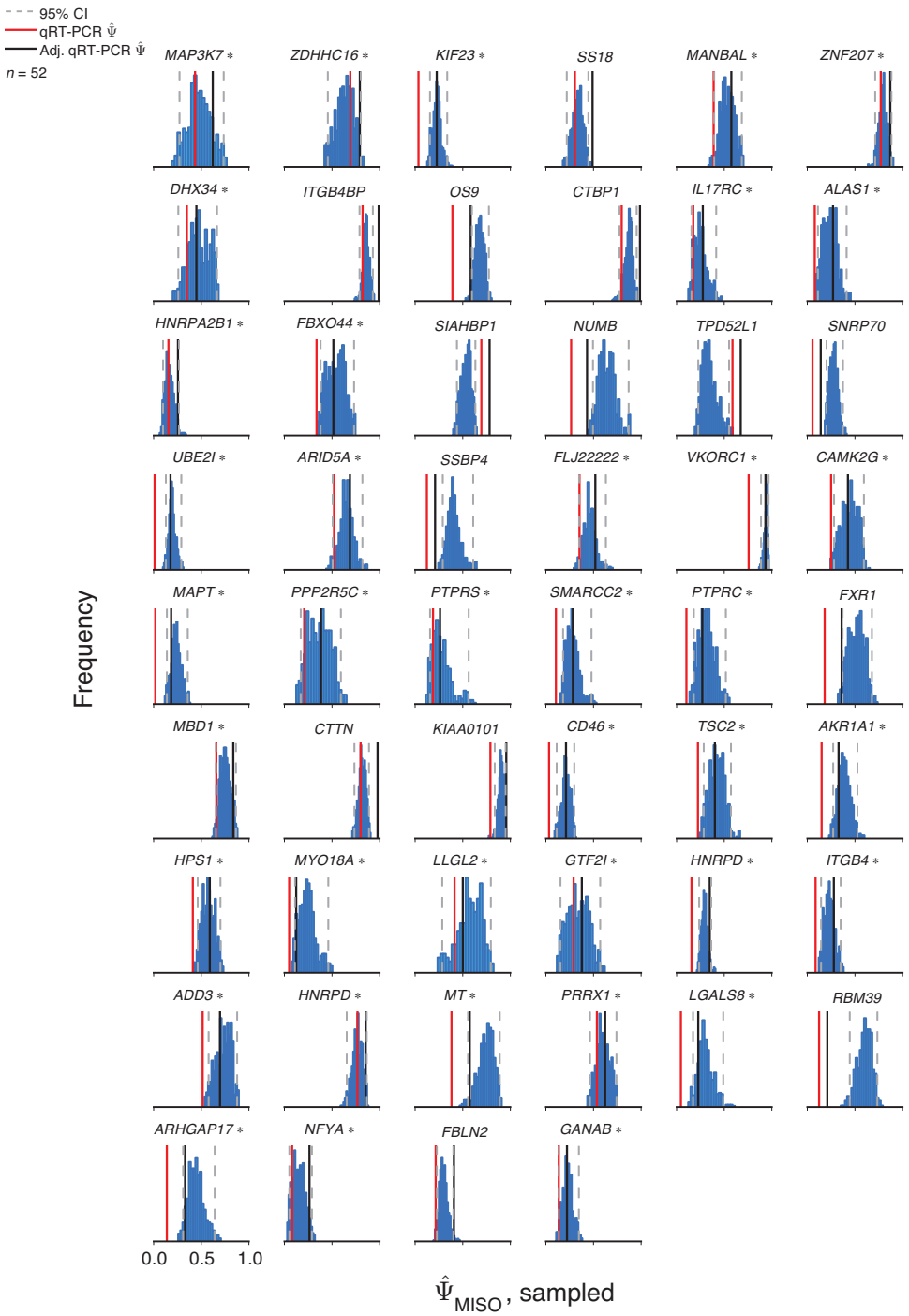
Supplementary Figure 3. Steps of MISO statistical inference procedure from. Reads are aligned to the genome and a set of junctions that are either precomputed or discovered de novo using transcript annotation/discovery tools. Aligned reads are then mapped to isoforms, shown here for the case of a skipped exon, and represented as binary matrices that correspond to their compatibility with isoforms. Each each row i in the isoform compatibility matrix corresponds to a read, and each column j to an isoform, where the ijth entry is 1 if read i is consistent with isoform j and 0 otherwise. In this example, read R_1 is consistent only with the inclusive isoform (containing the white exon), R_2 consistent only with the exclusive isoform (excluding the white exon), while R_N consistent with both. Inference is performed by computing a probability distribution (the posterior) over Ψ given the reads. Bayes' rule states that this distribution is proportional to the product of our expectation about the value of Ψ (the prior, here taken to be uniformly distributed over $[0, 1]$) and the likelihood of observing the reads given Ψ (the likelihood). By summing over all possible assignments of reads to isoforms, weighting each assignment by its probability, the posterior distribution over Ψ is obtained. Inferences are then summarized by the mean of the posterior distribution, used as an estimate of Ψ, and confidence intervals that quantitate the confidence in the estimate (as described in Online Methods.)

Nature Methods: doi: 10.1038/nmeth.1528



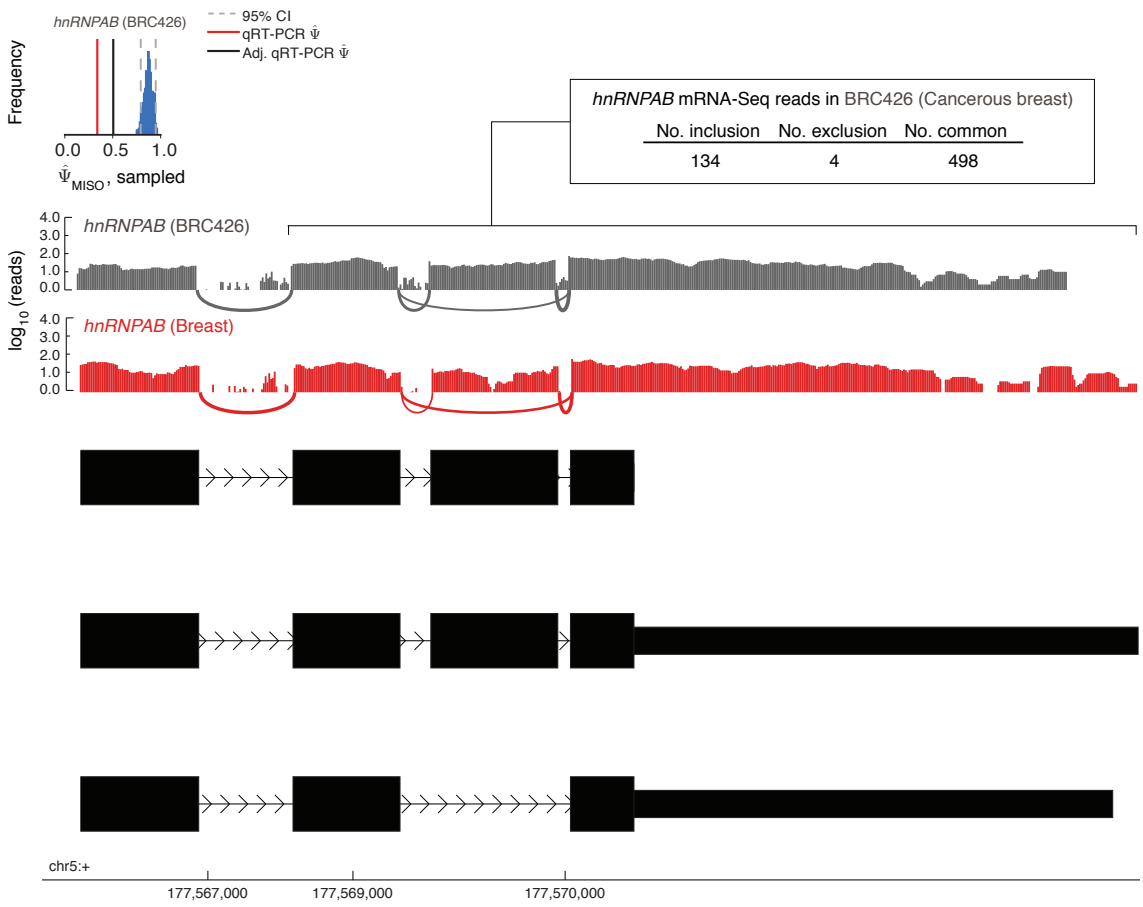
Supplementary Figure 4. Evidence for exon size-dependent qRT-PCR bias. Posterior marginals for each of the 52 alternative splicing events used in the breast cancer tissue sample[4].

Nature Methods: doi: 10.1038/nmeth.1528



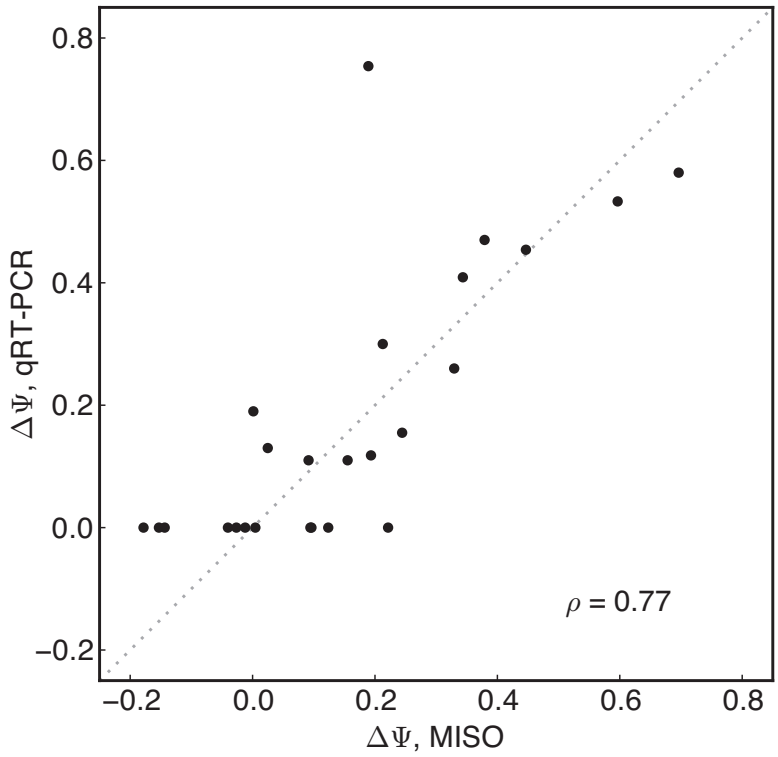
Supplementary Figure 5. MISO Ψ estimation for 52 alternative exons in breast cancer tissue. Posterior distributions for each of the 52 alternative splicing events used in the breast cancer tissue sample[4].

Nature Methods: doi: 10.1038/nmeth.1528



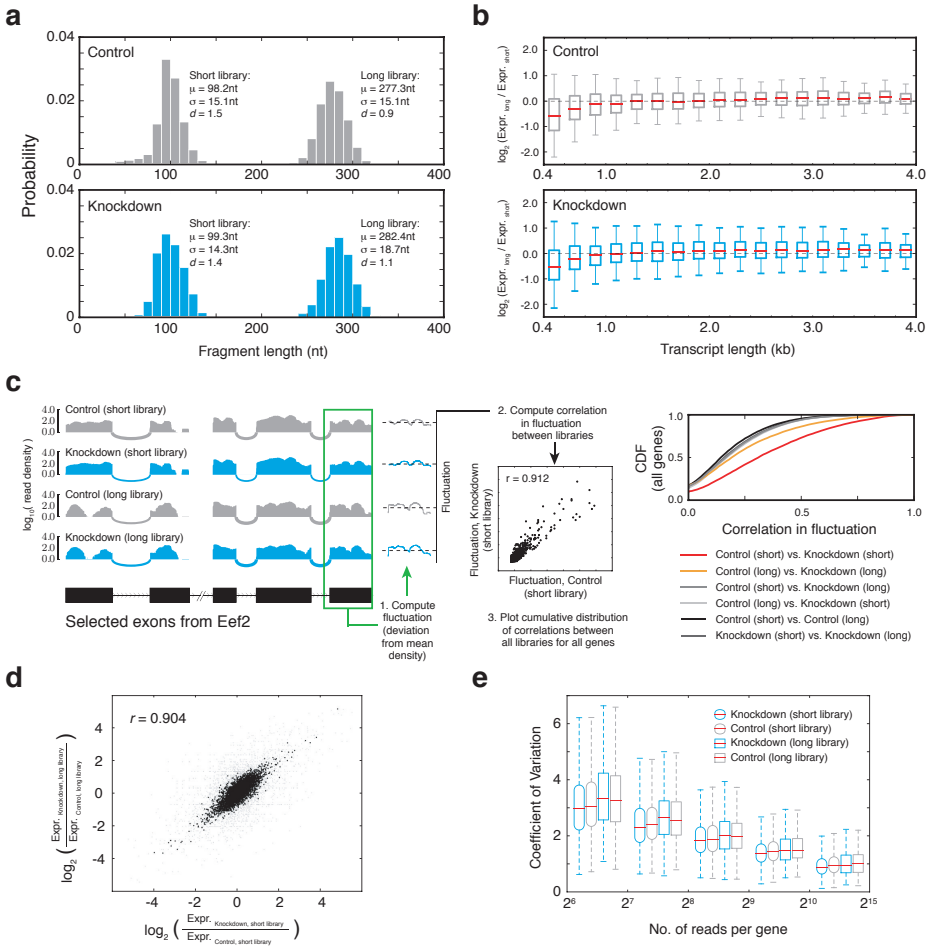
Supplementary Figure 6. mRNA-Seq data for hnRNPAB in breast cancer tissue. Data from breast cancer tissue (sample BRC426) from[4] and normal breast tissue (provided by Illumina, available on request).

Nature Methods: doi: 10.1038/nmeth.1528



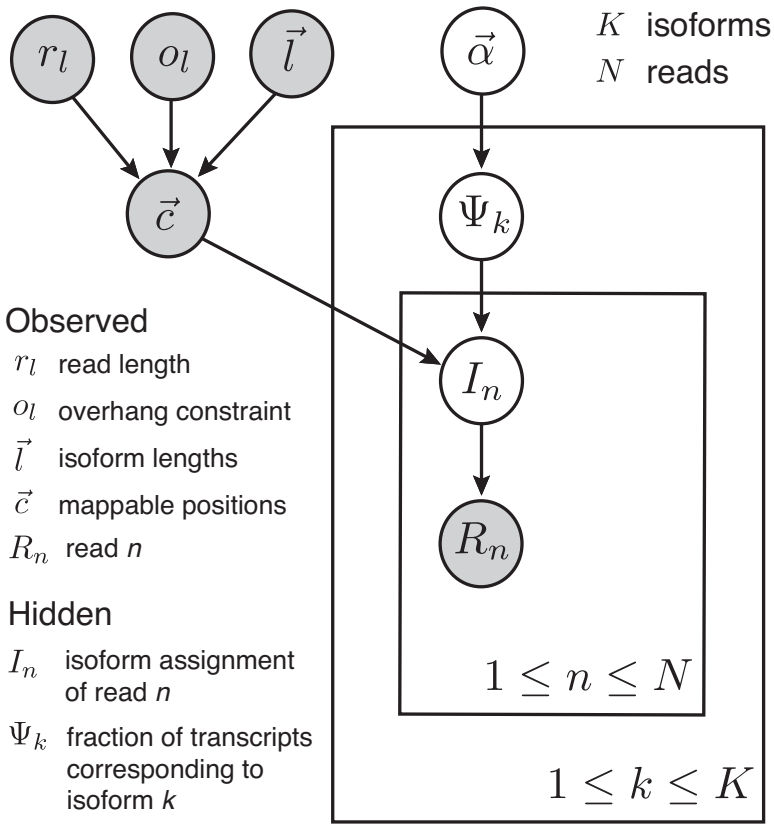
Supplementary Figure 7. Comparison of MISO $\Delta\Psi$ and qRT-PCR $\Delta\Psi$ values for hnRNP H dataset. Change in Ψ value based on MISO (x-axis) and qRT-PCR estimates (y-axis) for a set of 25 alternative exons in the hnRNP H control and knockdown data set are shown ($\rho = 0.77$).

Nature Methods: doi: 10.1038/nmeth.1528



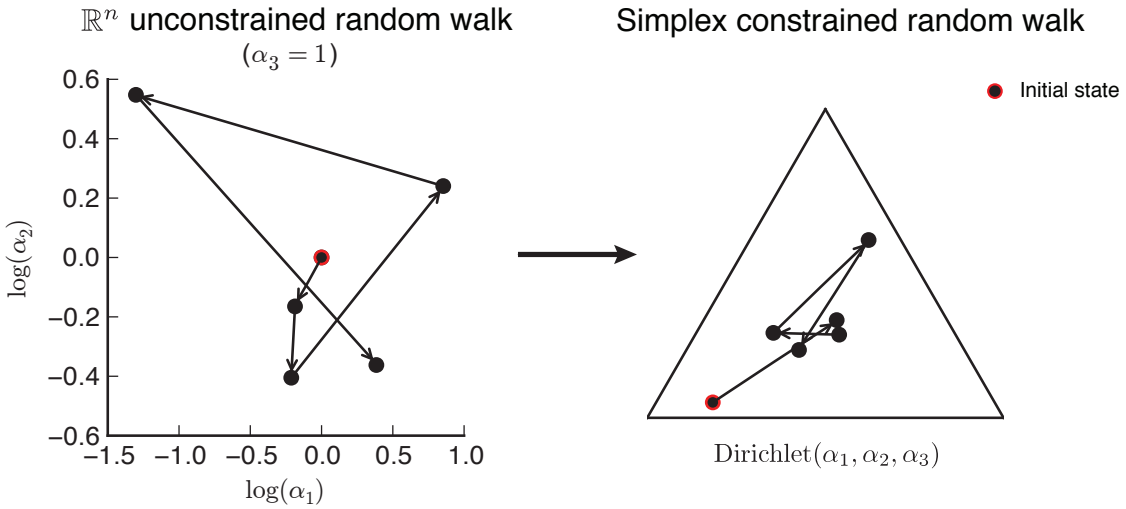
Supplementary Figure 8. Comparison of read coverage fluctuations and gene expression for technical replicate libraries with short and long insert lengths. (**a**) The insert length distribution of pairs of libraries made from the same batch of RNA with mean fragment lengths of ∼98 nt and ∼277 nt. These are shown in both control (black) and CUGBP1 knockdown conditions. (**b**) Fold change between the short and long technical replicate libraries in control (top, black) and knockdown (bottom, blue) conditions. (**c**) Read coverage varies across exons, as illustrated on selected exons on Eef2. Deviation of sequence coverage from the mean can be computed for each gene. These values can be correlated between libraries, and the cumulative distribution function of correlation values can be plotted. (**d**) Changes in gene expression are preserved between libraries of differing insert lengths. (**e**) The coefficient of variation in read coverage across genes does not markedly differ between libraries of differing insert lengths.

Nature Methods: doi: 10.1038/nmeth.1528



Supplementary Figure 9. Graphical model representation of MISO for single-end reads. A graphical representation of the probabilistic dependencies between variables in MISO, for a single gene with K isoforms. Shaded nodes represent observed variables, which include all the reads for the gene of interest, the parameters of the mRNA-Seq experiment and alignment procedure (the read length and the overhang length constraint) and features of the gene of interest (lengths of isoforms and the number of mappable positions in each isoform). The unshaded nodes represent random variables whose value are to be inferred from data, namely the Ψ value of each isoform k (Ψ_k) and the isoform from which each read was generated (I_n). The vector $\vec{\alpha}$ corresponds to the parameters of the Dirichlet prior distribution on isoform abundances, which is fixed to encode a uniform prior. MISO models the joint inference problem of finding the best set of Ψ values for the isoforms and the correct assignment of reads to the isoforms from which they were generated. For paired-end data, the probability of assigning a read to an isoform also depends on the parameters μ, σ of the insert length distribution, and are incorporated into the model as described in main text (for simplicity, these are not shown in the graphical model.)

Nature Methods: doi: 10.1038/nmeth.1528



Supplementary Figure 10. Random walk sampling scheme for inference in MISO. A five-step random walk sampled from a Logistic-Normal proposal distribution in log space of the parameters of a Dirichlet distribution (left). Each step in this random walk parameterizes a Dirichlet distribution, from which corresponding points on the simplex can be drawn (right). The use of the Logistic-Normal proposal distribution allows efficient exploration of the space of $\vec{\Psi}$ values.

Nature Methods: doi: 10.1038/nmeth.1528

Input: Set of reads R, set of isoforms G of a gene, number of iterations to run M
Output: Set S of sampled $\vec{\Psi}$ values

Initialize $\vec{\Psi}_t = \vec{\Psi}_0$ randomly

Initialize assignments of reads to isoforms consistently

Set $S = \{\}$

foreach Iteration $m = 1,...,M$ **do**
    Propose $\vec{\Psi}_{new}$ from a distribution centered around $\vec{\Psi}_t$
    Compute the probability α of accepting $\vec{\Psi}_{new}$ (using Metropolis-Hastings ratio)
    With probability α, set $\vec{\Psi}_{t+1} = \vec{\Psi}_{new}$, otherwise set $\vec{\Psi}_{t+1} = \vec{\Psi}_t$

    foreach Read $r \in R$ **do**

        foreach Isoform $g \in G$ **do**
            Compute probability $p_{r,g}$ of reassigning read r to isoform g
        end

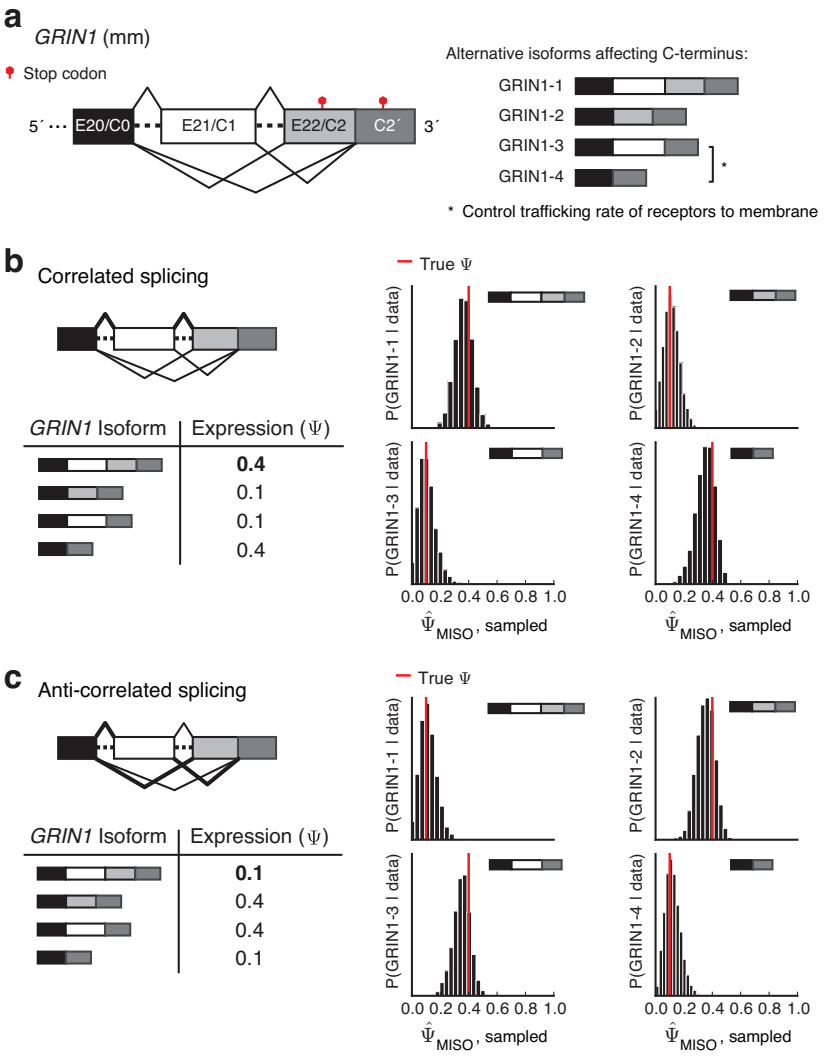
        Sample reassignment of read r to an isoform $g \in G$ based on computed probabilities
    end

    Set $S = S \cup \{\Psi_{t+1}\}$
end

return S

Supplementary Figure 11. Sampling-based MISO inference algorithm. Algorithm for estimating the posterior distribution over $\vec{\Psi}$ by Markov chain Monte Carlo sampling is shown. The algorithm begins with a random initialization of isoform distributions and assignments of reads to isoforms, and then repeatedly proposes new isoform distributions. These proposals are probabilistically accepted or rejected. If rejected, the previous isoform distribution is used in the next step. Each read is then probabilistically reassigned to one of the gene's isoforms, based on the new isoform distribution. As the algorithm converges, it is expected that an isoform distribution and associated assignment of reads to isoforms will be sampled in proportion to their probability under the model.

Nature Methods: doi: 10.1038/nmeth.1528



Supplementary Figure 12. Isoform abundance estimation for four isoforms of *GRIN1* gene. (**a**) A combination of alternative 3′ splice sites and exon skipping produces four isoforms in the region encoding the C-terminus of *GRIN1*[5,6]. (**b**) MISO Ψ estimates for reads simulated from an underlying isoform distribution where splicing of exons 21/C1 and 22/C2 is correlated, causing the two exons to be frequently included together (simulated isoform abundances shown at left). The posterior marginal distribution, estimated using the Monte Carlo algorithm described in Online Methods, is shown for each isoform, with the correct value shown as a vertical red line. (**c**) Estimation of isoform abundance for the case where reads were simulated from an *GRIN1* isoform distribution where splicing of exons 21/C1 and 22/C2 is anti-correlated, so the exons are rarely included together. The Ψ values for 21/C1 and 22/C2 are equal (0.5) in both conditions, but in the correlated condition the conditional probability of including 21/C2 given that 22/C2 is included is 0.8, while in the anti-correlated condition it is 0.2.

Nature Methods: doi: 10.1038/nmeth.1528

Supplementary Tables

Supplementary Table 1. Short read datasets used in study

Source	No. mapped reads	Run type
Human heart (Illumina, HiSeq 2000)	156 M	PE, 2x50nt
Human heart (Illumina, GA2)	30 M	PE, 2x54nt
Human testes (Illumina, GA2)	17 M	PE, 2x54nt
Human breast cancer tissue[13]	11 M	PE, 2x36nt
HEK 293T cells, control[14]	16 M	SE, 36nt
HEK 293T cells, hnRNP H knockdown[14]	21 M	SE, 36nt
HEK 293T cells, hnRNP H CLIP-Seq (this work)	4 M	SE, 36nt

Supplementary Table 2. Events used in qRT-PCR validation of MISO on breast cancer data set

Gene	Chromosome	Strand	ASE coords	ASE size	PCR Psi	Adj. PCR Psi	MISO Psi	PsiSJ
MAP3K7	6	-	91311072-91310992	81	0.43	0.62	0.49	0.43
ZDHHC16	10	+	99203546-99203593	48	0.69	0.79	0.64	0.62
KIF23	15	+	67520161-67520472	312	0.03	0.23	0.23	0.18
SS18	18	-	21869885-21869793	93	0.3	0.49	0.34	0.4
MANBAL	20	+	35360580-35360696	117	0.39	0.57	0.55	0.54
ZNF207	17	+	27712600-27712647	48	0.77	0.87	0.8	0.78
DHX34	19	+	52571970-52572044	75	0.35	0.45	0.48	0.47
ITGB4BP	20	-	33332046-33331871	176	0.82	0.99	0.86	0.91
OS9	12	+	56400149-56400313	165	0.39	0.58	0.69	0.73
CTBP1	4	-	1225307-1225113	195	0.8	0.99	0.88	0.87
IL17RC	3	+	9937609-9937653	45	0.17	0.27	0.26	0.22
ALAS1	3	+	52207728-52207904	177	0.08	0.27	0.24	0.18
HNRPA2B1	7	-	26204011-26203976	36	0.16	0.26	0.17	0.48
FBXO44	1	+	11641177-11641272	96	0.34	0.51	0.55	0.49
SIAHBP1	8	-	144974874-144974824	51	0.7	0.78	0.54	0.58
NUMB	14	-	72815885-72815742	144	0.27	0.43	0.65	0.51

TPD52L1	6	+	125619943-125620003	61	0.59	0.67	0.35	0.33
SNRP70	19	+	54297183-54297254	72	0.05	0.14	0.28	0.46
UBE2I	16	+	1302349-1302605	257	0.01	0.18	0.2	0.31
ARID5A	2	+	96578785-96578923	139	0.52	0.69	0.66	0.6
SSBP4	19	+	18403163-18403228	66	0.12	0.21	0.42	0.44
FLJ22222	17	-	77945708-77945496	213	0.35	0.52	0.47	0.38
VKORC1	16	-	31012243-31012134	110	0.76	0.93	0.94	0.88
CAMK2G	10	-	75249409-75249296	114	0.25	0.43	0.45	0.47
MAPT	17	+	41423081-41423278	198	0.02	0.18	0.24	0.23
PPP2R5C	14	+	101453921-101454037	117	0.21	0.38	0.38	0.32
PTPRS	19	-	5167778-5167731	48	0.19	0.26	0.29	0.26
SMARCC2	12	-	54853080-54852988	93	0.11	0.28	0.29	0.32
PTPRC	1	+	196938139-196938282	144	0.1	0.27	0.32	0.28
FXR1	3	+	182175795-182175886	92	0.18	0.36	0.53	0.35
MBD1	18	-	46053839-46053702	138	0.66	0.84	0.76	0.81
HNRPAB	5	+	177569739-177569879	141	0.34	0.51	0.88	0.86
CTTN	11	+	69945224-69945334	111	0.8	0.98	0.82	0.87
KIAA0101	15	-	62456157-62455995	163	0.79	0.96	0.9	0.73
CD46	1	+	206030221-206030313	93	0.03	0.21	0.22	0.12
TSC2	16	+	2067600-2067728	129	0.22	0.4	0.43	0.57
AKR1A1	1	+	45790695-45790822	128	0.15	0.33	0.4	0.26

HPS1	10	-	100179636-100179538	99	0.41	0.59	0.58	0.62
MYO18A	17	-	24436792-24436748	45	0.05	0.12	0.25	0.15
LLGL2	17	+	71082129-71082171	43	0.41	0.5	0.59	0.55
GTF2I	7	+	73771134-73771196	63	0.29	0.38	0.35	0.22
HNRPD	4	-	83496860-83496714	147	0.16	0.35	0.3	0.24
ITGB4	17	+	71262731-71262889	159	0.09	0.28	0.24	0.28
ADD3	10	+	111882053-111882148	96	0.51	0.7	0.74	0.59
HNRPD	4	-	83511761-83511705	57	0.77	0.85	0.77	0.83
MT	22	-	41863248-41863031	218	0.38	0.57	0.76	0.53
PRRX1	1	+	168966042-168966113	72	0.54	0.62	0.61	0.51
LGALS8	1	+	234772838-234772963	126	0.04	0.23	0.32	0.22
RBM39	20	-	33791933-33791861	73	0.12	0.21	0.61	0.71
ARHGAP17	16	-	24858419-24858186	234	0.14	0.33	0.46	0.41
NFYA	6	+	41156528-41156614	87	0.08	0.26	0.16	0.09
FBLN2	3	+	13638276-13638416	141	0.22	0.41	0.31	0.4
GANAB	11	-	62158423-62158358	66	0.14	0.22	0.23	0.43

Supplementary Note: Details of statistical estimators and MISO inference algorithm

Estimates of Ψ. Multiple approaches for estimating Ψ values have been proposed. One simple estimate, $\hat{\Psi}_{\mathsf{A3SS}}$, considers the splicing event as a choice between the competing $3'$ splice sites ($3'$ss) of the alternative exon and the downstream constitutive exon, estimating the inclusion of the exon by the relative numbers of reads that join these $3'$ss to the upstream constitutive exon (Supplementary Fig. 2). This estimate was previously used for a subset of alternative splicing events.[1]

A more comprehensive estimate is $\hat{\Psi}_{\mathsf{SJ}}$, which estimates exon inclusion based on the combined read density in the body of the alternative exon and in the two junctions that involve the alternative exon, relative to the density of junction reads that join the upstream and downstream constitutive exons[1] (Figure 1c). The remaining reads that align to the bodies of the flanking constitutive exons could have derived from either isoform and are not used in $\hat{\Psi}_{\mathsf{SJ}}$.

More formally, $\hat{\Psi}_{\mathsf{SJ}} = \frac{D_I}{D_I + D_E}$, where D_I is the density of inclusion reads and D_E the density of exclusion reads. Let e_l be the length of the alternatively spliced exon, r_l be the length of mRNA-seq reads, and o_l the overhang constraint placed on splice junctions. Assuming all positions in the gene of interest are uniquely mappable, D_I and D_E are computed as follows:

$$D_I = \frac{N_I}{e_l - r_l + 1 + 2(r_l + 1 - 2o_l)}, D_E = \frac{N_E}{r_l + 1 - 2o_l}$$

where N_I and N_E are the number of reads supporting inclusion and exclusion reads, respectively. If non-uniquely mappable read starting positions exist, these are simply subtracted from the denominators of D_I and D_E.

Computing the maximum a posteriori estimate $\hat{\Psi}_{\mathsf{MISO}}$ for single-end reads. As explained in the main text, constitutive reads contain latent information about Ψ, and can be used to improve and stabilize Ψ estimates (Figure 1d). For exon-centric analyses, an analytic estimate can be computed, if only single-end reads are used and certain assumptions are made about the prior distribution $P(\Psi)$. This estimate is denoted $\hat{\Psi}_{\mathsf{MISO}}$ and can be derived by computing the *maximum a posteriori* (MAP) estimate of Ψ under the MISO model. (Note that in a subset of figures in the main text, $\hat{\Psi}_{\mathsf{MISO}}$ is used alternatively to denote the mean of the posterior distribution over Ψ obtained by MCMC-based inference, as indicated by the figure legends.)

Since the prior $P(\Psi)$ is 1 when the hyperparameters $\alpha = \beta = 1$, the MAP estimate and the MLE estimates are equal, and so we proceed by finding the MLE. Given $R_{1:N}$ reads and their

Nature Methods: doi: 10.1038/nmeth.1528

isoform assignments $I_{1:N}$, the likelihood function $P(R_{1:N} \mid \Psi)$ is:

$$\begin{aligned} P(R_{1:N} \mid \Psi) &= \prod_{n=1}^{N} \sum_{I_n=1}^{2} P(R_n \mid I_n) P(I_n \mid \Psi) \\ &= \prod_{n=1}^{N} \left[\, P(R_n \mid I_n = 1)\Psi_f + P(R_n \mid I_n = 2)(1 - \Psi_f) \, \right] \end{aligned}$$

where $1 \leq n \leq N$. We'd like to find a value of $\hat{\Psi}$ that maximizes this likelihood, which is written as a function of Ψ_f. By the equivariance property of maximum likelihood, we can simply find the MLE $\hat{\Psi}_f$ of Ψ_f and transform it into $\hat{\Psi}$, since the two are one-to-one, using:

$$\hat{\Psi} = \frac{c_1 \hat{\Psi}_f}{c_1 - c_1 \hat{\Psi}_f + c_2 \hat{\Psi}_f} \tag{1}$$

To simplify the notation, let p_1 and p_2 stand for the probabilities of a read being generated from the first and second isoforms, respectively, assuming read lengths r_l:

$$p_1 = \frac{1}{m(rl, I_1)}, p_2 = \frac{1}{m(rl, I_2)}$$

Substituting our observation model into the likelihood gives:

$$\begin{aligned} P(R_{1:N} \mid \Psi_f) &= \prod_{n=1}^{N} \left(P(R_n \mid 1, \Theta)\Psi_f + P(R_n \mid 2, \Theta)(1 - \Psi_f) \right) \\ &= \prod_{n=1}^{N} \left[\, p_1 R_n^1 \Psi_f + p_2 R_n^2 (1 - \Psi_f) \, \right] \end{aligned}$$

Taking the log, we have:

$$\hat{\Psi}_f = \arg\max_{\Psi_f} \sum_{n=1}^{N} \log \left(p_1 R_n^1 \Psi_f + p_2 R_n^2 (1 - \Psi_f) \right)$$

Differentiating and setting the derivative to zero yields:

$$\begin{aligned} \frac{d}{d\Psi_f} \sum_{n=1}^{N} \log \left(p_1 R_n^1 \Psi_f + p_2 R_n^2 (1 - \Psi_f) \right) &= 0 \\ \sum_{n=1}^{N} \frac{p_1 R_n^1 - p_2 R_n^2}{p_1 R_n^1 \Psi_f + p_2 R_n^2 (1 - \Psi_f)} &= 0 \end{aligned}$$

Nature Methods: doi: 10.1038/nmeth.1528

This sum can be rewritten in terms of three sufficient statistics: the number of reads supporting only isoform 1 (N_I), the number of reads supporting only isoform 2 (N_E), and the number of reads supporting both isoforms (N_C), to get:

$$\frac{N_I}{\Psi_f} - \frac{N_E}{1-\Psi_f} + \frac{N_C(p_1-p_2)}{p_1\Psi_f + p_2(1-\Psi_f)} = 0$$

This equation reduces to solving a quadratic equation, whose relevant solution is:

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

$$\begin{aligned}
A &= N_I p_1 + N_C p_1 - 2N_I p_2 - N_E p_2 - N_C p_2 \\
B &= 4N_I p_2(N_I p_1 + N_E p_1 + N_C p_1 - N_I p_2 - N_E p_2 - N_C p_2) \\
C &= (-N_I p_1 - N_C p_1 + 2N_I p2 + N_E p_2 + N_C p_2)^2 \\
D &= 2(N_I p_1 + N_E p_1 + N_C p_1 - N_I p_2 - N_E p_2 - N_C p_2)
\end{aligned}$$

Now, $\hat{\Psi}_f$ can be plugged in to Equation 1 to obtain $\hat{\Psi}$, which is our MAP/MLE estimate. This resulting estimate is simply a function of the read counts N_I, N_E, N_C and the probabilities p_1 and p_2.

Proof that $\hat{\Psi}_{\mathsf{A3SS}}$ is unbiased. As an example of analytic estimates of Ψ, we show that the simplest estimate, $\hat{\Psi}_{\mathsf{A3SS}}$, is unbiased. Recall that $\hat{\Psi}_{\mathsf{A3SS}}$ uses only the reads from one inclusion junction and from the exclusion junction (Supplementary Fig. 2). Given a read length r_l and an overhang constraint of o_l, let J be the number of possible read starting positions in a junction:

$$J = r_l + 1 - 2o_l$$

Then $\hat{\Psi}_{\mathsf{A3SS}}$ can be defined as follows:

$$\begin{aligned}
\hat{\Psi}_{\mathsf{A3SS}} &= \frac{\frac{N_I}{J}}{\frac{N_I}{J} + \frac{N_E}{J}} \\
&= \frac{N_I}{N_I + N_E}
\end{aligned}$$

Proposition 1 (Unbiasedness of $\hat{\Psi}_{\mathsf{A3SS}}$) *The symmetric splice junction estimator $\hat{\Psi}_{\mathsf{A3SS}}$ is unbiased, i.e. $E(\hat{\Psi}_{\mathsf{A3SS}}) = \Psi$ for all Ψ.*

Nature Methods: doi: 10.1038/nmeth.1528

Proof Fix Ψ. Let l_1 be the length of the inclusive isoform, and l_2 be the length of the exclusive isoform. Then the number of reads possible from the two isoforms are c_1, c_2, respectively:

$$c_1 = l_1 - r_l + 1$$
$$c_2 = l_2 - r_l + 1$$

These constants are used to compute Ψ_f, the probability of sequencing a read from the inclusive isoform, which is defined as follows:

$$\Psi_f = \frac{c_1\Psi}{c_1\Psi + c_2(1-\Psi)}$$

Recall that $J = r_l + 1 - 2o_l$. In general, a read generated in our model falls into one of four mutually categories. It could support: (1) the inclusive isoform, (2) the exclusive isoform, (3) both isoforms, or (4) be thrown out due to an overhang violation. Relative to this space of outcomes, the expected probabilities of inclusion and exclusion reads are as follows:

$$\begin{aligned} P(N_I) &= P(\text{inclusive isoform})P(\text{inclusion junction read} \mid \text{inclusive isoform}) \\ &= \Psi_f \frac{J}{c_1} \\ P(N_E) &= P(\text{exclusive isoform})P(\text{exclusion junction read} \mid \text{exclusive isoform}) \\ &= (1-\Psi_f)\frac{J}{c_2} \end{aligned}$$

To show unbiasedness, it suffices to show that $E(\frac{N_I}{N_I+N_E}) = \Psi$. Since $\hat{\Psi}_{\mathsf{A3SS}}$ uses only the N_I and N_E reads, we know that $N_I + N_E = n$, where n is the total number of reads used in the estimate. Therefore,

$$E\left(\frac{N_I}{N_I+N_E}\right) = E\left(\frac{N_I}{n}\right) = \frac{1}{n}E(N_I)$$

The expected number of inclusion reads in a sample of n reads, $E(N_I)$, is simply $n \times P(N_I)$. Since reads other than inclusion or exclusion reads are discarded in the $\hat{\Psi}_{\mathsf{A3SS}}$ estimate, the probability of an inclusion read must be normalized to account for the fact that these are the only two outcomes:

$$\begin{aligned} E(N_I) &= n \times \frac{P(N_I)}{P(N_I) + P(N_E)} \\ &= n \times \frac{\Psi_f \frac{J}{c_1}}{\Psi_f \frac{J}{c_1} + (1-\Psi_f)\frac{J}{c_2}} \end{aligned}$$

Nature Methods: doi: 10.1038/nmeth.1528

Substituting Ψ_f with its definition results in:

$$
\begin{aligned}
E(N_I) &= n \times \frac{\frac{c_1\Psi}{c_1\Psi + c_2(1-\Psi)}\frac{J}{c_1}}{\frac{c_1\Psi}{c_1\Psi + c_2(1-\Psi)}\frac{J}{c_1} + \left(1 - \frac{c_1\Psi}{c_1\Psi + c_2(1-\Psi)}\right)\frac{J}{c_2}} \times \frac{c_1\Psi + c_2(1-\Psi)}{c_1\Psi + c_2(1-\Psi)} \\
&= n \times \frac{\Psi J}{\Psi J + (c_1\Psi + c_2(1-\Psi) - c_1\Psi)\frac{J}{c_2}} \\
&= n \times \frac{\Psi J}{\Psi J + (1-\Psi)J} \\
&= n \times \frac{\Psi}{\Psi + (1-\Psi)} \\
&= n \times \Psi
\end{aligned}
$$

Thus, $\frac{1}{n}E(N_I) = \Psi$, which demonstrates that $\hat{\Psi}_{\mathsf{A3SS}}$ is unbiased. A similar argument holds for the $\hat{\Psi}_{\mathsf{SJ}}$ estimate used in[1].

Efficient estimation of isoform distributions for genes with many isoforms. We devised a Markov chain Monte Carlo (MCMC) inference scheme based on a novel proposal distribution. Considering the length information and length correction in our problem leads to violations of the mathematically convenient conjugacy properties of traditional Dirichlet-Multinomial mixture models. For this reason, the use of a standard Gibbs sampler is not possible. Instead, we use a hybrid MCMC sampler that combines the Metropolis-Hastings (MH) algorithm with a Gibbs sampler[2]. 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 be easily 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:

$$
\alpha = \min\left(\frac{P(\vec{x}_{t+1})Q(\vec{x}_t; \vec{x}_{t+1})}{P(\vec{x}_t)Q(\vec{x}_{t+1}; \vec{x}_t)}, 1\right) \qquad \text{(MH ratio)}
$$

where α is the probability of transitioning to the proposed state $\vec{x}_{t+1}$ from the current state $\vec{x}_t$. The better the proposal distribution Q is at proposing probable values under P, the faster the sampling algorithm will converge to the correct distribution.

In our case, the target distribution is the posterior distribution on $\vec{\Psi}$ given a set of reads, $P(\vec{\Psi} \mid R_{1:N})$. In general, we expect any set of reads from a gene with many isoform to be well-explained by only a small set of closely related isoform distributions. In other words, we expect the model's probability mass to be peaked on a small set of $\vec{\Psi}$ values that explain the data, with little probability mass on other $\vec{\Psi}$ that encode a very different set of isoform abundances.

Nature Methods: doi: 10.1038/nmeth.1528

In light of this unimodal probability landscape, a proposal distribution that uniformly proposes random isoform distributions is unlikely to find values that fit the data. A standard strategy for solving problems of this form using sampling is to use a proposal distribution that "drifts"—or forms a *random walk*—over the sampled variable's state space. In a random walk proposal, the proposed value is typically the previously sampled value plus some noise (e.g. the previous proposal, corrupted by normally distributed noise.)

A challenge in defining a random walk proposal in our case is that isoform distributions are constrained to sum to one—i.e., they must be probability distributions. Therefore, a random walk where a proposal is drawn from a normal distribution centered on the previously sampled isoform distribution will not work. To overcome this, we formulated a random walk using the *Logistic-Normal distribution*[3], a distribution on the simplex that generalizes the more commonly used Dirichlet distribution. With the Logistic-Normal it is possible to to formalize the idea that the newly proposed isoform distribution is drawn from a distribution whose mean is the previously sampled distribution, meaning that only small changes to the current isoform distribution are proposed, while still respecting the constraint that proposed values must sum to one. Intuitively, this allows the algorithm to 'hone in' on the region of highly probable isoform distributions for a given data set, and move around in that space, without spending too much time sampling lower probability regions.

The random walk is defined over the parameters of the distribution from which $\vec{\Psi}$ is drawn, in log space, allowing the sampled values to range unconstrained over the space of real numbers. Each draw of a set of parameters then parameterizes a Dirichlet distribution from which an isoform distribution is drawn. Supplementary Figure 10 shows proposals drawn according to this process, illustrating how a five-step unconstrained random walk on the parameters of the distribution in log space induces a random walk in the constrained space of the 2D simplex, where each point represents a probability vector. Our algorithm exploits the fact that a random walk over the parameters of a distribution—which can be conveniently 'drifted over' unconstrained—can be used to define a random walk over the values drawn from this distribution, which in this case are constrained to lie within the simplex.

Our sampling algorithm, shown in Supplementary Figure 11, proceeds by repeatedly proposing new values for $\vec{\Psi}$, which are stochastically accepted in proportion to their probability under the model. First, a new isoform distribution is proposed, which is probabilistically accepted or rejected based on the MH ratio, as described above. For each proposed isoform distribution, the algorithm then probabilistically reassigns each read to a new isoform, which completes one iteration of the

Nature Methods: doi: 10.1038/nmeth.1528

algorithm. As the number of iterations increases, the algorithm is guaranteed to eventually sample isoform distributions and assignments of reads to isoforms in proportion to their posterior probability in our model.

A distribution that can capture the desired random walk in simplex space is the *Logistic-Normal* distribution, parameterized by a mean $\vec{\mu}$ and covariance matrix Σ, and denoted $L_k(\vec{\mu}, \Sigma)$ for the k-dimensional case[3]. A k-dimensional vector $\vec{\theta}$ can be sampled from $L_{k-1}(\vec{\mu}, \Sigma)$ by first sampling a vector v from a multivariate normal distribution Normal$(\vec{\mu}, \Sigma)$ and then taking its inverse logistic transform, logit^{-1}:

$$\vec{\theta} = \text{logit}^{-1}(v) = \frac{e^v}{1 + \sum_{k=1}^{K} e^{v_k}}$$

The vector v can be obtained back via the logit transform $v = \log(\theta/\theta_{k+1})$, where $\theta_{k+1} = 1 - \sum_{k=1}^{K} \theta_k$. The probability density for $\vec{\theta}$ with parameters $\vec{\mu}, \Sigma$ is:

$$|2\pi\Sigma|^{-\frac{1}{2}} \left(\prod_{j=1}^{k+1} \theta_j\right)^{-1} \exp\left[-\frac{1}{2}\{\log(\theta/\theta_{k+1}) - \vec{\mu}\}^{\mathrm{T}} \Sigma^{-1} \{\log(\theta/\theta_{k+1})\}\right] \tag{2}$$

Given a fixed covariance matrix Σ for the proposal distribution, the sampling scheme is as follows:

1. Initialize $\vec{\mu}_t, \vec{\Psi}_t = \vec{\mu}_0, \vec{\Psi}_0$, and initialize I to a consistent assignment
2. For $m = 1, \ldots, M$ iterations,
   (a) Propose $\vec{\mu}_{t+1}, \vec{\Psi}_{t+1}$:

$$\vec{\mu}_{t+1} \sim \text{Normal}(\vec{\mu}_t, \Sigma)$$
$$\vec{\Psi}_{t+1} = \text{logit}^{-1}(\vec{\mu}_{t+1})$$

   (b) Let $\vec{\mu}_t, \vec{\Psi}_t = \vec{\mu}_{t+1}, \vec{\Psi}_{t+1}$ with probability α (otherwise, keep values from step t):

$$\alpha = \min\left(\frac{P(\vec{\Psi}_{t+1}, I_{1:N}, R_{1:N}) Q(\vec{\Psi}_t; \vec{\Psi}_{t+1})}{P(\vec{\Psi}_t, I_{1:N}, R_{1:N}) Q(\vec{\Psi}_{t+1}; \vec{\Psi}_t)}, 1\right)$$

   Here, Q is a Logistic-Normal (following the probability density given in Equation 2) with mean equal to the previous time step's $\vec{\Psi}$, excluding its last element:

$$Q(\vec{\Psi}_{t+1}; \vec{\Psi}_t) \sim L_{k-1}(\log([\Psi_t^1, \ldots, \Psi_t^{k-1}]), \Sigma)$$

   And similarly for $Q(\vec{\Psi}_t; \vec{\Psi}_{t+1})$. The joint distributions in the MH ratio factor into a product of conditionals, as explained in Online Methods.

Nature Methods: doi: 10.1038/nmeth.1528

(c) Gibbs step: for $j = 1, \ldots, N$ reads,

i. Compute the probability $a_{j,k}$ of reassigning read R_j to the kth isoform, for every isoform $1 \leq k \leq K$:

$$a_{j,k} = P(I_j = k \mid R_{1:N}, I_{1:N} \backslash \{I_j\}, \vec{\Psi}_{t+1})$$

ii. Sample reassignment of $I_j \sim \text{Multinomial}(1, [a_{j,1} \ \cdots \ a_{j,K}])$

Note that Step (c) is the usual Gibbs sampling step for reassigning data points to components in mixture models.

Cross-validation adjustment of qRT-PCR values. Given the apparent length bias in the qRT-PCR estimates of Ψ in Figure 2, we computed an adjusted set of qRT-PCR Ψ values using cross-validation, in order to estimate the overlap between these values and the Bayesian confidence intervals of $\hat{\Psi}_{\mathsf{MISO}}$. Exons were binned by length ($n = 3$ bins) and the events used for validation were split in four sets. An adjusted qRT-PCR estimate of Ψ was then computed for each set by adding the average $\Delta\Psi$ of MISO and qRT-PCR estimates in the remaining three sets, bounding the resulting value in $[0, 1]$. The adjusted and raw Ψ estimates for MISO and qRT-PCR, along with the Bayesian confidence intervals, are shown in Supplementary Figure 5.

Nature Methods: doi: 10.1038/nmeth.1528

References

1. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456** (2008).

2. Liu, J. S. *Monte Carlo Strategies in Scientific Computing (Springer Series in Statistics)* (Springer, 2008).

3. Aitchison, J. & Shen, S. M. Logistic-normal distributions: Some properties and uses. *Biometrika* **67** (1980).

4. Venables, J. P. *et al.* Cancer-associated regulation of alternative splicing. *Nature Structural & Molecular biology* **16**, 670–676 (2009).

5. Zukin, R. & Bennett, M. Isoforms of the NMDAR1 receptor subunit. *Trends in Neurosciences* **18** (1995).

6. Lee, J.-A. *et al.* Depolarization and CaM Kinase IV Modulate NMDA Receptor Splicing through Two Essential RNA Elements. *PLoS Biology* **5** (2007).

Nature Methods: doi: 10.1038/nmeth.1528